Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Phasing *via* SAD/MAD data: the method of the joint probability distribution functions

Carmelo Giacovazzo[a,b]* and Dritan Siliqi[b,c]

[a]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Istituto di Cristallografia, CNR, c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and [c]Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana, Albania

Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Received 14 July 2003
Accepted 7 October 2003

The method of the joint probability distribution functions is applied to derive a probabilistic formula which is able to phase reflections in the MAD case accurately, under the condition that the anomalous-scatterer substructure has been defined previously. The mathematical approach takes into account both measurement and model errors, which are treated as primitive random variables, as well as the atomic positions defining the unknown part of the crystal structure. The probabilistic formula has the classical tangent expression. All the parameters influencing the phase estimation are immediately interpretable in terms of experimental quantities: *i.e.* anomalous and dispersive differences, magnitude of the errors and normalized structure-factor moduli. The formula has been applied to several practical cases: a procedure has also been designed which is able to refine the phases and lead to easily interpretable electron-density maps.

1. Notation

N: number of atoms in the unit cell.
a: number of anomalous scatterers in the unit cell.
$na = N - a$: number of non-anomalous scatterers.
$f_j = f_j^0 + \Delta f_j + if_j'' = f_j' + if_j''$: scattering factor of the jth atom. f' is its real and f'' its imaginary part. The thermal factor is included.
Σ_a, Σ_{na}, $\Sigma_N = \sum(f_j'^2 + f_j''^2)$, where the summation is extended to a, na and N atoms.
$F^+ = |F^+| \exp(i\varphi^+) = F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j)$.
$F_a^+ = |F_a^+| \exp(i\varphi_a^+) = \sum_a f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j)$.
$F^- = |F^-| \exp(i\varphi^-) = F_{-\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(-2\pi i \mathbf{h}\mathbf{r}_j)$.
$F_a^- = |F_a^-| \exp(i\varphi_a^-) = \sum_a f_j \exp(-2\pi i \mathbf{h}\mathbf{r}_j)$.
$\Delta_{\rm ano} = |F^+| - |F^-|$.

2. Introduction

The tunability and increased power of modern synchrotron beamlines has made MAD (multiwavelength anomalous dispersion) techniques a very important tool in protein crystallography. Some traditional approaches consider SAD (single-wavelength anomalous dispersion) and MAD data as special SIR (single isomorphous replacement) and MIR (multiple isomorphous replacement) cases (Blow & Crick, 1959; Terwilliger & Eisenberg, 1987). Other approaches apply probabilistic criteria (Pähler *et al.*, 1990; Chiadmi *et al.*, 1993) to the algebraic analysis provided by Karle (1980), which uses the wavelength-dependence of the atomic structure factor of the anomalous scatterers.

More recently, the rigorous method of the joint probability distribution functions has found a wide range of applications when SAD/MAD data are available.

(i) To find the anomalous scatterer substructure (Burla *et al.*, 2002, 2003). The method is able to carefully estimate the structure-factor moduli corresponding to the normal scattering of the anomalous-scatterer substructure, to which Patterson or direct methods may be applied in order to locate the anomalous scatterers (for other techniques, see Blow & Rossmann, 1961; North, 1965; Matthews, 1966; Terwilliger *et al.*, 1987; Miller *et al.*, 1994; Sheldrick & Gould, 1995).

(ii) To phase reflections in the SAD case (Giacovazzo & Siliqi, 2001*a*; referred to as paper I in the following) on the assumption that the anomalous-scatterer substructure has been previously found.

(iii) To phase reflections in the two-wavelength case (Giacovazzo & Siliqi, 2001*b*; referred to as paper II in the following). In spite of the complex mathematical apparatus very simple conclusive formulas were derived, estimating phases in terms of anomalous and of dispersive differences. This paper describes a further step of the method, which is generalized to the *n*-wavelength case. An approximation used in the paper II for the two-wavelength case, particularly rough for small structures, is avoided and a general unbiased probabilistic formula is provided that is valid for any wavelength number. The theoretical results are implemented in an automatic procedure and have been successfully applied to several practical cases.

3. The joint probability distribution $P(F_1^+, \ldots, F_n^+, F_1^-, \ldots, F_n^- | F_{a1}^+, \ldots, F_{an}^+, F_{a1}^-, \ldots, F_{an}^-)$

In accordance with the premises and the results obtained in Appendix *A*, the conditional joint probability distribution $P(E_1^+, \ldots, E_n^+, E_1^-, \ldots, E_n^- | E_{a1}^+, \ldots, E_{an}^+, E_{a1}^-, \ldots, E_{an}^-)$ (in short P) is given by

$$
\begin{aligned}
P \simeq \pi^{-(2n)} q^{-1} \prod_{i=1}^{n} (R_i G_i) \\
\times \exp\Bigg\{ & -\frac{1}{q}\sum_{i=1}^{n}\lambda_{ii}[R_i^2 + R_{ai}^2 - 2R_iR_{ai}\cos(\varphi_i^+ - \varphi_{ai}^+)] \\
& -\frac{1}{q}\sum_{i=n+1}^{2n}\lambda_{ii}[G_i^2 + G_{ai}^2 - 2G_iG_{ai}\cos(\varphi_i^- - \varphi_{ai}^-)] \\
& -\frac{2}{q}\sum_{i,j=1,i<j}^{n}\lambda_{ij}[R_iR_j\cos(\varphi_i^+ - \varphi_j^+) - R_iR_{aj}\cos(\varphi_i^+ - \varphi_{aj}^+) \\
& \quad - R_jR_{ai}\cos(\varphi_j^+ - \varphi_{ai}^+) + R_{ai}R_{aj}\cos(\varphi_{ai}^+ - \varphi_{aj}^+)] \\
& -\frac{2}{q}\sum_{i,j=1,i<j}^{n}\lambda_{n+i,n+j}[G_iG_j\cos(\varphi_i^- - \varphi_j^-) - G_iG_{aj}\cos(\varphi_i^- - \varphi_{aj}^-) \\
& \quad - G_jG_{ai}\cos(\varphi_j^- - \varphi_{ai}^-) + G_{ai}G_{aj}\cos(\varphi_{ai}^- - \varphi_{aj}^-)] \\
& -\frac{2}{q}\sum_{i,j=1}^{n}\lambda_{i,n+j}[R_iG_j\cos(\varphi_i^+ + \varphi_j^-) - R_iG_{aj}\cos(\varphi_i^+ + \varphi_{aj}^-) \\
& \quad - G_jR_{ai}\cos(\varphi_j^- + \varphi_{ai}^+) + R_{ai}G_{aj}\cos(\varphi_{ai}^+ + \varphi_{aj}^-)]\Bigg\}. \qquad (1)
\end{aligned}
$$

Equation (4) in paper II is a particular case (for $n = 2$) of (1).

