RESEARCH PAPERS

Acta Cryst. (1998). D**54**, 159–174

A General Phasing Algorithm for Multiple MAD and MIR Data

JORDI BELLA AND MICHAEL G. ROSSMANN*

Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA.
E-mail: mgr@indiana.bio.purdue.edu

(*Received 10 April 1997; accepted 22 July 1997*)

Abstract

A phasing algorithm is presented for combining multiple wavelength anomalous dispersion (MAD) data from multiple types of anomalous scatterers, either in the same or in different derivative crystals, as well as for combining MAD data with multiple isomorphous replacement (MIR) data from different derivative crystals. A heavy-atom phasing and refinement program originally written by Rossmann [(1967) *HATOMLSQ* program, Purdue University, West Lafayette, Indiana, USA] has been modified to refine the parameters that define the anomalous and isomorphous scatterers and to determine protein phases by using all MAD and MIR derivatives simultaneously. The technique allows for appropriate weighting of every data set, including the native data, which contains neither an anomalous nor an isomorphous component. This method is a generalization of currently used heavy-atom methods. Numerical tests are presented for different experimental scenarios, including a double MAD experiment on the same crystal (diffraction data at two absorption edges), combination of two MAD experiments on different crystals, and combination of MAD data with MIR data from multiple crystals. An appendix shows how the Karle equations used in MAD phasing can be reformulated as a particular case of this algorithm.

1. Introduction

The Harker diagram (Harker, 1956) is the direct or implied basis for all methods used to determine and refine phases with isomorphous or anomalous dispersion data. Blow & Crick (1959) suggested a general method for treating error in such phase determination. A simplification of the error treatment was given by Dickerson, Kendrew & Strandberg (1961), which also served to popularize the procedure. This assumes that there is no error in the native structure amplitudes and that all the error is contained in the parameters that describe the heavy-atom or anomalous scatterers, as well as in the amplitudes of the isomorphous or anomalous diffraction data. Furthermore, the method requires a knowledge of the native structure amplitude if a phase has to be determined. Hence, the frequent practice is, when using popular programs such *MLPHARE* (Otwinowski, 1991) for the refinement of anomalous phases, to treat one of the available data sets (usually the one with the smaller Bijvoet differences) as the native data set with no error. Cullis, Muirhead, Perutz, Rossmann & North (1961) presented an alternative geometrical construction of the Harker diagram which allowed for equivalent treatment of all isomorphous data sets. We present here the Cullis *et al.* (1961) technique adapted to the simultaneous refinement of isomorphous and anomalous dispersion data and show its relationship to the solution of the Karle equations (Karle, 1980), often used for the interpretation of multiple wavelength anomalous dispersion (MAD) data.

The genesis of the work described here originates with our attempts to solve the crystal structure of the first two domains of intercellular adhesion molecule-1 (Kolatkar, Oliveira *et al.*, 1992) using the MAD phasing method in a poorly substituted single-site selenomethionyl derivative.

2. Description of the phasing method

2.1. *The phasing method of Blow & Crick (1959) and Dickerson et al. (1961)*

If $\mathbf{F}_{\mathrm{PH}i}$ and $\mathbf{F}_{\mathrm{H}i}$ are the complex structure factors of the ith isomorphous heavy-atom derivative and its respective heavy-atom constellation, and if $\mathbf{F}_{\mathrm{P}}$ is the corresponding unsubstituted parent (native) structure factor, then

$$\mathbf{F}_{\mathrm{PH}i} = \mathbf{F}_{\mathrm{P}} + \mathbf{F}_{\mathrm{H}i}.$$

From this it follows (Harker, 1956) that

$$F^2_{\mathrm{PH}i} = F^2_{\mathrm{P}} + F^2_{\mathrm{H}i} + 2F_{\mathrm{P}}F_{\mathrm{H}i}\cos(\psi_i - \alpha), \qquad (1)$$

where ψ_i is the phase of $\mathbf{F}_{\mathrm{H}i}$ and α the phase of $\mathbf{F}_{\mathrm{P}}$. $F_{\mathrm{PH}i}$, $F_{\mathrm{H}i}$ and F_{P} are the respective amplitudes of the derivative, heavy-atom, and protein structure factors. If the positions of the heavy atoms in the unit cell are known, $\mathbf{F}_{\mathrm{H}i}$ can be calculated and then α can be determined using (1).

Every derivative equation has two possible solutions for α, which correspond to two equally probable protein phase angles. In an error-free situation, there will be a unique common solution for all the derivatives, which in the Argand diagram will result in all circles from protein and derivatives intersecting at the same point. In an experimental situation, however, several types of errors concur, and the different circles do not intersect exactly. This is the case represented in Fig. 1. Every pair parent-ith derivative will produce a solution $\mathbf{F}_{Pi} = F_P \exp i\alpha_i$, which will be close to, but not exactly the same as, the $\mathbf{F}_{Pj}$ solution from the jth derivative. Thus, some uncertainty as to the correct value of the protein phase remains. The simplest estimate of α would be an average between all the calculated α_i's. The most common approach (Blow & Crick, 1959; Dickerson *et al.*, 1961) involves the calculation of the right-hand side of (1) at regular intervals of α, using the experimental values of F_P and calculated values for F_{Hi} and ψ_i, and then comparing the result with the experimental value of F_{PHi}, where

$$\varepsilon_i(\alpha) = F_{PHi} - [F_P^2 + F_{Hi}^2 + 2F_P F_{Hi}\cos(\psi_i - \alpha)]^{1/2}. \quad (2)$$

Here $\varepsilon_i(\alpha)$ is the calculated lack-of-closure error between the parent and the ith derivative when the parent phase angle has a value of α. The smaller the lack-of-closure error, the higher is the probability that the protein phase angle is correct.

The probability $P_i(\alpha)$ of the protein phase angle having a value of α, as determined from the ith derivative, can be evaluated by assuming a Gaussian error distribution for $\varepsilon_i(\alpha)$ (Blow & Crick, 1959; Dickerson *et al.*, 1961) such that

$$P_i(\alpha) \propto \exp\frac{-\varepsilon_i^2(\alpha)}{2E_i^2}, \quad (3)$$

where E_i is the estimated standard error of the $\varepsilon_i(\alpha)$ distribution. The joined probability can then be calculated as the product of the individual probabilities from every derivative (Rossmann & Blow, 1961), where

$$P(\alpha) = \prod_{i=1}^{N} P_i(\alpha) = K \exp\left[-\sum_{i=1}^{N}\frac{\varepsilon_i^2(\alpha)}{2E_i^2}\right]. \quad (4)$$

K is a normalization factor that ensures that the sum of probabilities for all possible α values is equal to 1, and N is the number of heavy-atom derivatives.

The usual Dickerson *et al.* (1961) phase determination assumes that all errors accumulate on the experimental quantities F_{PHi}, without regard for any error in the measured F_P amplitudes or the parameters that determine $\mathbf{F}_{Hi}$. Indeed, it is usually assumed that $\mathbf{F}_{Hi}$ can be parametrized by spherically symmetrical atoms, which is certainly not the case in the event of any loss of isomorphism. Nevertheless, Blow & Crick (1959) showed that, to a good approximation, errors in F_P and F_{PHi} may be treated as a single error. They also showed that errors in $\mathbf{F}_{Hi}$ can be convoluted with the error in F_P and F_{PHi}, thus demonstrating that, to a reasonable approximation, all errors could be considered to reside in F_{PHi}.

2.2. *The Cullis et al. (1961) method*

In this method, all F_{PHi} and F_P observations are treated in an equivalent manner, implicitly assuming the existence of errors in all sets of amplitudes. It was first used by Cullis *et al.* (1961) for the phase determination of horse oxy-hemoglobin at 5.5 Å resolution.

First, circles of radii F_{PHi} are drawn for every derivative, including the measured native, as shown in Fig. 1. Then, the amplitude of the native structure factor is calculated for every derivative and phase angle, α, as the length of a radius vector of slope $\tan\alpha$ that intercepts the circle of radius F_{PHi} (Fig. 2). Analytically, $F_{Pi}(\alpha)$ can be expressed as

$$\begin{aligned} F_{Pi}(\alpha) = &- F_{Hi}\cos(\psi_i - \alpha) \\ &\pm [F_{PHi}^2 - F_{Hi}^2\sin^2(\psi_i - \alpha)]^{1/2}. \end{aligned} \quad (5)$$

Next, the mean value of these native structure-factor amplitudes, $\langle F_P(\alpha)\rangle$, is calculated at every phase angle α,

$$\langle F_P(\alpha)\rangle = \frac{1}{N+1}\sum_{i=0}^{N} F_{Pi}(\alpha), \quad (6)$$

where the measured native structure factor is treated as one more circle of radius F_{P0} centered at the coordinate origin, so that $\mathbf{F}_{H0} = 0$. $N + 1$ is the total number of circles used for phase determination, from N derivative and one native (if present) measurements.

The most probable phase is the one that gives the minimum variance of the mean value $\langle F_P(\alpha)\rangle$, that is, the intersections of the different circles with the radius vector of phase α are minimally spread. The probability function at every value of α is calculated by

$$P_i(\alpha) = \exp\frac{-[\langle F_P(\alpha)\rangle - F_{Pi}(\alpha)]^2}{2E_i^2}, \quad (7)$$

and the joined probability as the product of the individual probabilities,

$$P(\alpha) = K \exp\left\{-\sum_{i=0}^{N}\frac{1}{2E_i^2}[\langle F_P(\alpha)\rangle - F_{Pi}(\alpha)]^2\right\}. \quad (8)$$

The double sign in (5) comes from the double intersection of a circle of radius F_{PHi} with a straight line of slope $\tan\alpha$ (Fig. 3). It is easy to demonstrate that, for those cases where one solution of F_{Pi} is positive and the other negative, the positive solution corresponds to a phase of α and the negative solution corresponds to a phase of $\pi + \alpha$. Some special situations arise (Cullis *et al.*, 1961) when both solutions are of the same sign or when the discriminant in (5) is negative. Such cases are described in *Appendix A*.