The coefficients λ_{ij}/q are related to the elements (Λ_{ij}) of the matrix $\mathbf{k}^{-1}$ (see Appendix *A*) by the following relationships:

$$
\begin{aligned}
\frac{\lambda_{ii}}{q} &= \frac{\Lambda_{ii}}{e_i^+} \quad \text{for } i = 1, 2, \ldots, n, \\
\frac{\lambda_{ii}}{q} &= \frac{\Lambda_{ii}}{e_i^-} \quad \text{for } i = n+1, \ldots, 2n, \\
\frac{\lambda_{ij}}{q} &= \frac{\Lambda_{ij}}{(e_i^+ e_j^+)^{1/2}} \quad \text{for } i, j \le n, \\
\frac{\lambda_{ij}}{q} &= \frac{\Lambda_{ij}}{(e_i^+ e_j^-)^{1/2}} \quad \text{for } i \le n, j > n, \\
\frac{\lambda_{ij}}{q} &= \frac{\Lambda_{ij}}{(e_i^- e_j^-)^{1/2}} \quad \text{for } i, j > n.
\end{aligned}
$$

We now give the explicit expressions for the elements λ_{ij} and for the factor q. Denoting

$$\prod = \prod_{p=1}^{n}(\sigma_p^{+2}\sigma_p^{-2}),$$

and

$$S = \sum_{p=1}^{n}(\sigma_p^{+2} + \sigma_p^{-2}),$$

we have

$$
\begin{aligned}
\lambda_{ij} &= -\frac{\prod}{\sigma_i^{+2}\sigma_j^{+2}} \quad \text{for } j \ne i \text{ and } i, j \le n, \\
\lambda_{ij} &= -\frac{\prod}{\sigma_i^{+2}\sigma_{j-n}^{-2}} \quad \text{for } j \ne i \text{ and } i \le n, n < j \le 2n, \\
\lambda_{ij} &= -\frac{\prod}{\sigma_{i-n}^{-2}\sigma_{j-n}^{-2}} \quad \text{for } j \ne i \text{ and } n < i, j \le 2n, \\
\lambda_{ii} &= \frac{\prod}{\sigma_i^{+2}}\left(1 - \frac{1}{\sigma_i^{+2}} + S\right) \quad \text{if } i \le n, \\
\lambda_{ii} &= \frac{\prod}{\sigma_{i-n}^{-2}}\left(1 - \frac{1}{\sigma_{i-n}^{-2}} + S\right) \quad \text{if } n < i \le 2n, \\
q &= \prod(1 + S).
\end{aligned}
$$

It is worthwhile noting that the coefficients λ_{ij}, for $i \ne j$, are always negative, no matter what the value of n is.

Let us now sum the elements of the ith row of the matrix $\boldsymbol{\lambda}$: for a fixed i value,

$$
\begin{aligned}
\sum_{j=1,j\ne i}^{4n}\lambda_{ij} &= -\frac{1}{\sigma_i^{+2}}\prod\left(-\frac{1}{\sigma_i^{+2}} + S\right) \quad \text{for } i \le n \\
\sum_{j=1,j\ne i}^{4n}\lambda_{ij} &= -\frac{1}{\sigma_{i-n}^{-2}}\prod\left(-\frac{1}{\sigma_{i-n}^{-2}} + S\right) \quad \text{for } n < i \le 2n.
\end{aligned}
$$

We can then establish the following relationship:

$$\lambda_{ii} = \frac{1}{\sigma_i^{+2}}\prod - \sum_{j=1,j\ne i}^{4n}\lambda_{ij} \quad \text{for } i \le n, \qquad (2a)$$

$$\lambda_{ii} = \frac{1}{\sigma_{i-n}^{-2}}\prod - \sum_{j=1,j\ne i}^{4n}\lambda_{ij} \quad \text{for } n < i \le 2n. \qquad (2b)$$

It is worthwhile noting that in paper II we introduced the approximation $\lambda_{ii} = -\sum_{j=1}^{4n}\lambda_{ij}$: this was justified by the fact that

$\sigma^2 = \langle|\mu|^2\rangle/\Sigma_{na}$ is usually quite a small quantity. Here the more rigorous equations (2) are used, which will introduce in the conclusive probabilistic formula an additional Sim-like contribution (Sim, 1959*a*,*b*), which was neglected in paper II.

The relation (2) allows us to rewrite P in a more appealing way,

$$\begin{aligned}P \approx (\pi)^{-2n} q^{-1} \prod_{i=1}^{n}(R_i G_i) \\ \times \exp\Bigg[-\frac{1}{q}\prod\Bigg(\sum_{i=1}^{n}\frac{1}{\sigma_i^{+2}}|E_i^+ - E_{ai}^+|^2 + \sum_{i=1}^{n}\frac{1}{\sigma_{i-n}^{-2}}|E_i^- - E_{ai}^-|^2\Bigg) \\ + \frac{1}{q}\sum_{i,j=1,i<j}^{n}\lambda_{ij}|(E_i^+ - E_j^+) - (E_{ai}^+ - E_{aj}^+)|^2 \\ + \frac{1}{q}\sum_{i,j=1,i<j}^{n}\lambda_{n+i,n+j}|(E_i^- - E_j^-) - (E_{ai}^- - E_{aj}^-)|^2 \\ + \frac{1}{q}\sum_{i,j=1}^{n}\lambda_{i,n+j}|(E_i^+ - E_j^{-*}) - (E_{ai}^+ - E_{aj}^{-*})|^2\Bigg], \end{aligned} \tag{3}$$

where E^* represents the complex conjugate of E. The above equation suggests that: (i) the joint probability distribution will attain its maximum value when the squared moduli in the exponential assume their minimum value, which complies perfectly with expectations, (*b*) the coefficients λ_{ij}/q modulate the probability function in accordance with the error distribution and (*c*) the moduli of the structure-factor differences play the role of lack-of-closure criterion.