2.3. *Multiple wavelength anomalous dispersion*

The MAD formalism that is commonly applied to biological macromolecules was initially developed by Karle (1980), and later modified by Hendrickson (1985, 1991; reviewed in Smith, 1991). It can be shown that, if only one type of anomalous scatterer is present,

$$
\begin{aligned}
{}^{\lambda}F_{\mathrm{PH}i}^2 = {}^{0}F_{\mathrm{PH}i}^2 &+ \frac{({}^{\lambda}f_i')^2 + ({}^{\lambda}f_i'')^2}{({}^{0}f_i)^2}\,{}^{0}F_{\mathrm{H}i}^2 \\
&+ \frac{2\,{}^{\lambda}f_i'}{{}^{0}f_i}\,{}^{0}F_{\mathrm{PH}i}\,{}^{0}F_{\mathrm{H}i}\cos({}^{0}\alpha_{\mathrm{PH}i} - {}^{0}\psi_{\mathrm{H}i}) \qquad (9) \\
&\pm \frac{2\,{}^{\lambda}f_i''}{{}^{0}f_i}\,{}^{0}F_{\mathrm{PH}i}\,{}^{0}F_{\mathrm{H}i}\sin({}^{0}\alpha_{\mathrm{PH}i} - {}^{0}\psi_{\mathrm{H}i}),
\end{aligned}
$$

where ${}^{\lambda}F_{\mathrm{PH}i}$ is the experimentally observed structure-factor amplitude of the ith derivative, collected at wavelength λ, including any anomalous diffraction contribution; ${}^{0}F_{\mathrm{PH}i}$ is the amplitude of the normal structure factor of the same derivative (that is, excluding anomalous diffraction effects and therefore independent of the wavelength); ${}^{0}F_{\mathrm{H}i}$ is the amplitude of the normal structure factor from the constellation of anomalous scatterers; and ${}^{0}\alpha_{\mathrm{PH}i}$ and ${}^{0}\psi_{\mathrm{H}i}$ are the phases of their respective structure factors. The $\pm$ sign in the last term is dependent on whether the data for $F(hkl)$ or $F(\bar{h}\bar{k}\bar{l})$ is being considered. In this way, each Friedel mate at each wavelength provides one of a set of simultaneous equations.

This formalism separates the non-anomalous, wavelength independent, quantities from the wavelength-dependent anomalous ones. The only wavelength information in equation (9) is contained in the coefficients ${}^{\lambda}f_i'$ and ${}^{\lambda}f_i''$, which can be obtained with reasonable approximation from theoretical calculations (Cromer & Liberman, 1970), or determined *in situ* experimentally from X-ray fluorescence measurements on the crystal to be diffracted (Hendrickson, Smith, Phizackerley & Merritt, 1988). The ${}^{\lambda}F_{\mathrm{PH}i}$ amplitudes are experimentally measured quantities, and the only unknown parameters are ${}^{0}F_{\mathrm{PH}i}$, ${}^{0}F_{\mathrm{H}i}$ and $({}^{0}\alpha_{\mathrm{PH}i} - {}^{0}\psi_{\mathrm{H}i})$. In general, two Karle equations (9) can be written for each wavelength corresponding to the two Friedel opposites. Hence, error-free data at two different wavelengths should be more than sufficient to solve the equations. In practice, three or four data sets are typically used to overdetermine the problem, and the unknown quantities are derived for every reflection by a least-squares fit to the multiple measurements (Hendrickson, 1985).

The ${}^{0}F_{\mathrm{H}i}$ values derived in this way can be used to locate the anomalous scatterers by calculating a Patterson map with ${}^{0}F_{\mathrm{H}i}^2$ coefficients. Knowing the positions of the anomalous scatterers permits calculation of ${}^{0}\psi_{\mathrm{H}i}$ and then ${}^{0}\alpha_{\mathrm{PH}i}$ for each reflection. The calculated values of ${}^{0}F_{\mathrm{PH}i}$ and ${}^{0}\alpha_{\mathrm{PH}i}$ are then used to compute an electron-density map of the protein plus the non-anomalous contribution of the anomalous scatterers. Equation (9) applies to the case when only one type of anomalous scatterer is present in the crystal. If more scatterer types are present, two additional unknowns are required for each new type, and the number of terms in equation (9) increases to $(n + 1)^2$, where n is the number of anomalous scatterer types (Karle, 1980; Hendrickson, 1985). The Karle equations (9) cannot be applied simultaneously to two or more MAD data sets collected on isomorphous crystals, each with a different anomalous scatterer type. In such cases, the equations must be solved separately for each crystal, and then the phases can be averaged externally.

The treatment described so far is the basis for the algebraic formalism developed by Hendrickson and coworkers in their program *MADLSQ* (Hendrickson, 1985; Hendrickson *et al.*, 1988). An alternative that has become popular in recent years (Ramakrishnan & Biou, 1997), treats multiwavelength data as arising from a conventional multiple isomorphous replacement (MIR) experiment with the inclusion of anomalous scattering (North, 1965; Matthews, 1966). This approach was used by Sweet and coworkers in their determination of the structure of the globular domain of histone H5 by MAD phasing (Ramakrishnan, Finch, Graziano, Lee & Sweet, 1993). In this procedure, diffraction data from one wavelength – $\lambda 1$ for example – are treated as a 'native' data set that has observable Bijvoet differences. Data collected at the other wavelengths are treated as derivatives, where the isomorphous differences arise from the dispersive terms $f_j' - f_1'$ and the Bijvoet differences result from anomalous terms f_j''. Thus, the real and imaginary parts of the atomic structure factors for the jth wavelength are $f_j' - f_l'$ and f_j'', respectively (Ramakrishnan & Biou, 1997).

Heavy-atom refinement and phasing is then carried out with an MIR refinement program, typically *MLPHARE*, which uses a maximum-likelihood algorithm (Otwinowski, 1991). The need for using the data from one wavelength as a native data set imposes limitations on the MIR treatment, in that it does not permit usage of multiple anomalous scattering species in the same crystal, nor multiple wavelength data collected on different crystals. Terwilliger (1994) has proposed a variation of the MIR treatment of dispersive differences that allows for combination of more than one multiwavelength data set. His method involves the approximation that the magnitude of the structure factor corresponding to the anomalous scattering atoms is small compared with that from all other atoms in the structure (Terwilliger, 1994).

The Cullis *et al.* (1961) method described in this paper allows for simultaneous handling of multiple MAD data sets, each from a different constellation of anomalous or isomorphous atoms, in a very straightforward way and with no approximations (Fig. 4). Each Friedel mate from every wavelength and from every derivative can contribute – if the measurement is available – to the phase diagram. The general expression for the derivative

structure factor is

$$
\begin{aligned}
{}^{\lambda}\mathbf{F}_{\mathrm{PH}i} &= \mathbf{F}_{\mathrm{P}} + {}^{\lambda}\mathbf{F}_{\mathrm{H}i} \\
&= \mathbf{F}_{\mathrm{P}} + \sum_{k} \left(1 + \frac{{}^{\lambda}f'_{ki}}{{}^{0}f_{ki}} \pm \frac{i\,{}^{\lambda}f''_{ki}}{{}^{0}f_{ki}}\right) {}^{0}\mathbf{F}_{\mathrm{H}ki},
\end{aligned} \tag{10}
$$

where the i subscript identifies the derivative, λ the wavelength and the double sign the Friedel dependence. ${}^{\lambda}\mathbf{F}_{\mathrm{H}i}$ is the overall heavy-atom structure factor and includes contributions from all the heavy-atom sites, denoted by the k subscript. ${}^{0}\mathbf{F}_{\mathrm{H}ki}$ is the normal diffraction contribution from the kth heavy-atom site, ${}^{\lambda}f'_{ki}$ and ${}^{\lambda}f''_{ki}$ the anomalous atomic scattering factors, and ${}^{0}f_{ki}$ the normal atomic scattering factor. In general, any heavy-atom site is treated as a potential anomalous scatterer site. If ${}^{\lambda}f'_{ki}$ and ${}^{\lambda}f''_{ki}$ are zero, the site corresponds to a normal (non-anomalous), isomorphous scatterer. Different types of anomalous scatterers can be present in the same derivative, or some of the heavy-atom sites may show anomalous scattering whereas others do not. Some of the 'heavy atoms' might be S or Se atoms in the protein if these had been previously located. Every equation (10) corresponds to two circles (one for each Friedel opposite) that can be used in the phase determination as shown in Figs. 1 and 2. MAD, MIR and native data can be readily added to the Argand diagram by calculating the proper ${}^{0}\mathbf{F}_{\mathrm{H}ki}$, ${}^{\lambda}f'_{ki}$ and ${}^{\lambda}f''_{ki}$ values (all of them zero for the measured native data).

${}^{\lambda}F_{\mathrm{P}i}(\alpha)$ are determined by the intersections of the circles obtained from (10) with the radius vector of phase α. Equation (5) can be rewritten as

$$
\begin{aligned}
{}^{\lambda}F_{\mathrm{P}i}(\alpha) = &- {}^{\lambda}F_{\mathrm{H}i}\cos({}^{\lambda}\psi_{\mathrm{H}i} - \alpha) \\
&\pm [{}^{\lambda}F^{2}_{\mathrm{PH}i} - {}^{\lambda}F^{2}_{\mathrm{H}i}\sin^{2}({}^{\lambda}\psi_{\mathrm{H}i} - \alpha)]^{1/2}.
\end{aligned} \tag{11}
$$

Mean amplitudes $\langle F_{\mathrm{P}}(\alpha)\rangle$ are then calculated using (6), individual phase probabilities using (7), and combined phase probabilities using (8). As previously defined, the most probable phases and amplitudes are obtained from the value of α that gives a minimum in the variance of $\langle F_{\mathrm{P}}(\alpha)\rangle$. Under this formulation, there is no need to treat the data from one particular wavelength as 'native', and both amplitude and phase estimates are obtained for an average native structure factor.

This procedure requires the previous knowledge of the positions of the anomalous scatterers in the unit cell, so that the respective ${}^{\lambda}\mathbf{F}_{\mathrm{H}i}$ vectors can be calculated for every derivative, wavelength and hand. Anomalous scatterers can be located from an anomalous Patterson map (Rossmann, 1961), which uses $[F(hkl) - F(\overline{hkl})]^2$ as coefficients. Alternatively, simultaneous Karle equations can be solved for each ith derivative, multiwavelength data set, to calculate ${}^{0}F_{\mathrm{H}i}$ and then to produce an ${}^{0}F^{2}_{\mathrm{H}i}$ Patterson map (see *Appendix B*). Finally, normal isomorphous scatterers have to be identified by conventional Patterson or Fourier isomorphous difference methods.

In most MIR phasing algorithms, centric reflections receive special consideration as their phase is limited to only two possible values (0 or π if the projected center of symmetry is at the origin). Thus, the probability $P(\alpha)$ of α being the phase of F_{P} must be zero for all non-allowed values of α, although, in the presence of anomalous dispersion, the phases of the different $F_{\mathrm{PH}i}$ are not subject to such a restriction.

3. Least-squares refinement of the heavy-atom parameters

The isomorphous or anomalous atomic parameters can be refined by minimizing

$$
\begin{aligned}
\Xi &= \sum_{h}^{H}\sum_{i}^{N} w_{hi}[\langle F_{\mathrm{P}h}(\alpha_{\max})\rangle - F_{\mathrm{P}hi}(\alpha_{\max})]^2 \\
&= \sum_{h}^{H}\sum_{m}^{M}\sum_{\lambda}^{L_m}\sum_{\tau}^{2} w_{hm\lambda\tau}[\langle F_{\mathrm{P}h}(\alpha_{\max})\rangle \\
&\quad - F_{\mathrm{P},hm\lambda\tau}(\alpha_{\max})]^2.
\end{aligned} \tag{12}
$$

The sum is extended to all H unique reflections, represented by the h index, and all N derivatives, represented by the i index. Every derivative can be described in terms of three indices: compound, $m = 1,\ldots, M$; wavelengths of the mth compound, $\lambda = 1,\ldots, L_m$; and Friedel mate, $\tau = 1$ or 2, where a specific value of m, λ and τ defines one circle in the Argand diagram. Every term in (12) is multiplied by a weight $w_{hi} \equiv w_{hm\lambda\tau}$ calculated from the standard measurement error in each reflection. $\langle F_{\mathrm{P}h}(\alpha_{\max})\rangle$ and $F_{\mathrm{P}hi}(\alpha_{\max})$ are, respectively, the mean native protein amplitude and the calculated native protein amplitude for each derivative, both at the most probable value for the protein phase $\alpha_{\max}$. If only two circles are available for a given reflection, then the centroid of the probability-weighted phase distribution, α_{best}, is used instead.