4. The conditional probability $P(\varphi_1^+, \ldots, \varphi_n^- | \ldots)$

The conditional probability $P(\varphi_1^+, \ldots, \varphi_n^- | \ldots)$ is easily derived from (1) by standard techniques. The use of the relationships (2) leads to

$$\begin{aligned}P(\varphi_1^+, \ldots, \varphi_n^- | \ldots) \approx \\ L^{-1}\exp\Bigg\{ -\frac{2}{q}\sum_{i,j=1,i<j}^{n}\lambda_{ij}[R_i R_j\cos(\varphi_i^+ - \varphi_j^+) \\ - R_i R_{aj}\cos(\varphi_i^+ - \varphi_{aj}^+) - R_j R_{ai}\cos(\varphi_j^+ - \varphi_{ai}^+)] \\ -\frac{2}{q}\sum_{i,j=1,i<j}^{n}\lambda_{n+i,n+j}[G_i G_j\cos(\varphi_i^- - \varphi_j^-) \\ - G_i G_{aj}\cos(\varphi_i^- - \varphi_{aj}^-) - G_j G_{ai}\cos(\varphi_j^- - \varphi_{ai}^-)] \\ -\frac{2}{q}\sum_{i,j=1}^{n}\lambda_{i,n+j}[R_i G_j\cos(\varphi_i^+ + \varphi_j^-) - R_i G_{aj}\cos(\varphi_i^+ + \varphi_{aj}^-) \\ - G_j R_{ai}\cos(\varphi_j^- + \varphi_{ai}^+)] \\ -\frac{2}{q}\sum_{i=1}^{n}\Bigg[-\frac{1}{\sigma_i^{+2}}\prod + \sum_{j=1,j\neq i}^{2n}\lambda_{ij}\Bigg]R_i R_{ai}\cos(\varphi_i^+ - \varphi_{ai}^+) \\ -\frac{2}{q}\sum_{i=1}^{n}\Bigg[-\frac{1}{\sigma_{i-n}^{-2}}\prod + \sum_{j=1,j\neq(n+i)}^{2n}\lambda_{n+i,j}\Bigg]G_i G_{ai}\cos(\varphi_i^- - \varphi_{ai}^-)\Bigg\}. \end{aligned} \tag{4}$$

5. The conditional probability $P(\varphi_1^+ | E_{ai}^+, E_{ai}^-, R_i, G_i, i = 1, \ldots, n)$

In paper II, we explored three different ways of obtaining from $P(\varphi_1^+, \ldots, \varphi_2^- | \ldots)$ a sensible expression for the conditional distribution $P(\varphi_1^+ | E_{ai}^+, E_{ai}^-, R_i, G_i, i = 1, \ldots, 2)$. The most effective way involves the approximation

$$\varphi_1^+ = \varphi_2^+ = -\varphi_1^- = -\varphi_2^- .$$

In accordance with paper II, we will assume

$$\varphi_1^+ = \varphi_2^+ = \ldots = \varphi_n^+ = -\varphi_1^- = \ldots = -\varphi_n^- .$$

(3) then reduces to

$$P(\varphi_1^+ | \ldots) \approx [2\pi I_0(L_1)]^{-1}\exp[L_1\cos(\varphi_1^+ - \theta_1^+)], \tag{5}$$

where

$$\tan\theta_1^+ = \frac{\sum_{j=1}^{n} c_j R_{aj}\sin\varphi_{aj}^+ + \sum_{j=1}^{n} c_{n+j} G_{aj}\sin\varphi_{aj}^{-*}}{\sum_{j=1}^{n} c_j R_{aj}\cos\varphi_{aj}^+ + \sum_{j=1}^{n} c_{n+j} G_{aj}\cos\varphi_{aj}^{-*}} = \frac{T}{B}, \tag{6}$$

$$L_1 = (T^2 + B^2)^{1/2}, \tag{7}$$

$$\varphi_{aj}^{-*} = -\varphi_{aj}^- ,$$

$$c_j = 2\Bigg[\frac{\prod}{\sigma_j^{+2}}R_j + \sum_{p=1,p\neq j}^{n}\lambda_{jp}(R_p - R_j) + \sum_{p=1}^{n}\lambda_{j,n+p}(G_p - R_j)\Bigg]\Bigg/ q$$

if $j < n$, (8*a*)

$$c_j = 2\Bigg[\frac{\prod}{\sigma_{j-n}^{-2}}G_{j-n} + \sum_{p=1}^{n}\lambda_{jp}(R_p - G_{j-n}) + \sum_{p=1,p\neq j-n}^{n}\lambda_{j,p+n}(G_p - G_{j-n})\Bigg]\Bigg/ q$$

for $n < j \leq 2n$. (8*b*)

The reader can easily verify that equations (16)–(19) in paper II are approximated forms (for $n = 2$) of (6)–(8).

The terms

$$\frac{2\prod}{\sigma_j^{+2}}\frac{R_j}{q} \text{ and } \frac{2\prod}{\sigma_{j-n}^{-2}}\frac{G_j}{q},$$

components of the c_j coefficients in (8), were omitted in equations (16)–(19) in paper II owing to the approximations introduced there. In the tangent formula (6) they are multiplied by R_{aj} and G_{aj}, respectively, and constitute the Sim-type contribution.

(5) is a von Mises distribution: θ_1^+ is the most probable value of φ_1^+ and L_1 is its concentration parameter. The value θ_1^+ is defined as a function of the normalized structure factors of the anomalous-scatterer substructure: the terms c_j may be considered as weights, the values of which depend on the observed dispersive and anomalous differences and on the errors at the various wavelengths.

To familiarize the reader with (8), in Appendix *B* we briefly treat the case $n = 3$ as an example. According to this appendix, we can rewrite (8*a*) and (8*b*) in a simplified form,

$$c_j = 2\left\{\frac{\prod}{\sigma_j^{+2}} R_j + \sum_{p=1}^{n}[\lambda_{jp}(R_p - R_j) + \lambda_{j,n+p}(G_p - R_j)]\right\}\Big/ q$$
for $j \le n$, (9a)

$$c_j = 2\left\{\frac{\prod}{\sigma_{j-n}^{-2}} G_j + \sum_{p=1}^{n}[\lambda_{jp}(R_p - G_{j-n}) + \lambda_{j,p+n}(G_p - G_{j-n})]\right\}\Big/ q$$
for $n < j \le 2n$. (9b)

The tangent formula (6) and its concentration parameter L_1 may be rewritten in an alternative form which reveals additional probabilistic properties. Let us introduce (9*a*) and (9*b*) directly into the numerator and denominator of (6). The calculations, briefly described in Appendix *C*, show the following.