The refinable parameters are the atomic positions for each heavy-atom and anomalous scatterer, their occupancy factor and their temperature factor. If anomalous diffraction is included, then the anomalous scattering factors ${}^{\lambda}f'$ and ${}^{\lambda}f''$ can also be refined for every type of anomalous scatterer. Furthermore, since several data sets need to be scaled together prior to any phase determination, additional refinement of the scaling parameters may be needed.

The 'observed' quantities $\langle F_{\mathrm{P}h}(\alpha_{\max})\rangle$ in (12) are, in fact, dependent on the parameters. Thus, equation (12) can be rewritten as

$$
\Xi = \sum_{h}^{H}\sum_{i}^{N} w_{hi}(0 - \Phi_{hi})^2, \tag{13}
$$

where

$$
\Phi_{hi} = F_{\mathrm{P}hi}(\alpha_{\max}) - \langle F_{\mathrm{P}h}(\alpha_{\max})\rangle. \tag{14}
$$

Equation (13) becomes a normal residual function in which every known value is zero, and the absolute minimum in absence of error will be reached when all terms $(\langle F_{Ph}\rangle - F_{Phi})^2$ are also zero. In practice that does not happen, and a normal least-squares procedure is defined, where the normal equations are,

$$\sum_h^H \sum_i^N \sum_j^J w_{hi} \frac{\partial \Phi_{hi}}{\partial \zeta_j} \frac{\partial \Phi_{hi}}{\partial \zeta_k} \Delta\zeta_k = -\sum_h^H \sum_i^N w_{hi} \Phi_{hi} \frac{\partial \Phi_{hi}}{\partial \zeta_j}. \quad (15)$$

There are as many equations (15) as there are refinable parameters ζ_j ($j = 1,...,J$). The solutions of the system of equations are the parameter shifts $\Delta\zeta_k$. The derivatives of Φ_{hi} can be easily derived from equations (6) and (14) as,

$$\frac{\partial \Phi_{hi}}{\partial \zeta_j} = \frac{\partial F_{hi}}{\partial \zeta_j} - \left\langle \frac{\partial F_h}{\partial \zeta_j} \right\rangle, \quad (16)$$

where

$$\left\langle \frac{\partial F_h}{\partial \zeta_j} \right\rangle = \frac{1}{N_h} \sum_i^N \frac{\partial F_{hi}}{\partial \zeta_j}. \quad (17)$$

N_h is the overall number of derivative circles in which the phase determination of the hth reflection is based, as some of the observations from different derivatives, wavelengths or Friedel opposites may not be usable for that particular reflection.

The Cullis *et al.* (1961) phasing method has been implemented into the heavy-atom phasing and refinement program *HATOMLSQ* (Rossmann, 1967), which has been extensively modified to handle multiwavelength data. The new program, *WHALESQ* (Wavelength-Heavy-Atom-LEast-SQuares) has a least-squares procedure as described above. The program can also apply non-crystallographic symmetry constraints where appropriate to the refinement of the heavy-atom parameters (Rossmann, 1976).

Several numerical tests have been run to verify that the program produces reliable phase information from data with simulated measurement errors and moderate non-isomorphism effects. In all these tests, individual and joined probability distributions have been calculated using equations (7) and (8), where the standard errors E_i have been estimated in a simple way as the root-mean-square of Φ_{hi}, calculated in resolution shells from all the reflections from the same compound and wavelength. Amplitudes and phases for the 'best' electron-density map have been calculated for each reflection at the centroid of the probability distribution,

$$\mathbf{F}_{best} = \langle F_P\rangle \frac{\int_0^{2\pi} P(\alpha)\exp(i\alpha)\mathrm{d}\alpha}{\int_0^{2\pi} P(\alpha)\mathrm{d}\alpha} = \langle F_P\rangle m \exp(i\alpha_{best}), \quad (18)$$

where m is usually referred to as 'figure of merit'.

The program also calculates two additional quantities analogous to those used in lack-of-closure-based MIR methods,

$$R_{m\lambda}(\text{Cullis}) = \left[\frac{\sum_h^H \sum_\tau^2 \Phi^2_{hm\lambda\tau}}{\sum_h^H \sum_\tau^2 (F_{PHhm\lambda\tau} - \langle F_{Ph}\rangle)^2} \right]^{1/2}, \quad (19)$$

and

$$\mathcal{K}_{m\lambda}(\text{Phasing power}) = \left(\frac{\sum_h^H \sum_\tau^2 F^2_{Hhm\lambda\tau}}{\sum_h^H \sum_\tau^2 \Phi^2_{Hhm\lambda\tau}} \right)^{1/2}. \quad (20)$$

Sums in equations (19) and (20) are extended to all the reflections, overall and in resolution shells. Both $R_{m\lambda}$ and $\mathcal{K}_{m\lambda}$ are calculated for every compound and every wavelength used in the phase determination, and can be useful to monitor the quality of the heavy-atom contribution. Typically, bad derivatives show values of $R_{m\lambda}$ around 1.0 and $\mathcal{K}_{m\lambda}$ close to zero.

Despite the simplicity in the error treatment, *WHALESQ* has produced satisfactory results on several numerical tests as described below, with average values of the figures of merit in pretty good agreement with the cosines of the mean phase errors. The phasing method is general enough to allow for several improvements, like more accurate treatments of the propagation of errors into the protein phase angles, as described by Terwilliger & Eisenberg (1987), or implementation of maximum-likelihood algorithms into the refinement of the heavy-atom parameters (Otwinowski, 1991). These and other improvements will be introduced in future versions of the program. The source code for the program is available from the authors.

4. Numerical tests

We have analyzed the performance of the phasing algorithm described above in four numerical tests, representing different experimental situations. The first test, a single MAD experiment, is intended to assess the accuracy of the phases obtained with this method and to compare them with those obtained using widely used phasing programs like *MLPHARE* (Otwinowski, 1991). The second test shows how the presence of two types of anomalous scatterers in the same crystal can be exploited for phasing purposes in a very straightforward way. The third test explores the possibility of using MAD data from two different crystals, each one with a different type of anomalous scatterer, to improve the phases with respect to those obtained from each MAD data set separately. The impact of non-isomorphism between the

Table 1. *Anomalous scattering factors and diffraction ratios*

Diffraction ratios are calculated as follows. Diagonal terms: $\sum_{hkl} |F(hkl) - F(\overline{hkl})| / \sum_{hkl} \langle F(hkl) \rangle$ (centric data in parentheses). Off-diagonal terms: $\sum_{hkl} |^{\lambda i}F(hkl) - {}^{\lambda j}F(hkl)| / \sum_{hkl} \langle F(hkl) \rangle$.

(*a*) Anomalous scattering factors for the Fe *K* absorption edge, and diffraction ratios from error-free data used in the first numerical test (resolution range 25.0–2.5 Å)

	Scattering factors (e^-)		Diffraction ratios (%)				
			λ_j (Å)				
λ_i (Å)	f'	f''	1.8000	1.7402	1.7380	1.6500	1.5000
1.8000	−3.17	0.50	0.3	2.3	1.9	1.0	1.2
1.7402	−8.97	2.48		1.5	1.3	2.6	3.0
1.7380	−6.49	5.19			3.1	1.7	2.2
1.6500	−2.18	3.56				2.1	0.5
1.5000	−0.95	3.05					1.8

(*b*) Diffraction ratios (%) from data with simulated error (resolution range 25.0–3.0 Å)

	λ_j (Å)				
λ_i (Å)	1.8000	1.7402	1.7380	1.6500	1.5000
1.8000	3.8 (3.5)	4.4	4.2	3.9	4.0
1.7402		4.1 (3.5)	4.0	4.5	4.7
1.7380			4.9 (3.4)	4.2	4.3
1.6500				4.4 (3.6)	3.8
1.5000					4.2 (3.3)

two crystals is analyzed there. The fourth test explores the possibility of phasing using multiple data sets originating from different derivatives and wavelengths in a combined MAD and MIR approach.

For all these tests, we have calculated different sets of structure factors from a known protein structure with known heavy-atom positions. We have introduced non-isomorphous errors when combining simulated data from different crystals. To make the tests more realistic, we have simulated measurement errors for the structure-factor amplitudes as a function of resolution (*Appendix C*). In multiple crystal tests, these measurement errors have been introduced on top of the non-isomorphous errors. The protein model is the crystal structure of the homotetrameric hemoglobin from *Urechis caupo* (Kolatkar, Ernst *et al.*, 1992). This hemoglobin contains 141 amino-acid residues plus one heme group per monomer, and the crystal contains two monomers per asymmetric unit. The structure was solved initially to 5 Å using MIR methods (Kolatkar, Meador, Stanfield & Hackert, 1988), and its phases were refined to 2.5 Å using MAD data (Kolatkar, Ernst *et al.*, 1992). Atomic coordinates of the structure are available from the Protein Data Bank (entry code 1ITH).

4.1. *First numerical test: a single MAD experiment*

Atomic coordinates from protein and heme groups were used in the calculation of structure-factor amplitudes. Anomalous dispersion effects were introduced for the Fe atoms at five diffraction wavelengths around the Fe *K* absorption edge at 1.740 Å. The values of f' and f'' at each wavelength were obtained from Hendrickson *et al.* (1988). Table 1(*a*) shows the calculated anomalous signal for the five simulated wavelengths used in this test, assuming an error-free situation. The diffraction ratios for the data with simulated error (Table 1*b*) are larger than the calculated anomalous signal, and in the same range as those observed in the real experiments reported by Hendrickson *et al.* (1988) and Kolatkar, Ernst *et al.* (1992). Diagonal terms in parentheses relate to the differences between Friedel mates of centric reflections as a consequence of the simulated error, and give a measure of the noise level introduced in the data.

Data with simulated error were used in *WHALESQ* to calculate phases and amplitudes for the native structure factors $\mathbf{F}_P$. Each reflection was represented by ten circles, two for each wavelength, with a phase angle interval of 5°. For centric reflections, the structure-factor amplitudes of any two Friedel mates are the same (radii of the circles), but the centers of the circles still are different for each Friedel mate. The *R* factor in Table 2 is an index of the agreement between the mean $\langle F_P \rangle$ of the estimated F_{Pi} amplitudes and the error-free F_P values. In this experiment, the native structure included all protein plus non-metal heme atoms, whereas the two Fe sites were considered heavy atoms. It can be seen that the estimates of F_P are very good. The mean phase error, $\overline{\Delta\alpha}$, was calculated from the differences between the centroid phases and the actual phases of the parent structure factors. Table 2 and Fig. 5 compare the mean phase error with the arc cosine of $\overline{m}$, the mean figure of merit, which should be the expected value of the cosine of the mean phase error. The $\overline{m}$ values – normally available in an

Table 2. *Phasing statistics from the first numerical test, using amplitudes greater than 1σ(F) and within the resolution range 25.0–3.0 Å*

For comparison, statistics are shown from two *MLPHARE* calculations, each using a different wavelength as 'native' data set.

	Overall	Centric	Acentric
WHALESQ			
R factor†	0.030	0.018	0.033
$\overline{\Delta\alpha}$‡	41.2	31.3	42.9
$\cos^{-1}\overline{m}$§	44.9	44.1	45.1
Number of reflections	6622	965	5657
MLPHARE			
λ5 as native			
$\overline{\Delta\alpha}$	44.6	34.2	46.4
$\cos^{-1}\overline{m}$	84.1	79.8	84.8
Number of reflections	6594	953	5641
λ1 as native			
$\overline{\Delta\alpha}$	44.3	37.1	45.5
$\cos^{-1}\overline{m}$	85.8	84.3	86.0
Number of reflections	6591	950	5641

† R factor = $\sum_{hkl}|\langle F_P\rangle - F_P|/\sum_{hkl}F_P$, where F_P is the error-free native structure factor and $\langle F_P\rangle$ is the mean native structure factor as defined in the text. ‡ $\overline{\Delta\alpha}$ is the mean phase error between the estimated phase and its true value. § $\cos^{-1}\overline{m}$ is the cosine of the mean figure of merit.

actual experiment – agree reasonably well with $\overline{\Delta\alpha}$ – available in this test but normally unknown in an actual experiment.

For comparison, phases were also calculated with *MLPHARE* using the MIR analysis of the dispersive differences (Ramakrishnan & Biou, 1997). Data selection criteria were the same for both programs. The data from the remote wavelength (λ5) were used as 'native' in first instance, and then phases were redetermined using the 'pre-edge' wavelength λ1 as native (see Hendrickson *et al.*, 1988, for details about the actual positioning of the different wavelengths in the absorption spectrum curve around the Fe *K* edge). Overall, *MLPHARE* phases between 25 and 3 Å were slightly worse than those calculated from *WHALESQ*, independently of the 'native' wavelength choice (Table 2). The *MLPHARE* figures of merit are seriously underestimated, overall and by resolution ranges (Fig. 5), and seem to depend on which wavelength is used as a 'native'.

The differences in the number of reflections used for phasing statistics in Table 2 arise from the $1\sigma(F)$ cutoff applied. This highlights another potential disadvantage of 'native'-based MAD phasing methods. For a given reflection, the structure-factor amplitude at the wavelength chosen as 'native' can be unobserved or rejected on grounds of σ-based cutoffs. This may result in reflections being inaccurately phased or not phased at all. Instead, *WHALESQ* can reject a poor observation by applying σ-cutoffs, and still determine the phase as accurately as the remaining well determined observations allow. In this numerical test, the total number of unique reflections between 25.0 and 3.0 Å is 6660. Table 2 statistics are based on the reflections surviving the $1\sigma(F)$ cutoff. For *WHALESQ*, those include any reflection for which there are at least any two observations (circles) greater than $1\sigma(F)$. For *MLPHARE*, reflections are included only when 'native' and at least one more wavelength observations are greater than $1\sigma(F)$.

4.2. *Second numerical test: a double MAD experiment with perfect isomorphism*

In this numerical test, two different types of anomalous scatterers are present simultaneously in the same crystal. The heavy-atom derivatives of the hemoglobin crystals (Kolatkar, Ernst *et al.*, 1992) are real-life examples for this case. A selenomethionyl-protein crystal soaked in one heavy-atom solution, or an unmodified protein crystal soaked in two different heavy-atom solutions,

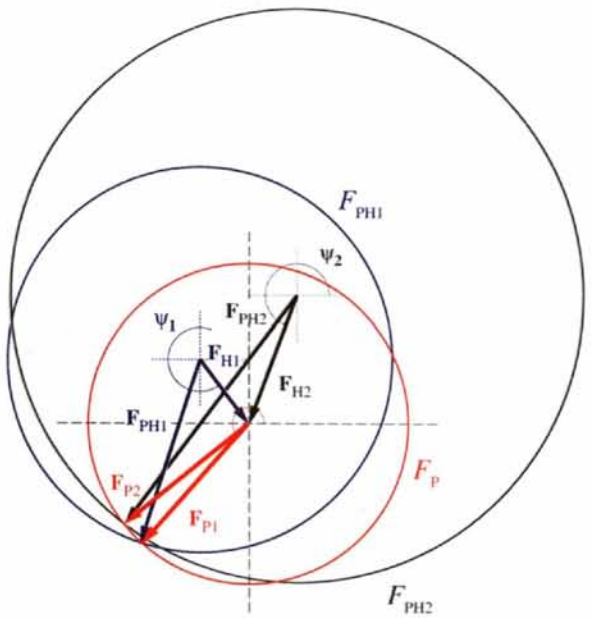

Fig. 1. Argand diagram of the determination of unknown phases from a protein crystal structure using multiple isomorphous derivatives (Harker, 1956). Throughout this paper we will adopt a different convention from what is normally used. In this convention, protein structure factors $\mathbf{F}_P$ (red) are vectors that radiate from the origin of coordinates, with amplitude F_P and phase α; heavy-atom structure factors $\mathbf{F}_{Hi}$ (blue for $i = 1$, black for $i = 2$) are represented by vectors of amplitudes F_{Hi} and phases ψ_i, that start at $(-F_{Hi}\cos\psi_i, -F_{Hi}\sin\psi_i)$ and end at the origin of coordinates; derivative structure factors are represented by vectors beginning at the origin of their respective heavy-atom vectors and finishing at the end of the protein structure factor vectors, such that $\mathbf{F}_{PHi} = \mathbf{F}_{Hi} + \mathbf{F}_P$. Possible solutions for the protein complex structure factor $\mathbf{F}_{Pi}$ are given by the intersections between the circle of radius F_P centered at the origin and the derivative circles of radii F_{PHi} centered at their respective $\mathbf{F}_{Hi}$ origins. In absence of any error, all circles should intersect at only one point, which would give the correct solution for $\mathbf{F}_P$. In practice, every derivative circle may intersect with the native circle in a different position or may not even intersect at all.

would also fit into this category. The experiment simulated here consisted of the measurement of anomalous diffraction data at the absorption edge of each anomalous scatterer. In this simulation, we chose a platinum derivative from Kolatkar, Ernst *et al.* (1992) (third derivative on Table 1 from that reference), and kept the occupancies of the two Pt sites at the levels experimentally determined: 0.14 and 0.17, respectively. Structure-factor amplitudes were calculated at six different wavelengths, three around the Fe K edge and three around the Pt L_{III} edge at 1.072 Å. The f' and f'' values for platinum at the L_{III} edge were estimated from X-ray absorption spectra experimentally measured (unpublished data). Anomalous scattering factors for Pt at the Fe K edge and for Fe at the Pt L_{III} edge were calculated theoretically (Cromer, 1983). The diffraction ratios show that the signal from the Pt L_{III} edge data is quite weak (Table 3).

Phases were determined using *WHALESQ* in a two-derivative approach. The first 'derivative' set consisted of the simulated data for the three wavelengths of the Fe K edge. The second 'derivative' consisted of the simulated data for the three wavelengths of the Pt L_{III} edge. Each 'derivative' contained two Fe and two Pt atoms, and their anomalous scattering factors corresponded to the particular wavelength in use. Thus, a total of 12 circles were used for phase determination. Overall, inclusion of anomalous dispersion data from a low-occupancy secondary anomalous scatterer (Pt) improved the accuracy

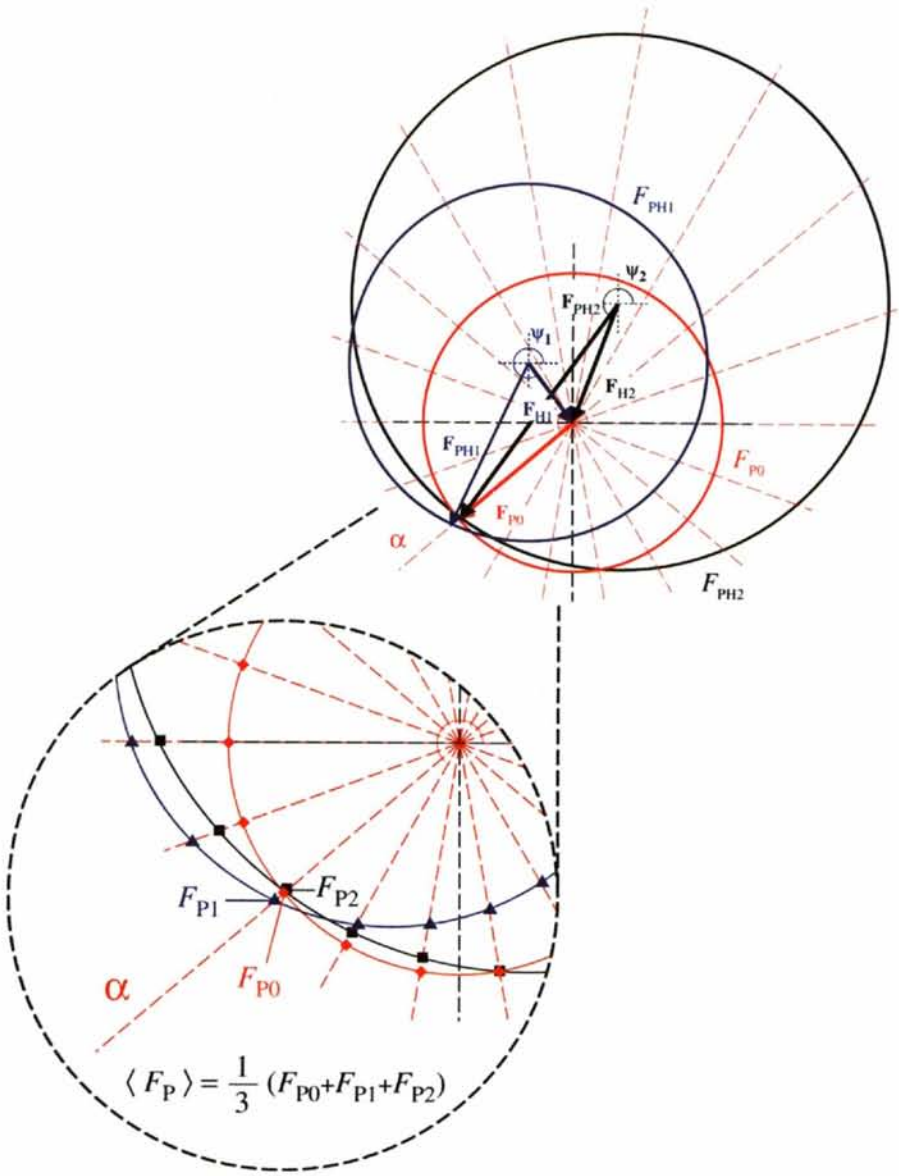

Fig. 2. The protein structure factor $\mathbf{F}_P$ is calculated from each derivative at each sampled value of the phase angle α (every 20° in this example). The most probable phase angle is that which gives the minimum spread (variance) between the calculated amplitudes F_{Pi}. The enlarged area shows the way in which the different F_{Pi} values are calculated at the intersections between the respective circles and straight lines with slope tanα. Different F_{Pi} solutions are shown as triangles, squares or diamonds, respectively. The native circle is treated as another derivative (F_{P0} in this figure) that has a null heavy-atom vector. The best estimate for the protein amplitude is calculated as the average $\langle F_P \rangle$ between all the calculated F_{Pi}. Thus, protein phases and amplitudes can be straightforwardly determined from any three derivative circles, even in the absence of any protein native data. The color scheme for native and derivatives is the same as in Fig. 1.