(i) The Sim-like contributions to T and B may be written as

$$\frac{2}{q}\prod\sum_{j=1}^{n}\left[\frac{R_j}{\sigma_j^{+2}}\mathrm{Im}(E_{aj}^{+}) + \frac{G_j}{\sigma_j^{-2}}\mathrm{Im}(E_{aj}^{-*})\right]$$

and

$$\frac{2}{q}\prod\sum_{j=1}^{n}\left[\frac{R_j}{\sigma_j^{+2}}\mathrm{Re}(E_{aj}^{+}) + \frac{G_j}{\sigma_j^{-2}}\mathrm{Re}(E_{aj}^{-*})\right],$$

respectively. Accordingly, the Sim-like term tries to drive the value of θ_1^+ towards the phase of the vector

$$\sum_{j=1}^{n}(w_j^+ E_{aj}^+ + w_j^- E_{aj}^{-*}), \tag{10}$$

where

$$w_j^+ = \frac{2}{q}\prod\frac{R_j}{\sigma_j^{+2}} = \frac{2}{1+S}\frac{R_j}{\sigma_j^{+2}},$$

$$w_j^- = \frac{2}{q}\prod\frac{G_j}{\sigma_j^{-2}} = \frac{2}{1+S}\frac{G_j}{\sigma_j^{-2}}.$$

(ii) The contribution to T arising from anomalous and dispersive differences may be written as

$$\frac{2}{q}\Bigg\{\sum_{j,p=1,p>j}^{n}[-\lambda_{jp}(R_j - R_p)\mathrm{Im}(E_{aj}^{+} - E_{ap}^{+}) - \lambda_{n+j,n+p}(G_j - G_p)\mathrm{Im}(E_{aj}^{-*} - E_{ap}^{-*})] - \sum_{j,p=1}^{n}\lambda_{j,n+p}(R_j - G_p)\mathrm{Im}(E_{aj}^{+} - E_{ap}^{-*})\Bigg\}.$$

The corresponding contribution to B may be written as

$$\frac{2}{q}\Bigg\{\sum_{j,p=1,p>j}^{n}[-\lambda_{jp}(R_j - R_p)\mathrm{Re}(E_{aj}^{+} - E_{ap}^{+}) - \lambda_{n+j,n+p}(G_j - G_p)\mathrm{Re}(E_{aj}^{-*} - E_{ap}^{-*})] - \sum_{j,p=1}^{n}\lambda_{j,n+p}(R_j - G_p)\mathrm{Re}(E_{aj}^{+} - E_{ap}^{-*})\Bigg\}.$$

Accordingly, the contribution arising from anomalous and dispersive differences drives the value of θ_1^+ to the phase of the vector

$$\sum_{j,p=1,p>j}^{n}[w_{jp}(E_{aj}^{+} - E_{ap}^{+}) + w_{n+j,n+p}(E_{aj}^{-*} - E_{ap}^{-*})] + \sum_{j,p=1}^{n} w_{j,n+p}(E_{aj}^{+} - E_{ap}^{-*}), \tag{11}$$

where

$$w_{jp} = -\frac{2}{q}\lambda_{jp}(R_j - R_p) = +\frac{2}{(1+S)\sigma_j^{+2}\sigma_p^{+2}}(R_j - R_p),$$

$$w_{n+j,n+p} = -\frac{2}{q}\lambda_{n+j,n+p}(G_j - G_p) = +\frac{2}{(1+S)\sigma_j^{-2}\sigma_p^{-2}}(G_j - G_p),$$

$$w_{j,n+p} = -\frac{2}{q}\lambda_{j,n+p}(R_j - G_p) = +\frac{2}{(1+S)\sigma_j^{+2}\sigma_p^{-2}}(R_j - G_p).$$

Combining (10) and (11) allows us to state the following rule: the most probable phase of φ_1^+, say θ_1^+, is the phase of the vector

$$\sum_{j=1}^{n}(w_j^+ E_{aj}^+ + w_j^- E_{aj}^{-*}) + \sum_{j,p=1,p>j}^{n}[w_{jp}(E_{aj}^{+} - E_{ap}^{+}) + w_{n+j,n+p}(E_{aj}^{-*} - E_{ap}^{-*})] + \sum_{j,p=1}^{n} w_{j,n+p}(E_{aj}^{+} - E_{ap}^{-*}) \tag{12}$$

and the reliability parameter L_1 of the phase estimate is nothing else but its modulus. It may be noted that the larger the number of wavelengths, the larger the number of terms in (10) and (11) and therefore the larger (on the average) the reliability of the phase estimate.

6. The least-squares procedure

Let us now return to analyse (12). While (6)–(9) seem to indicate that the contribution arising from anomalous and dispersive differences depends on the E_{aj}^{+} and on the E_{aj}^{-} values (and therefore on the values of Δf_j and f_j'' at the various wavelengths), the algebraic form of (11) reveals that the most probable value of φ_1^+, say θ_1^+, actually depends on the *differences* $(E_{aj}^{+} - E_{ap}^{+})$, $(E_{aj}^{+} - E_{ap}^{-*})$ and $(E_{aj}^{-*} - E_{ap}^{-*})$. This result justifies the practice (see, for example, Otwinowski, 1991) adopted in our procedure of refining the anomalous-scatterer substructure and the anomalous components of the scattering factors of the anomalously scattering atoms by minimizing the quantities

$$\sum_{H}\sum_{j}[|\Delta_{\mathrm{ano}j}| - K_j(|F_{aj}^{+} - F_{aj}^{-*}|)]^2 \tag{13}$$

and

$$\sum_{H}\sum_{j,p}[|\overline{\Delta}_{\mathrm{disp}j,p}| - K_{j,p}(|\overline{F}_{aj} - \overline{F}_{ap}|)]^2, \tag{14}$$

where j and p denote the wavelengths, K_m and $K_{j,p}$ are suitable scale factors and

$$\overline{\Delta}_{\mathrm{disp}j,p} = \overline{F}_j - \overline{F}_p, \quad \overline{F}_j = \frac{|F_j^{+}| + |F_j^{-}|}{2}, \quad \overline{F}_{aj} = \frac{F_{aj}^{+} + F_{aj}^{-*}}{2}.$$

Table 1
Set of test structures.

PDB is the Protein Data Bank code, SG the space group, NRES is the number of residues, solv is the percentage solvent content, nwl is the number of wavelengths used in the experiment, An. scatt. is the atomic species of the anomalous scatterers (the number of the anomalous scatterers per asymmetric unit is given in pranetheses) and Res is the limiting resolution to which the data are measured. When native data are available, the resolution is quoted in parentheses.