Table 3. *Anomalous scattering factors and diffraction ratios for the Fe K and Pt L_{III} absorption edge data used in the second numerical test*

Resolution range: 25.0–3.0 Å. Definitions as in Table 1.

	Scattering factors (e^-)				Diffraction ratios (%) λ_j (Å)					
	Fe		Pt		Fe	Fe	Fe	Pt	Pt	Pt
λ_i (Å)	f'	f''	f'	f''	1.7420	1.7380	1.5000	1.0720	1.0714	0.9800
1.7402	−8.97	2.48	−4.77	8.38	4.1 (3.4)	4.0	4.8	5.1	5.0	5.1
1.7380	−6.49	5.19	−4.77	8.38		5.0 (3.6)	4.4	4.8	4.7	4.7
1.5000	−0.95	3.07	−5.39	6.64			4.3 (3.4)	4.1	4.0	3.9
1.0720	0.18	1.75	−25.12	10.20				4.1 (3.5)	3.9	3.9
1.0714	0.18	1.75	−17.56	14.26					4.2 (3.4)	3.9
0.9800	0.25	1.50	−8.26	8.74						4.0 (3.4)

Table 4. *Phasing statistics from the second numerical test*

Phases calculated with *WHALESQ* using amplitudes greater than $1\sigma(F)$ and within the resolution range 25.0–3.0 Å. Definitions as in Table 2.

	Overall	Centric	Acentric
R factor	0.033	0.018	0.036
$\overline{\Delta\alpha}$	33.6	19.9	35.9
$\cos^{-1}\overline{m}$	39.0	32.5	40.0
Number of reflections	6628	970	5658

of the phases obtained from the high-occupancy main anomalous scatterer (Fe) (Table 4).

In practice, it is not uncommon to prepare derivatives of protein crystals already containing one potential anomalous scatterer. Thus, in theory it would be possible to perform real experiments like the one described above and collect anomalous data from the same crystal at two different absorption edges, ensuring perfect isomorphism. However, the arbitrary choice of hand used to define the anomalous scatterer constellation must be reconciled with the absolute (right) hand chosen to index the reflections (Blow & Rossmann, 1961).

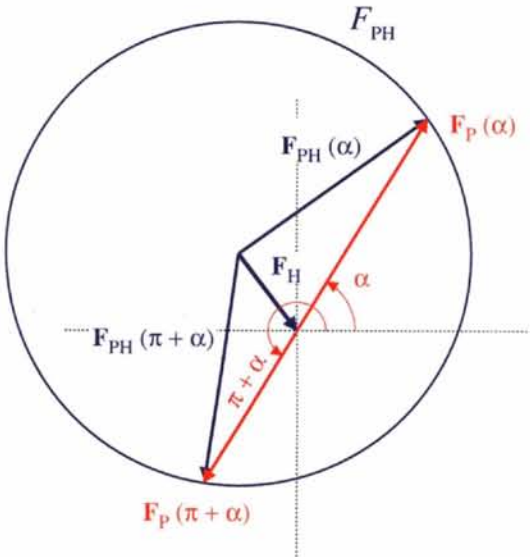

Fig. 3. Origin of the double solution of equation (5) for a particular derivative. The positive solution in this case corresponds to the calculated F_p when the phase angle is α. The negative solution corresponds to minus the calculated F_P when the phase angle is $\pi+\alpha$. As in Fig. 1, native is shown in red.

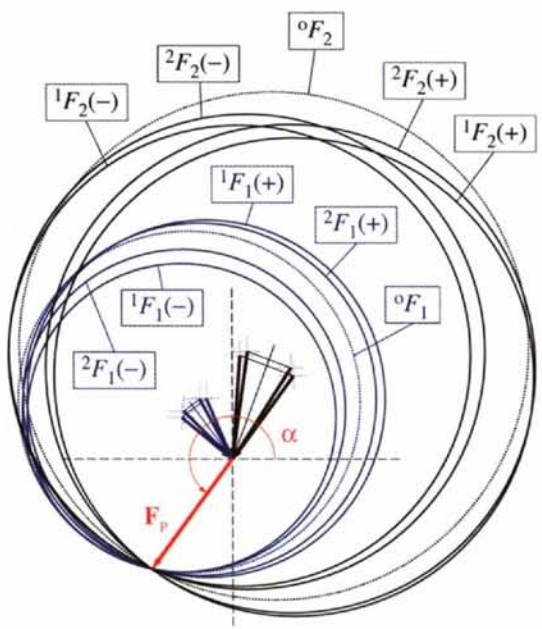

Fig. 4. Treatment of MAD data by the method described in this paper. The size of the anomalous effects has been exaggerated for clarity. The contribution of every kind of anomalous scatterer is calculated independently for every ith compound, λ wavelength and Friedel mate, as if it were an independent derivative. Circles of radii $^\lambda F_i(\pm)$ represent experimentally observed amplitudes. Dashed circles of radii ^0F_i correspond to derivative structure-factor amplitudes in absence of anomalous scattering effects. No native circle is present in this example. The common intersection of all circles gives the phase and amplitude of the protein structure factor, $\mathbf{F}_p$, which includes the contribution of all atoms in the asymmetric unit except the anomalous scatterers. In this way, several MAD data sets can be reduced to the same native protein structure factor. The color scheme is as in Fig. 1.

Table 5. *Anomalous scattering factors and diffraction rations (%) for the Pt L_{III} and Au L_{III} absorption edge data used in the third numerical test*

Resolution range: 25.0–3.5 Å. Definitions as in Table 1.

	Scattering factors (e^-)				Diffraction ratios (%) λ_j (Å)					
	Pt		Au		Pt	Pt	Pt	Au	Au	Au
λ_i (Å)	f'	f''	f'	f''	1.0720	1.0714	0.9800	1.0402	1.0390	0.9500
1.0720	−25.12	10.20			4.0 (3.6)	3.7	4.0	11.7	11.9	12.2
1.0714	−17.57	14.26				4.2 (3.4)	3.8	12.0	12.1	12.4
0.9800	−8.26	8.74					3.9 (3.4)	12.3	12.5	12.8
1.0402			−25.04	10.20				3.8 (3.3)	3.8	3.9
1.0390			−16.43	14.25					4.1 (3.7)	3.8
0.9500			−8.17	8.71						3.9 (3.6)

4.3. *Third numerical test: a two-crystal, two-MAD experiment*

In this test, we explored the possibilities of combining MAD data from two different crystals, each having a different anomalous scattering species. The accuracy of the combined phase determination will depend heavily on the degree of isomorphism between the two crystals. Two heavy-atom derivatives were chosen for this simulation, one containing two Pt sites and the other containing a single Au site (third and fifth derivatives in Table 1 from Kolatkar, Ernst *et al.*, 1992). To enhance the anomalous signal, occupancies of these sites were increased to 0.30, 0.30 and 0.33, respectively. Data were simulated at the Pt L_{III} and Au L_{III} absorption edges, respectively, and Fe atoms were considered 'non-anomalous' in all the calculations. Non-isomorphism between the two derivatives was modeled by changing the unit-cell dimensions of the Au derivative (a was increased by 0.2%, b and c were decreased by 0.2%); by translating each molecule in the reference asymmetric unit by 0.10 Å in x, −0.05 Å in y, −0.10 Å in z; and by rotating each molecule by 0.5° around the z axis. The size of the non-isomorphous errors can be estimated from the R factors between the two crystals (Table 5). Anomalous scattering factors for Pt and Au were theoretically determined or, when possible, estimated from actual X-ray absorption spectra.

The phase improvement obtained by combining two MAD data sets was evaluated by calculating phases with *WHALESQ* using (*a*) data from the Pt derivative alone (six circles), (*b*) data from the Au derivative alone (six circles), and (*c*) data from both derivatives together (12 circles). Phasing statistics for the three cases at different resolution intervals are shown in Table 6. At low resolution, there was a significant improvement in phasing accuracy when combining the Pt and Au data, compared with the phases obtained by using each of them separately. At higher resolution, the overall improvement decreased, and when extended to 3.5 Å resolution, phases calculated using both derivatives were worse than those based on the Pt derivative alone. The decrease of phase improvement with resolution (Table 6 and Fig. 6) is the

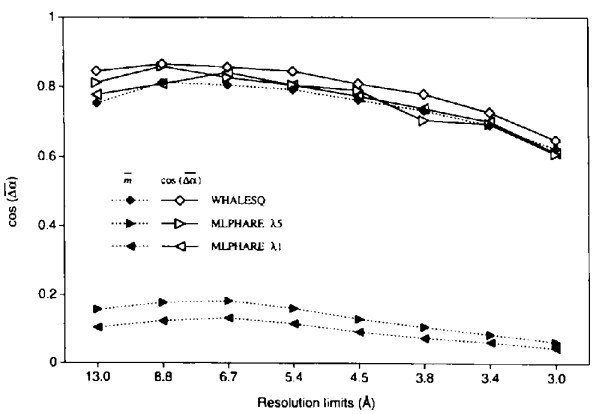

Fig. 5. Variation with resolution of the cosine of the mean phase error, $\cos(\overline{\Delta\alpha})$, and average figure of merit, $\overline{m}$, in the first numerical test. Phases were calculated with *WHALESQ* (diamonds) and *MLPHARE* (triangles). λ5 or λ1 indicate which wavelength was used as native in *MLPHARE* phase determinations.

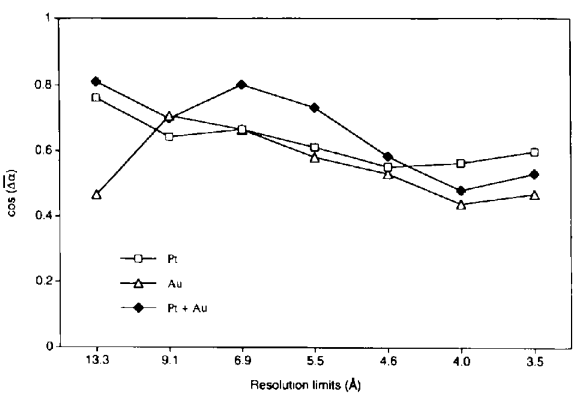

Fig. 6. Variation with resolution of the cosine of the mean phase error, cos $(\overline{\Delta\alpha})$, for the third numerical test, using the Pt L_{III} data alone (squares), the Au L_{III} data alone (triangles), or both Pt L_{III} and Au L_{III} data sets combined (diamonds).