Structure code	PDB	SG	NRES	Solv. (%)	nwl	An. scatt.	Res (Å)	Reference
TTG	1srv	$C222_1$	145	53	4	Se (3)	2.27 (1.7)	Walsh *et al.* (1999)
JIA	1c8u	$C222_1$	570	68	4	Se (8)	1.90	Li *et al.* (2000)
PSCP	1ga1	$P6_2$	372	59	3	Br (13)	1.40	Dauter *et al.* (2001)
CYANASE	1dw9	$P1$	1560	44	4	Se (40)	2.40 (1.65)	Walsh *et al.* (2000)
TGEV	1lvo	$P2_1$	1812	53	4	Se (60)	2.70 (1.95)	Anand *et al.* (2002)
KPR	1ks9	$P4_22_12$	291	43	3	Se (8)	1.70	Silinski *et al.* (2001)
AEPT	1m32	$P2_1$	2196	59	3	Se (66)	2.60 (2.2)	Chen *et al.* (2002)
TM0665	1j6n	$P2_1$	1212	52	3	Se (45)	2.60 (1.8)	Joint Centre for Structural Genomics (to be published)
MDD	1fi4	$P2_22_12$	832	55	3	Se (9)	2.28	Bonanno *et al.* (2001)
IDI	1i9a	$P4_12_12$	364	64	2	Se (8)	2.40	Bonanno *et al.* (2001)
CAUFD	2fdn	$P4_32_12$	55	10	1	Fe (8)	0.94	Dauter *et al.* (1997)
GILU	8xia	$I222$	388	55	1	Mn (1); Mg (1)	1.50	Carrell *et al.* (1989)
HAPTBR	1fj2	$P2_1$	464	51	1	Br (22)	1.80	Betzel *et al.* (1988)
SAV3	1svn	$P2_12_12_1$	269	40	1	Ca (4); S (3); Cl (2)	1.74	Betzel *et al.* (1988)
LYSO2	1l78	$P4_32_12$	258	40	1	S (10); Cl (8)	1.53	Dauter *et al.* (1999)
DOROTA	1ick	$P2_12_12_1$	12	29	1	P (10)	0.95	Dauter & Adamiak (2001)

A simple computer program has been written to implement the approach described above. The program performs least-squares cycles minimizing the quantities (13) and (14) and applies the formula (12) to evaluate the phases. It operates as follows.

(i) The experimental values $|F_j^+|$, $\sigma(|F_j^+|)$, $|F_j^-|$, $\sigma(|F_j^-|)$ are read together with the expected $\Delta f_j'$, f_j'' values for each jth wavelength. If the reflection is centric, we set

$$|F_j^+| = |F_j^-| = (|F_j^+| + |F_j^-|)/2,$$
$$\sigma(|F_j^+|) = \sigma(|F_j^-|) = [\sigma^2(|F_j^+|) + \sigma^2(|F_j^-|)]^{1/2}.$$

(ii) All the diffraction intensities are normalized with respect to Σ_{na}.

(iii) A full-matrix least-squares program is applied: the atomic positional parameters of the anomalous scatterers, their occupancies and thermal factors are considered to be *global* parameters (a unique structural model is refined *via* all the measured intensities); the f'' and the $\Delta f'$ values are treated as *local* parameters (refined *via* the intensities collected at specific wavelengths).

(iv) The global parameters and the f'' values are refined by minimizing the quantity (13). The summation over H includes 70% of the measured reflections (those with the largest values of $\langle|\Delta_{\rm ano}|\rangle$, where the average is taken over all the wavelengths).

(v) The model obtained at step (iv), the occupancies excluded, is kept fixed when the quantity (14) is minimized for defining the differences $\Delta f_j' - \Delta f_p'$. In this case, the summation over H uses only centric reflections, if their number is sufficiently large.

(vi) The refinement is controlled by suitable weights. For step (iv), the weight associated with the intensity of the reflection **h** measured at the wavelength j is the product of two factors: the first is reflection-dependent and the second wavelength-dependent,

$$W_{\rm lsq}(\mathbf{h}, j) = (\sigma^2|F_{hj}^+| + \sigma^2|F_{hj}^-|)^{-1} \cdot {\rm RSD}_j^{-1}.$$

RSD_j is the crystallographic residual obtained in the preceding least-squares cycle for the jth wavelength. It is introduced in the procedure after some cycles of unweighted least squares.

For step (v), the least-squares weight for the reflection **h** corresponding to the intensities measured at the wavelength pair (j, p) is

$$W_{\rm lsq}(\mathbf{h}, j, p) = (\sigma^2|F_{hj}^+| + \sigma^2|F_{hj}^-| + \sigma^2|F_{hp}^+| + \sigma^2|F_{hp}^-|)^{-1/2} \cdot {\rm RSD}_{jp}^{-1},$$

where RSD_{jp} is the crystallographic residual corresponding to pair of wavelengths (j, p).

(vii) After least-squares convergence the formula (12) is calculated. Each term in the tangent expression is additionally weighted by the least-squares residual factors RSD.

The procedure has been written to allow an automatic refinement of the initial model: however, the user can modify the default if necessary.

Table 2
KPR calculated data: errors ($\langle\Delta\varphi\rangle$) of the phase estimates provided by (12).

Values in parentheses are the weighted phase error ($\langle\Delta\varphi\rangle_w$). For the case No. sites = 8, two additional phase errors are given: the first corresponds to a random error in the calculated data up to 10%$|F|$ and the second to a random error up to 50%$|F|$.

No. sites	$\langle\Delta\varphi\rangle$ ($\langle\Delta\varphi\rangle_w$) (°)		
8	12 (12)	28 (19)	64 (58)
7	24 (21)	—	—
6	33 (28)	—	—
5	40 (34)	—	—
4	47 (41)	—	—

Table 3
Test structures.

For each test structure, under the heading SAD/MAD we show the number (NREF) of symmetry-independent reflections phased at the end of the refinement approach described in §6, the corresponding phase error $\langle\Delta\varphi\rangle$ (the weighted phase error is given in parentheses), the correlation factor (CC) between our last electron-density map and the published map and the CPU time necessary to refine the substructure model (CPU). Under the heading FLEX we give the number of reflections phased after the application of the solvent-flattening procedure *FLEX*, the corresponding phase error ($\langle\Delta\varphi\rangle$) (the weighted phase error is given in parentheses) and the correlation factor (CC) between our final electron-density map and the published map.