Table 6. *Phasing statistics from the third numerical test*

Phases calculated with *WHALESQ* using amplitudes greater than $1\sigma(F)$ and within the resolution range 25.0–3.5 Å. Definitions as in Table 2.

Resolution (Å)	25.0–5.5			5.5–4.2			4.2–3.5		
	Overall	Centric	Acentric	Overall	Centric	Acentric	Overall	Centric	Acentric
R factor									
Pt only	0.036	0.029	0.038	0.039	0.034	0.040	0.047	0.040	0.048
Au only	0.034	0.029	0.036	0.034	0.034	0.034	0.039	0.039	0.038
Pt + Au	0.041	0.043	0.040	0.062	0.086	0.057	0.074	0.084	0.073
$\overline{\Delta\alpha}$									
Pt only	49.7	45.8	50.9	57.3	56.5	57.4	53.1	47.4	53.9
Au only	51.8	50.3	52.2	60.6	56.9	61.3	62.7	62.5	62.7
Pt + Au	41.2	36.5	42.7	57.4	56.5	57.5	58.8	54.4	59.4
$\cos^{-1}\overline{m}$									
Pt only	50.1	54.0	48.8	56.8	60.8	56.0	54.5	57.9	54.0
Au only	51.2	52.8	50.7	61.4	66.4	60.4	64.2	71.0	63.2
Pt + Au	33.5	25.2	35.9	42.0	32.9	43.6	40.4	34.2	41.2
Number of reflections									
Pt only	1119	271	848	1370	220	1150	1721	207	1514
Au only	1110	265	845	1362	215	1147	1714	206	1508
Pt + Au	1124	276	848	1374	223	1151	1724	210	1514

consequence of the increasing importance of non-isomorphous errors with resolution.

Overall, figures of merit matched reasonably well with the accuracy of the phases when single derivatives were considered for phasing. When phasing with two derivatives, however, figures of merit were overestimated (Table 6). The increase in the number of circles introduces a sharpening in the probability distribution and, hence, improves the calculated m. However, because of non-isomorphous errors, this increase in the number of circles does not correspond with the actual increase in phase accuracy, and as a result the calculated figures of merit are higher than they should be. This discrepancy becomes worse as the resolution of the data increases and the non-isomorphous errors become more important. The F_P amplitudes are still estimated reasonably well, although it seems that combining derivatives results in a small increase in the mean amplitude errors.

Calculation of the phasing power for each derivative as a function of resolution gave some indication of whether the phase accuracy was improved by combination of the two MAD data sets. At a resolution higher than 4.6 Å, the phasing power of the Au data dropped to 1.0–1.2 for the three wavelengths when used in combination with the Pt data, suggesting that their phasing contribution was too small compared to the increasing effect of non-isomorphous errors.

Combining real experimental MAD data from two or more different crystals can be beneficial in those cases where the phases obtained from every MAD experiment independently are poor, so that the impact of non-isomorphism errors is compensated by the phase improvement upon combination. If one of the MAD data sets is able to produce accurate phases on its own, then the introduction of a second MAD data set will only improve

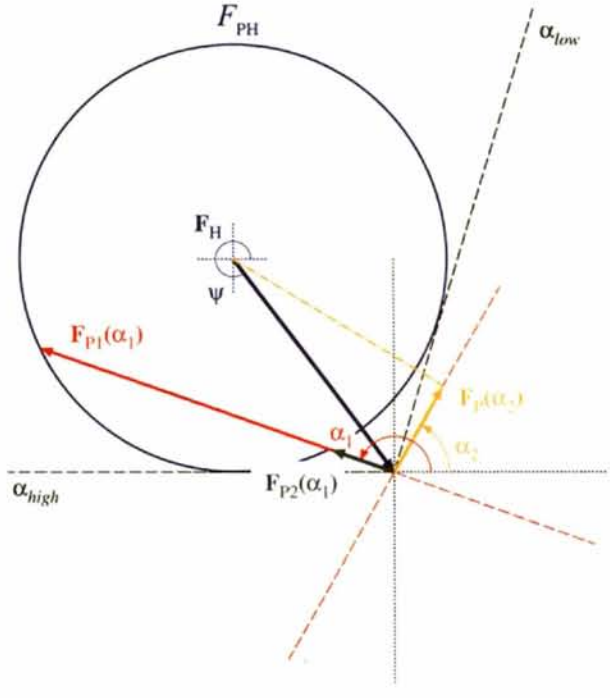

Fig. 7. Examples of special cases when F_H is larger than F_{PH}. At the phase angle α_1, the radius vector intersects the circle of radius F_{PH} (blue) in two points, giving two possible solutions for the protein structure factor: $\mathbf{F}_{P1}(\alpha_1)$ (red) and $\mathbf{F}_{P2}(\alpha_2)$ (green). The same will be true for any α angle between two limiting values α_{low} and α_{high}, defined by the two tangent lines from the coordinate origin to the circle of radius F_{PH}. At the phase angle α_2, there is no intersection between the radius vector and the circle of radius F_{PH}. An approximate solution for the protein structure factor can be obtained as the vector $\mathbf{F}_P(\alpha_2)$ (gold), which is drawn from the origin of coordinates to the point where the radius vector intercepts the perpendicular drawn from the center of the derivative circle.

the phases if there is good isomorphism between the two crystals. Comparison of R(Cullis), phasing power $\mathcal{K}$, and figures of merit at different resolution shells and using various combinations of data sets should generally show the resolution limits of useful combination of phase information.

4.4. *Fourth numerical test: MAD and MIR phase combination*

In this test, we combined MIR and MAD data sets into the same Harker diagram. The simulated experimental data consisted of two collections of structure-factor data sets. The first set contained data from a native crystal and from four low-occupancy derivatives. This set yielded MIR phases of low quality. The second set consisted of MAD data at three wavelengths from a fifth derivative, and was simulated so that the combination of measurement errors and low occupancy of the only anomalous scatterer also produced poor phases when used alone.

All the derivative parameters were taken from Table 1 in Kolatkar, Ernst *et al.* (1992). The MIR data included a native, designated as Nat; two platinum derivatives, Pt(1) and Pt(2); and two mercury derivatives, Hg(1) and Hg(2), with heavy-atom occupancies and coordinates taken from the original reference. The MAD data, Au (3λ), was simulated for the gold derivative at three wavelengths around the Au L_{III} edge. The occupancy, atomic coordinates, wavelengths and anomalous structure factors for this derivative were the same as used in the third numerical test (Table 5). All crystals were modeled with some degree of non-isomorphism, in a similar way to that described for the third numerical test, and then the resulting structure-factor amplitudes were modified with a simulated measurement error. R factors between the different data sets, with or without simulated noise, give an idea of the effect of the different errors (Table 7).

Table 7. *R factors (%) between different derivatives used in the fourth numerical test*

Values in parentheses correspond to noise-free data. Resolution range: 25.0–3.5 Å. $R_{ij} = \sum_{hkl} |F_i(hkl) - F_j(hkl)| / \sum_{hkl} \langle F(hkl) \rangle$.

	Native	Hg (1)	Hg (2)	Pt (1)	Pt (2)
Au (λ2)	10.9	9.9	16.6	9.4	12.0
	(10.5)	(9.1)	(16.3)	(8.6)	(11.3)
Native		16.5	15.2	15.0	7.2
		(16.1)	(14.7)	(14.6)	(6.2)
Hg (1)			21.7	7.7	17.7
			(21.4)	(6.7)	(17.3)
Hg (2)				21.6	15.3
				(21.2)	(14.8)
Pt (1)					15.7
					(15.3)

The MIR phases from the first set of data (native, platinum and mercury derivatives) were calculated using *WHALESQ*, with a mean phase error of 61° for data between 25.0 and 3.5 Å resolution. The overall figure of merit calculated with *WHALESQ* was 0.428, which agreed well with the mean phase error ($\cos^{-1}\overline{m}$) = 65°. For comparison, *MLPHARE* was used to calculate phases from exactly the same data. The mean phase error was 62°, with an overall figure of merit of 0.243 ($\cos^{-1}\overline{m}$ =

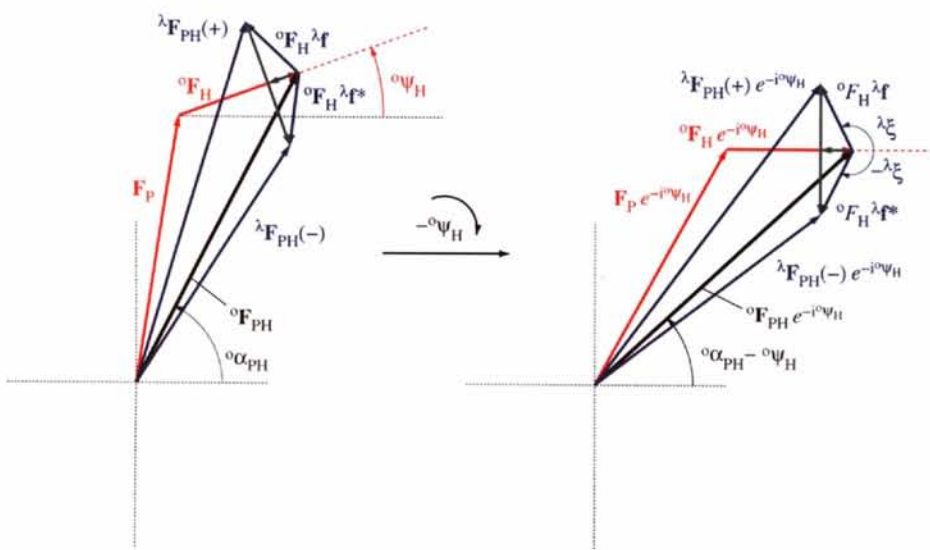

Fig. 8. Geometric relationships between the different complex structure factors resulting from a derivative with anomalous scatterers (left). Vectors in red and black are obtained exclusively from normal diffraction effects and are independent of the wavelength. Vectors in blue are wavelength-dependent and result from anomalous diffraction effects. $^{\lambda}\mathbf{f}$ (green) is the anomalous scattering vector of real component $^{\lambda}f'/^{0}f$ and imaginary component $^{\lambda}f''/^{0}f$ (see text). Rotation of the complex plane by minus the phase of $^{0}\mathbf{F}_{H}$ results in this vector being aligned with the origin of the phase angles (right).

76°). Again, *WHALESQ* phases were on average a little bit better, and the calculated figures of merit were much closer to the cosine of the mean phase error. It appears that *MLPHARE*'s figures of merit are too pessimistic.

WHALESQ phases calculated using only the three Au (3λ) data sets, in a single MAD experiment, were also poor, with a 60° mean phase error. Again, the calculated figure-of-merit calculation, 0.505, agreed with the calculated phase error ($\cos^{-1}\overline{m} = 60°$).