	SAD/MAD				FLEX		
Structure code	NREF	$\langle\Delta\varphi\rangle$ (°)	CC	CPU	NREF	$\langle\Delta\varphi\rangle$ (°)	CC
TTG	7139	62 (52)	0.62	18	15718	57 (51)	0.73
JIA	32830	57 (46)	0.54	361	74732	40 (30)	0.88
PSCP	40163	69 (59)	0.45	221	87701	44 (38)	0.83
CYANASE	63166	59 (58)	0.48	5677	187107	68 (61)	0.61
TGEV	43043	57 (49)	0.58	5200	146248	67 (63)	0.72
KPR	11536	57 (46)	0.62	45	32249	60 (55)	0.75
AEPT	82026	56 (44)	0.56	7311	82203	50 (41)	0.73
TM0665	73990	47 (35)	0.74	5508	97094	44 (37)	0.82
MDD	21425	64 (56)	0.53	65	22195	60 (57)	0.69
IDI	19391	63 (52)	0.55	50	21332	50 (42)	0.82
CAUFD	19266	42 (30)	0.60	50	29095	44 (34)	0.80
GILU	69780	65 (57)	0.35	31	74882	32 (26)	0.90
HAPTBR	33996	65 (58)	0.49	167	35247	54 (49)	0.68
SAV3	19838	62 (53)	0.42	18	25556	51 (46)	0.68
LYSO2	15132	56 (46)	0.57	54	17923	46 (38)	0.78
DOROTA	3360	48 (42)	0.53	6	16102	44 (34)	0.82

Table 4
Results obtained when (14) is omitted from the refinement procedure.

For defininitions of the headings of the various columns see Table 3.

	SAD/MAD				FLEX		
Structure code	NREF	$\langle\Delta\varphi\rangle$ (°)	CC	CPU	NREF	$\langle\Delta\varphi\rangle$ (°)	CC
TTG	5987	64 (57)	0.57	15	15718	64 (52)	0.66
JIA	32830	58 (48)	0.56	163	74732	42 (36)	0.88
PSCP	40163	70 (62)	0.40	93	87701	44 (38)	0.83
CYANASE	63166	59 (58)	0.48	5677	187107	68 (61)	0.61
TGEV	43043	57 (43)	0.64	3993	146248	65 (61)	0.72
KPR	11536	57 (41)	0.66	35	32249	57 (52)	0.78
AEPT	82026	50 (41)	0.63	6518	82203	48 (40)	0.76
TMO665	73815	47 (36)	0.74	4136	97044	43 (37)	0.84
MDD	21425	63 (53)	0.56	37	22195	56 (49)	0.77
IDI	19391	67 (53)	0.50	35	21332	54 (49)	0.77

7. Experimental tests

To check the correctness of our theoretical results and to evaluate the efficiency of the proposed least-squares approach, we have used experimental data from 16 test structures: they are listed, together with their main crystallochemical data, in Table 1. In this table we specify for each test structure the Protein Data Bank code, the space group, the number of residues, the number of wavelengths used for data collection, the type and the number of anomalous scatterers in the asymmetric unit and the data resolution. For ten structures the data were collected by MAD techniques; the six SAD cases are grouped in the last rows of the Table. In addition, to validate the approximations introduced in our mathematical approach, we have used for the three-wavelength case the calculated data of the structure KPR (32 249 reflections): we simulated cases in which a subset of or all the Se atoms were located (No. sites from 4 to 8). In order to verify the effect of the measurement errors on the efficiency of (12), we have introduced into the calculated data, for the case in which all the eight Se atoms are correctly located, random errors up to 10% $|F|$ and up to 50% $|F|$. The corresponding phase errors are in Table 2: the figures show the robustness of our conclusive formula (12).

Returning back to the observed data, for each test structure an initial substructure model has been provided in accordance with the method recently described by Burla *et al.* (2002, 2003). The least-squares approach described in the §6 of this paper was then applied and the protein phases were estimated *via* the formula (12). The results obtained by complete automation (without any user intervention) are described in Table 3, where under the heading SAD/MAD we quote the number of symmetry-independent reflections phased at the end of the least-squares procedure, the corresponding phase error (in degrees), the correlation factor between our last electron-density map and the published map, the CPU time necessary to perform the phasing procedure (in s for a Dell Precision 830 Pentium V 1.8 MHz). This step involves all the reflections up to the SAD/MAD data resolution (see the last column of Table 1).

The phases thus obtained were automatically submitted to the solvent-flattening procedure *FLEX* (Giacovazzo & Siliqi, 1997). In Table 3 we show for each test structure the number of reflections phased *via FLEX*, the corresponding phase error and the correlation factor between our final electron-density map and the published map. This step involves all reflections to the native data resolution, when available; otherwise, all reflections to the SAD/MAD data resolution are used (see the last column of Table 1).

We observe the following.

(i) In all cases, the combined use of the least-squares procedure and of formula (12) is able to provide phases that are sufficiently accurate to constitute a useful basis for the phase-extension procedure *FLEX*.

(ii) Good electron-density maps are obtained for SAD as well as for MAD data.

Table 5
A comparison between the results obtained by our procedure and corresponding results obtained by other groups using *SHARP* and *DM*.

Structure code	$\langle\Delta\varphi\rangle$ (°), *SHARP*/ our procedure	$\langle\Delta\varphi\rangle$ (°), *DM*/ our procedure	CC, *DM*/ our procedure
PSCP	—/69	—/44	0.76†/0.83
CAUFD	67/42	49/44	0.70/0.80
GILU	66/65	42/32	0.78/0.90
HAPTBR	65/65	49/54	0.76/0.68
SAV3	66/62	55/51	0.70/0.68
LYSO2	58/56	42/46	0.79/0.78
DOROTA	48/48	38/44	0.84/0.82

† Only the correlation is available in the literature; its value is obtained after the application of *SHARP*.

(iii) The presence of different anomalously scattering species does not hinder the success of the procedure.

(iv) The CPU time strongly depends on the number of measured symmetry-independent reflections, as well as on the structural and on the substructural complexity. For most of the test structures the phasing process requires CPU times of the order of tens or of hundreds of seconds. CYANASE, TGEV, AEPT and TMO are the most CPU time-consuming cases (up to 2 h of CPU time): indeed, their native data have relatively high resolution, their structural complexity is high and their substructures are constituted of more than 40 Se atoms.

To understand the role of the dispersive differences in the least-squares approach described in §6, we omitted (14) from the refinement procedure. The results, shown in Table 4, indicate that such an omission does not necessarily decrease the quality of the final electron-density maps (in four cases the map improves), while reducing the overall computing time. This effect may be caused by the minor experimental accuracy of the dispersive differences with respect to the anomalous differences. This conclusion is supported by the following experimental feature: for all the test structures the residuals RSD_{jp} are rarely smaller than and are usually much larger than the residuals RSD_j. When the difference is large, the omission of (14) is beneficial to the quality of the phase estimates. *Vice versa*, when the RSD_{jp} are comparable with or smaller than the the RSD_j, they provide additional information for the phasing process. Our practice of using a weighting scheme depending on the values of RSD_{jp} and RSD_j is a way to take the above considerations into account.