Table 8 shows how the phase accuracy can be increased by combination of different subsets of the data described above. The biggest improvement came from combining all the data in the phase determination [MIR + Au (3λ)], with a mean phase error of 40° and calculated figure of merit of 0.727 ($\cos^{-1}\overline{m} = 43°$). Several intermediate tests were calculated in which the MAD data were combined with different derivatives with various degrees of isomorphism (Table 8). For example, the combination of the three Au (3λ) data sets with just one platinum derivative (one extra circle) produced a clear improvement in the phase accuracy, which was bigger with the more isomorphous Pt(1) data set than with the less isomorphous Pt(2) data set. On the other hand, combination of the MAD Au (3λ) data with the least isomorphous derivative, Hg(2), actually decreased the phase accuracy.

Introduction of the native data into the phase determination, in what would be like a MAD + SIR combination, improved the phase accuracy both in the high-isomorphous case [Aug (3λ), Pt(1) and Nat] and in the low-isomorphous case [Au (3λ), Hg(2) and Nat]. However, in this latest case the MAD Au (3λ) data still did better on their own. Combination of the two least isomorphous derivatives [Pt(2) and Hg(2)] with the Au (3λ) data sets produced better phases than the MAD phases alone, but again the Hg(2) data did more harm than good. Phases were better if only the Pt(2) data were combined with the MAD Au (3λ) data.

Inspection of the phasing power for the different derivatives could help to identify Hg(2) as a potentially bad derivative. $\mathcal{K}$[Pt(1)] for example, ranged from 0.8 to 1.6 on the different tests, $\mathcal{K}$[Pt(2)] was 0.6–0.7, $\mathcal{K}$[Hg(1)] varied from 0.4 to 0.7, and $\mathcal{K}$[Hg(2)] was always below 0.3.

Table 8. *Mean phase error (°) from different combinations of data used in the fourth numerical test*

MIR: all single-wavelength derivatives in Table 7 plus the native; Au (3λ): three-wavelength Au data set; Hg(1), Hg(2), Pt(1) and Pt(2): single derivatives; Nat: native data. Only amplitudes greater than $1\sigma(F)$ have been used. Resolution range: 25.0–3.5 Å. Definitions as in Table 2.

	$\overline{\Delta\alpha}$	$\cos^{-1}\overline{m}$
MIR	60.8	64.7
Au (3λ)	60.4	59.7
MIR + Au (3λ)	40.0	43.4
Au (3λ) + Pt(2)	55.1	51.3
Au (3λ) + Pt(1)	49.4	46.0
Au (3λ) + Pt(1) + Hg(1)	46.5	42.3
Au (3λ) + Pt(1) + Nat	44.3	47.2
Au (3λ) + Hg(2)	63.3	57.0
Au (3λ) + Hg(2) + Nat	61.2	52.5
Au (3λ) + Hg(2) + Pt(2)	58.4	50.9

5. Conclusions

The procedure described here avoids approximations and treats all data equally. This allows for the simultaneous phasing and refinement of MAD data from different anomalous scatterers. Numerical tests verify the procedure and show the results on combining phase information from multiple heavy-atom, anomalously diffracting, isomorphous derivatives. Particularly useful is the possibility of combining MIR and MAD data when the MAD signal is weak and isomorphism is poor. The two indices R(Cullis) and $\mathcal{K}$(phasing power) have been found to be useful to identify bad derivatives or resolution ranges in which poor isomorphism is deleterious rather than helpful. This method is currently being applied in our laboratory to the structure determination of intercellular adhesion molecule-1.

APPENDIX *A*
Special cases

Special cases occur when solving equation (5): the two solutions can be of the same sign, or the discriminant can be negative so that there is no solution at all. These special cases result from the amplitude of the heavy-atom vector, F_{Hi}, being larger than the amplitude of its corresponding derivative structure factor, F_{PHi} (Fig. 7).

If the discriminant is negative, that is $F_{PHi}^2 - F_{Hi}^2 \sin^2(\psi_i - \alpha) < 0$, then the following approximation is used (Cullis *et al.*, 1961),

$$F_{P_i}(\alpha) \simeq -F_{Hi}\cos(\psi_i - \alpha),$$

which is represented in geometrical form in Fig. 7.

Cases with two positive solutions are solved in an iterative way. First, the most positive solution is chosen and an initial $\langle F_P(\alpha)\rangle$ is calculated. Subsequently, the ambiguity between the two solutions is solved by choosing the one closest to the previous $\langle F_P(\alpha)\rangle$, and a new average is calculated. This procedure is repeated until convergence is achieved. If a phase angle α results in an equation (5) with two negative solutions, the angle $\pi + \alpha$ will produce two positive solutions.

APPENDIX *B*
Graphical solution of the Karle equations

The Karle equations can be derived as a special case of the formalism discussed in this paper when there is only

one derivative with anomalous data measured at multiple wavelengths. For simplicity, all anomalous scatterers will be considered from the same type. For a given Friedel pair we have,

$$\begin{aligned} {}^{\lambda}\mathbf{F}_{PH}(+) &= \mathbf{F}_P + {}^{\lambda}\mathbf{F}_H \\ &= \mathbf{F}_P + {}^{0}\mathbf{F}_H\left(1 + \frac{{}^{\lambda}f'}{{}^{0}f} + \frac{i{}^{\lambda}f''}{{}^{0}f}\right) \\ &= {}^{0}\mathbf{F}_{PH} + {}^{0}\mathbf{F}_H\,{}^{\lambda}\mathbf{f}, \end{aligned} \tag{21}$$

and

$$ {}^{\lambda}\mathbf{F}_{PH}(-) = \mathbf{F}_P + {}^{\lambda}\mathbf{F}_H^* = {}^{0}\mathbf{F}_{PH} + {}^{0}\mathbf{F}_H\,{}^{\lambda}\mathbf{f}^*, \tag{22}$$

where ${}^{0}\mathbf{F}_{PH} = \mathbf{F}_P + {}^{0}\mathbf{F}_H$ and ${}^{\lambda}\mathbf{f} = ({}^{\lambda}f'/{}^{0}f) + i({}^{\lambda}f''/{}^{0}f)$, (Fig. 8).

$\mathbf{F}_P$ is the protein structure factor, ${}^{\lambda}\mathbf{F}_H$ is the overall diffraction contribution from the heavy-atom constellation at wavelength λ, ${}^{0}\mathbf{F}_H$ is the normal diffraction contribution from the heavy atoms, ${}^{0}\mathbf{F}_{PH}$ is the normal diffraction contribution from protein and heavy atoms, and ${}^{\lambda}\mathbf{F}_{PH}(\pm)$ are the overall structure factors whose amplitudes are experimentally measured. ${}^{\lambda}\mathbf{F}_H^*$ and ${}^{\lambda}\mathbf{f}^*$ are complex conjugates of ${}^{\lambda}\mathbf{F}_H$ and ${}^{\lambda}\mathbf{f}$.

Now, we rotate the complex plane by multiplying both sides of the equation (21) by $\exp(-i{}^{0}\psi_H)$, where ${}^{0}\psi_H$ is the phase of ${}^{0}\mathbf{F}_H$ (Fig. 8),

$$\begin{aligned} {}^{\lambda}\mathbf{F}_{PH}(+)\exp(-i{}^{0}\psi_H) &= {}^{0}\mathbf{F}_{PH}\exp(-i{}^{0}\psi_H) \\ &\quad + {}^{0}\mathbf{F}_H\,{}^{\lambda}\mathbf{f}\exp(-i{}^{0}\psi_H), \end{aligned}$$

or, in exponential form

$$\begin{aligned} &{}^{\lambda}F_{PH}(+)\exp\{i[{}^{\lambda}\alpha_{PH}(+) - {}^{0}\psi_H]\} \\ &\quad = {}^{0}F_{PH}\exp[i({}^{0}\alpha_{PH} - {}^{0}\psi_H)] \\ &\qquad + {}^{0}F_H\left[\frac{({}^{\lambda}f')^2 + ({}^{\lambda}f'')^2}{({}^{0}f)^2}\right]^{1/2}\exp(i{}^{\lambda}\xi). \end{aligned} \tag{23}$$

${}^{\lambda}\alpha_{PH}(+)$, ${}^{0}\alpha_{PH}$ and ${}^{\lambda}\xi$ are phases of ${}^{\lambda}\mathbf{F}_{PH}(+)$, ${}^{0}\mathbf{F}_{PH}$ and ${}^{\lambda}\mathbf{f}$, respectively. Equation (22) is treated in an identical manner to obtain an equation (23) for the $\overline{hkl}$ measurements, with ${}^{\lambda}F_{PH}(-)$, ${}^{\lambda}\alpha_{PH}(-)$, and $-{}^{\lambda}\xi$ instead of ${}^{\lambda}F_{PH}(+)$, ${}^{\lambda}\alpha_{PH}(+)$, and ${}^{\lambda}\xi$.

From equation (23) it follows that

$$\begin{aligned} {}^{\lambda}F_{PH}^2(\pm) &= {}^{0}F_{PH}^2 + \frac{({}^{\lambda}f')^2 + ({}^{\lambda}f'')^2}{({}^{0}f)^2}{}^{0}F_H^2 \\ &\quad + 2{}^{0}F_{PH}{}^{0}F_H\left[\frac{({}^{\lambda}f')^2 + ({}^{\lambda}f'')^2}{({}^{0}f)^2}\right]^{1/2} \\ &\quad \times \cos(\pm{}^{\lambda}\xi + {}^{0}\psi_H - {}^{0}\alpha_{PH}). \end{aligned} \tag{24}$$

Equation (24) is equivalent to equation (1), but with three unknowns: ${}^{0}F_{PH}$, ${}^{0}F_H$ and $({}^{0}\alpha_{PH} - {}^{0}\psi_H)$. Since there are two equations per wavelength, one for each Friedel pair, two wavelengths are enough in an error-free situation to determine all the parameters. Every observation provides a circle of radius ${}^{\lambda}F_{PH}(\pm)$ with a heavy-atom vector ${}^{0}F_H{}^{\lambda}\mathbf{f}$ or ${}^{0}F_H{}^{\lambda}\mathbf{f}^*$, of variable length and fixed phase. The same scale ${}^{0}F_H$ applies to all heavy-atom vectors. The solution in an Argand diagram is given by the intersection of all circles at the same point. Fig. 9 shows how to solve graphically the simultaneous equations arising from two wavelengths. The phase angle $({}^{0}\alpha_{PH} - {}^{0}\psi_H)$ is sampled from 0 to 360°, and ${}^{0}F_H$ is the scale factor that multiplies all the heavy-atom vectors. Only the correct values for $({}^{0}\alpha_{PH} - {}^{0}\psi_H)$ and ${}^{0}F_H$ bring the four circles to intersect at one point. This intersection gives the correct amplitude ${}^{0}F_{PH}$.

In practice, more than two wavelengths are needed to overcome the experimental errors, and the best estimates of ${}^{0}F_H$, ${}^{0}F_{PH}$ and $({}^{0}\alpha_{PH} - {}^{0}\psi_H)$ are obtained by minimizing the spread of $\langle{}^{0}F_{PH}\rangle$ as described in the main text.

It can be shown that every equation (24) is formally equivalent to a Karle equation (9) by substituting into (24) the angular relation

$$\begin{aligned} &\cos(\pm\xi_\lambda + {}^{0}\psi_H - {}^{0}\alpha_{PH}) \\ &\quad = \cos({}^{0}\alpha_{PH} - {}^{0}\psi_H)\cos(\pm\xi_\lambda) \\ &\qquad + \sin({}^{0}\alpha_{PH} - {}^{0}\psi_H)\sin(\pm\xi_\lambda) \end{aligned}$$

to obtain

$$\begin{aligned} {}^{\lambda}F_{PH}^2(\pm) &= {}^{0}F_{PH}^2 + \frac{({}^{\lambda}f')^2 + ({}^{\lambda}f'')^2}{({}^{0}f)^2}{}^{0}F_H^2 \\ &\quad + \frac{2{}^{\lambda}f'}{{}^{0}f}{}^{0}F_{PH}{}^{0}F_H\cos({}^{0}\alpha_{PH} - {}^{0}\psi_H) \\ &\quad + \frac{2{}^{\lambda}f''}{{}^{0}f}{}^{0}F_{PH}{}^{0}F_H\sin({}^{0}\alpha_{PH} - {}^{0}\psi_H). \end{aligned}$$

APPENDIX C
Error simulation

Errors in the structure-factor amplitudes have been simulated as follows. First, a given structure-factor amplitude is calculated in its exact form F_{calc}. Then, a sample of 100 'observed' structure factors is generated as,

$$F_{obs,i} = (A_iF_{calc} + B_i)(1 + Cs^2) \quad i = 1, \ldots, 100, \tag{25}$$

where

$$A_i = 1 + T_1[\text{ran}(i, 1) - \tfrac{1}{2}],$$

and

$$B_i = T_2[\text{ran}(i, 2) - \tfrac{1}{2}].$$

Measurement errors are simulated *via* A_i, and are proportional to the size of F_{calc}; B_i simulates background fluctuation errors, independent of the size of F_{calc}; T_1 and

T_2 are adjustable target parameters that govern the size of the introduced errors; and ran(i,1) and ran(i,2) are two independent series of randomly generated numbers uniformly distributed between 0 and 1. The $(1 + Cs^2)$ factor in (25) introduces a resolution dependence on the errors, where s is the length of the reciprocal lattice vector and C an adjustable parameter. For the numerical examples used in this paper, the typical values of the three adjustable parameters have been 0.08–0.16 for T_1, 200–300 for T_2 and 5 Å^2 for C.

After the sample of 100 'measurements' is generated, the 'observed' amplitude used in the numerical tests is calculated as the average of a subset of ten observations from the population

$$\langle F_{\rm obs}\rangle = \frac{1}{10}\sum_i F_{{\rm obs},i} \quad i = n, n+10, n+20, \ldots, n+90,$$

where the first value of the series, n, is chosen randomly from 1 to 10. The 'standard deviation' of the 'observed' amplitude is determined from the entire population,

$$\sigma\langle F_{\rm obs}\rangle = \left[\frac{1}{100}\sum_{i=1}^{100}(F_{{\rm obs},i} - F_{\rm calc})^2\right]^{1/2}.$$

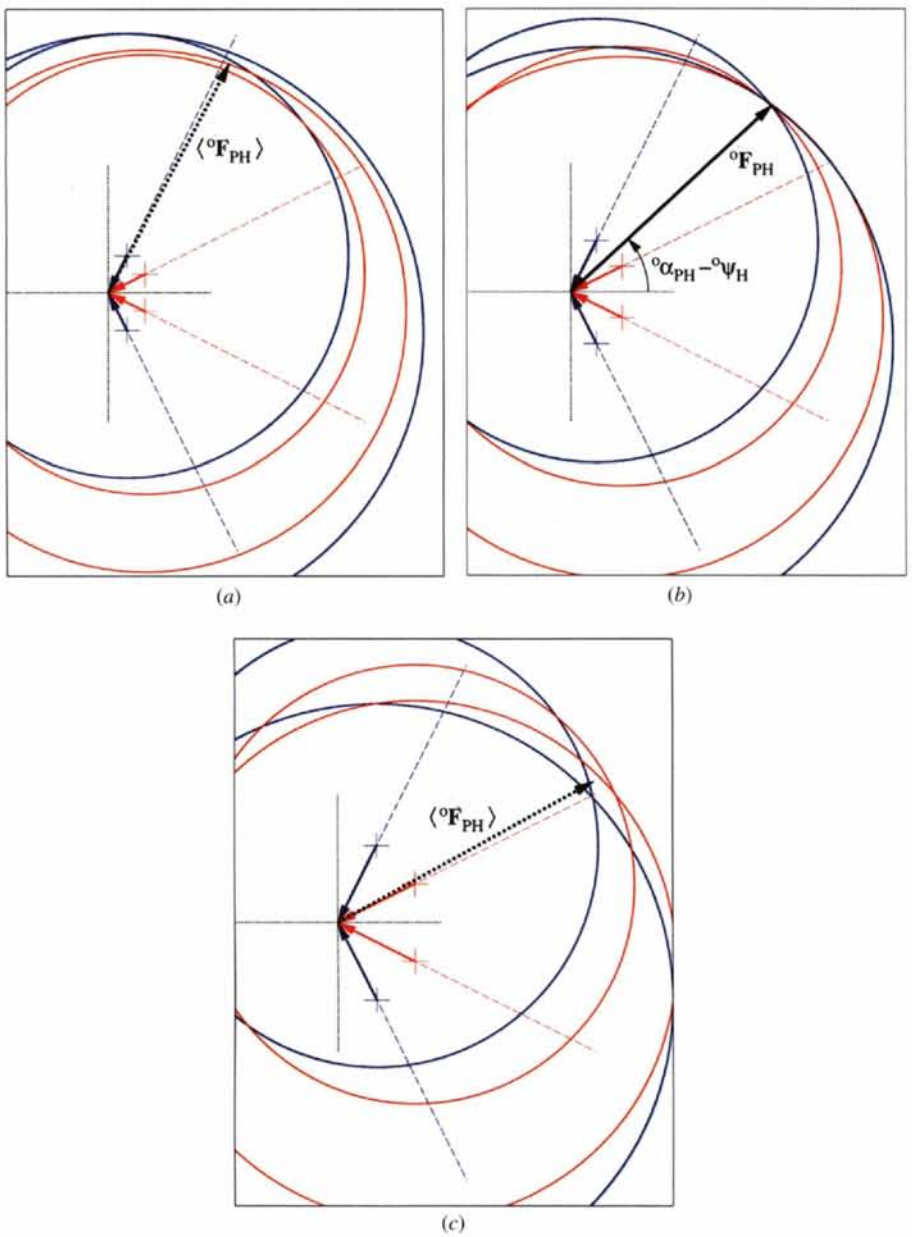

Fig. 9. Solving the Karle equations by Argand diagrams. The heavy-atom vectors are ${}^{0}F_{H}{}^{\lambda}\mathbf{f}$ and ${}^{0}F_{H}{}^{\lambda}\mathbf{f}^{*}$, and are represented as ending at the origin of coordinates (see caption to Fig. 1). The experimentally observed amplitudes are represented as circles of radii ${}^{\lambda}F_{PH}(+)$ and ${}^{\lambda}F_{PH}(-)$ centered at the origins of their respective heavy-atom vectors. Three vignettes illustrate the effect of ${}^{0}F_{H}$ as scaling factor in this two-wavelength example. Heavy-atom vectors in (a) are too short, and every given pair of circles intersects at two different points (the two red circles in fact do not intersect). Still, by evaluating the spread between the different circles at every phase angle, it is possible to find an approximate solution $\langle {}^{0}\mathbf{F}_{PH}\rangle$, as shown by the dashed vector. Heavy-atom vectors in (b) are of the correct size and all four circles intersect in a unique point corresponding to the correct ${}^{0}\mathbf{F}_{PH}$ solution (phase and amplitude). The correct scale of the heavy-atom vectors gives the right value for ${}^{0}F_{H}$. In (c), heavy-atom vectors are too long and there is no longer a unique intersection. The dashed vector indicates the approximate solution obtained at the minimum spread of the four circles. The approximate solutions in both (a) and (c) are close to the correct value, which indicates that ${}^{0}F_{H}$ can be refined if its initial value is not very far from the correct one. The data from the two different wavelengths are shown in red and blue.

We are grateful to Sharon Wilder for proofreading the manuscript. This work was supported by an NIH grant (AI 11219) to Michael G. Rossmann.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195–1202.
Cromer, D. T. (1983). *J. Appl. Cryst.* **16**, 437.
Cromer, D. T. & Liberman, D. (1970). *J. Chem. Phys.* **53**, 1891–1898.
Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1961). *Proc. R. Soc. London Ser. A*, **265**, 15–38.
Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Cryst.* **14**, 1188–1195.
Harker, D. (1956). *Acta Cryst.* **9**, 1–9.
Hendrickson, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Hendrickson, W. A., Smith, J. L., Phizackerley, R. P. & Merritt, E. A. (1988). *Proteins Struct. Funct. Genet.* **4**, 77–88.
Karle, J. (1980). *Int. J. Quantum Chem. Quantum Biol. Symp.* **7**, 357–367.
Kolatkar, P. R., Ernst, S. E., Hackert, M. L., Ogata, C. M., Hendrickson, W. A., Merritt, E. A. & Phizackerley, R. P. (1992). *Acta Cryst.* B**48**, 191–199.
Kolatkar, P. R., Meador, W. E., Stanfield, R. L. & Hackert, M. L. (1988). *J. Biol. Chem.* **263**, 3462–3465.
Kolatkar, P. R., Oliveira, M. A., Rossmann, M. G., Robbins, A. H., Katti, S. K., Hoover-Litty, H., Forte, C., Greve, J. M., McClelland, A. M. & Olson, N. H. (1992). *J. Mol. Biol.* **225**, 1127–1130.
Matthews, B. W. (1966). *Acta Cryst.* **20**, 82–86.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
Otwinowski, Z. (1991). *CCP4 Study Weekend on Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.
Ramakrishnan, V. & Biou, V. (1997). *Methods Enzymol.* **276**, 538–557.
Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. (1993). *Nature (London)*, **362**, 219–223.
Rossmann, M. G. (1961). *Acta Cryst.* **14**, 383–388.
Rossmann, M. G. (1967). Program *HATOMLSQ*, originally written by M. G. Rossmann with subsequent modifications by J. Bolin, G. Kamer & J.-B. Dai, Purdue University, West Lafayette, Indiana, USA.
Rossmann, M. G. (1976). *Acta Cryst.* A**32**, 774–777.
Rossmann, M. G. & Blow, D. M. (1961). *Acta Cryst.* **14**, 641–647.
Smith, J. L. (1991). *Curr. Opin. Struct. Biol.* **1**, 1002–1011.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 17–23.
Terwilliger, T. C. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 6–13.