It may be worthwhile to compare our results with the corresponding outcomes obtained by other research groups *via* well documented programs such as *SHARP* (de La Fortelle & Bricogne, 1997), *SOLVE*/*RESOLVE* (Terwilliger & Berenzen, 1999) and *DM* (Cowtan, 1994). We give in Table 5 (for those structures for which the data are available in the literature; see Dauter *et al.*, 2002) the following data.

(i) In column 2, the phase errors obtained by *SHARP* and by our procedure [*i.e.* in a default mode, by the application of (12) followed by the least-squares procedure described in §6].

(ii) In column 3, the results obtained by *DM* (after the application of *SHARP*) and by our *FLEX* program.

Only one data set is available in the literature for a comparison with the *SOLVE*/*RESOLVE* program and concerns TM0665 (González, 2003). The correlation coefficient to the refined model is 0.45 for the experimental map and 0.62 after density modification by *DM*. Our corresponding results are 0.62 and 0.82, respectively.

8. Conclusions

In this paper we have described the following.

(i) A new probabilistic approach able to phase protein reflections when the anomalous-scattering substructure is known. The final formula includes contributions arising from anomalous and dispersive differences and combines them with Sim-like terms. When necessary, such a combination allows us to overcome the phase ambiguity in the SAD case.

(ii) A simple least-squares procedure particularly designed for the automatic refinement of the anomalously scattering substructure model.

(iii) The applications of the above phasing process to 16 test structures, including both SAD and MAD cases. The tests have been made with complete automation and suggest that protein phasing can succeed fully even in the absence of user expertise.

The comparison between our results and corresponding results obtained *via* other well documented programs suggests that the method of the joint probability distribution functions is able to provide powerful phasing formulas, competitive with those derived by different mathematical approaches. Some steps of our procedure are rather weak: *e.g.* the scaling of the experimental data is obtained *via* simple Wilson plots (for more sophisticated approaches, see Blessing & Smith, 1999), the correction for absorption anisotropy (Blessing, 1995) is not applied, the resolution is not taken into account in the weighting scheme adopted for the least-squares step *etc*. It is likely that more robust least-squares procedures (Otwinowski, 1991; de La Fortelle & Bricogne, 1997; Pannu & Read, 1996) would improve the phase estimates further. It is, therefore, very encouraging that our mathematical approach provides, in spite of the weak steps, highly competitive results.

APPENDIX *A*

As in paper II, the positions of the non-anomalous scatterers will be the primitive random variables. The following assumptions are made.

(i) $F^+ = F_a^+ + F_{na}^+ + \mu^+$, where F_{na}^+ is the structure factor corresponding to the non-anomalous scatterers, all supposed non-located. Furthermore, $\mu^+ = |\mu|^+ \exp(i\theta^+)$ represents the cumulative error arising from errors in measurements and in the substructure model of the anomalous scatterers.

(ii) Equivalently, $F^- = F_a^- + F_{na}^- + \mu^-$.

(iii) F_a, F_{na} and μ^+ are uncorrelated with each other.

(iv) $\langle\mu^+\rangle = \langle\mu^-\rangle = 0$.

(v) $\langle\mu_i^+\mu_j^+\rangle = \langle\mu_i^-\mu_j^-\rangle = \langle\mu_i^+\mu_j^-\rangle = 0$ for any pair of wavelengths i, j. This implies that errors in F^+ and F^- are uncorrelated (this is not perfectly true, mostly because μ also contains errors in the model substructure, but the assumption

proved not to be critical and allows us to simplify the calculations).

Accordingly,

$$\langle|F^+|^2\rangle = |F_a^+|^2 + \Sigma_{na} + \langle|\mu^+|^2\rangle,$$
$$\langle|F^-|^2\rangle = |F_a^-|^2 + \Sigma_{na} + \langle|\mu^-|^2\rangle.$$

As in paper II, we will normalize the structure factors with respect to the unknown part of the structure. Accordingly,

$$R\exp(i\varphi^+) = (A^+ + iB^+) = E^+ = F^+/\Sigma_{na}^{1/2},$$
$$G\exp(i\varphi^-) = (A^- + iB^-) = E^- = F^-/\Sigma_{na}^{1/2},$$
$$R_a\exp(i\varphi_a^+) = (A_a^+ + iB_a^+) = E_a^+ = F_a^+/\Sigma_{na}^{1/2},$$
$$G_a\exp(i\varphi_a^-) = (A_a^- + iB_a^-) = E_a^- = F_a^-/\Sigma_{na}^{1/2,}$$

where R, G, R_a and G_a are the pseudo-normalized moduli (with respect to the non-anomalous scatterer substructure) of F^+, F^-, F_a^+ and F_a^-, respectively.

Under the assumptions specified above, we first calculate the characteristic function

$$C(u_1^+, \ldots, u_n^+, u_1^-, \ldots, u_n^-, v_1^+, \ldots, v_n^+, v_1^-, \ldots, v_n^-) \quad (15)$$

of the distribution

$$P(A_1^+, \ldots, A_n^+, A_1^-, \ldots, A_n^-, B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_n^-|A_{a1}^+, \ldots, B_{an}^-), \quad (16)$$

where $u_1^+, \ldots, u_n^+, u_1^-, \ldots, u_n^-, v_1^+, \ldots, v_n^+, v_1^-, \ldots, v_n^-$ are carrying variables associated with $A_1^+, \ldots, A_n^+, A_1^-, \ldots, A_n^-$, $B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_n^-$, respectively. For brevity, we do not specify the expression of (15).

The probability distribution (16) is obtained by Fourier inversion of (15). We have

$$P(A_1^+, \ldots, A_n^+, A_1^-, \ldots, A_n^-, B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_n^-|A_{a1}^+, \ldots, B_{an}^-) = \pi^{-(2n)}\left[\prod_{j=1}^{n}(e_j^+e_j^-)\right]^{-1}(\det\mathbf{k})^{-1/2}\exp\left(-\frac{1}{2}\overline{\mathbf{T}}\mathbf{k}^{-1}\mathbf{T}\right), \quad (17)$$

where

$$\overline{\mathbf{T}} = [(A_1^+ - A_{a1}^+)(2/e_1^+)^{1/2}, \ldots, (A_n^+ - A_{an}^+)/(2/e_n^+)^{1/2}, (A_1^- - A_{a1}^-)/(2/e_1^-)^{1/2}, \ldots, (A_n^- - A_{an}^-)/(2/e_n^-)^{1/2}, (B_1^+ - B_{a1}^+)(2/e_1^+)^{1/2}, \ldots, (B_n^+ - B_{an}^+)/(2/e_n^+)^{1/2}, (B_1^- - B_{a1}^-)/(2/e_1^-)^{1/2}, \ldots, (B_n^- - B_{an}^-)/(2/e_n^-)^{1/2}],$$

$$\mathbf{k} = \begin{vmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{vmatrix},$$

$$\mathbf{Q}_1 = \begin{vmatrix} 1 & \ldots & (e_1^+e_n^+)^{1/2} & (e_1^+e_1^-)^{-1/2} & \ldots & (e_1^+e_n^-)^{-1/2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ (e_n^+e_1^+)^{-1/2} & \ldots & 1 & (e_n^+e_1^-)^{-1/2} & \ldots & (e_n^+e_n^-)^{-1/2} \\ (e_1^-e_1^+)^{-1/2} & \ldots & (e_1^-e_n^+)^{-1/2} & 1 & \ldots & (e_1^-e_n^-)^{-1/2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ (e_n^-e_1^+)^{-1/2} & \ldots & (e_n^-e_n^+)^{-1/2} & (e_n^-e_1^-)^{-1/2} & \ldots & 1 \end{vmatrix},$$

$$\mathbf{Q}_2 = \begin{vmatrix} 1 & \ldots & (e_1^+e_n^+)^{-1/2} & -(e_1^+e_1^-)^{-1/2} & \ldots & -(e_1^+e_n^-)^{-1/2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ (e_n^+e_1^+)^{-1/2} & \ldots & 1 & -(e_n^+e_1^-)^{-1/2} & \ldots & -(e_n^+e_n^-)^{-1/2} \\ -(e_1^-e_1^+)^{-1/2} & \ldots & -(e_1^-e_n^+)^{-1/2} & 1 & \ldots & (e_1^-e_n^-)^{-1/2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -(e_n^-e_1^+)^{-1/2} & \ldots & -(e_n^-e_n^+)^{-1/2} & (e_n^-e_1^-)^{-1/2} & \ldots & 1 \end{vmatrix},$$

$$e_j^+ = 1 + \sigma_j^{+2}, \quad e_j^- = 1 + \sigma_j^{-2},$$

where

$$\sigma_j^{+2} = \langle|\mu_j^+|^2\rangle/\sum_{na}, \quad \sigma_j^{-2} = \langle|\mu_j^-|^2\rangle/\sum_{na}.$$

$\mathbf{Q}_1$ and $\mathbf{Q}_2$ are $(2n) \times (2n)$ matrices. In accordance with paper II,

$$(\det\mathbf{k}) = \left[\prod_{i=1}^{n}\frac{(\sigma_i^{+2}\sigma_i^{-2})}{e_i^+e_i^-}\right]^2\left[1 + \sum_{j=1}^{n}\left(\frac{1}{\sigma_j^{+2}} + \frac{1}{\sigma_j^{-2}}\right)\right]^2.$$

The element Λ_{ij} of the matrix $\mathbf{k}^{-1}$ may be obtained by observing that

$$\mathbf{k}^{-1} = \begin{vmatrix} \mathbf{Q}_1^{-1} & 0 \\ 0 & \mathbf{Q}_2^{-1} \end{vmatrix}.$$

The change of variables

$$A_i^+ = R_i\cos\varphi_i^+, \quad A_i^- = G_i\cos\varphi_i^-,$$
$$A_{ai}^+ = R_{ai}\cos\varphi_{ai}^+, \quad A_{ai}^- = G_{ai}\cos\varphi_{ai}^-$$

leads to expression (1) of the main text.

APPENDIX *B*

As an example, let us specify the expressions of the coefficient c_j defined by the equations (8) for $n = 3$. The extension to cases $n = 4, 5, \ldots$ is trivial. We do not consider the Sim-type contributions. We have

$$c_1 = 2[\lambda_{12}(R_2 - R_1) + \lambda_{13}(R_3 - R_1) + \lambda_{14}(G_1 - R_1) + \lambda_{15}(G_2 - R_1) + \lambda_{16}(G_3 - R_1)]/q,$$
$$c_2 = 2[\lambda_{21}(R_1 - R_2) + \lambda_{23}(R_3 - R_2) + \lambda_{24}(G_1 - R_2) + \lambda_{25}(G_2 - R_2) + \lambda_{26}(G_3 - R_2)]/q,$$
$$c_3 = 2[\lambda_{31}(R_1 - R_3) + \lambda_{32}(R_2 - R_3) + \lambda_{34}(G_1 - R_3) + \lambda_{35}(G_2 - R_3) + \lambda_{36}(G_3 - R_3)]/q,$$
$$c_4 = 2[\lambda_{41}(R_1 - G_1) + \lambda_{42}(R_2 - G_1) + \lambda_{43}(R_3 - G_1) + \lambda_{45}(G_2 - G_1) + \lambda_{46}(G_3 - G_1)]/q,$$
$$c_5 = 2[\lambda_{51}(R_1 - G_2) + \lambda_{52}(R_2 - G_2) + \lambda_{53}(R_3 - G_2) + \lambda_{54}(G_1 - G_2) + \lambda_{56}(G_3 - G_2)]/q,$$
$$c_6 = 2[\lambda_{61}(R_1 - G_3) + \lambda_{62}(R_2 - G_3) + \lambda_{63}(R_3 - G_3) + \lambda_{64}(G_1 - G_3) + \lambda_{65}(G_2 - G_3]/q.$$

Each c_j coefficient is therefore the sum of the elements of the jth line of the skew-symmetric matrix

$$\begin{vmatrix} 0 & \lambda_{12}(R_2-R_1) & \dots & \lambda_{1n}(R_n-R_1) & \lambda_{1,n+1}(G_1-R_1) & \dots & \lambda_{1,2n}(G_n-R_1) \\ \lambda_{21}(R_1-R_2) & 0 & \dots & \lambda_{2n}(R_n-R_2) & \lambda_{2,n+1}(G_1-R_2) & \dots & \lambda_{2,2n}(G_n-R_2) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_{n1}(R_1-R_n) & \lambda_{n2}(R_2-R_n) & \dots & 0 & \lambda_{n,n+1}(G_1-R_n) & \dots & \lambda_{n,2n}(G_n-R_n) \\ \lambda_{n+1,1}(R_1-G_1) & \dots & & \lambda_{n+1,n}(R_n-G_1) & 0 & & \lambda_{n+1,2n}(G_n-G_1) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_{2n,1}(R_1-G_n) & \dots & & & \lambda_{2n,n}(R_n-G_n) & \lambda_{2n,n+1}(G_1-G_n) & 0 \end{vmatrix}.$$

Since the diagonal elements of the matrix vanish, we can rewrite (8*a*) and (8*b*) as the more simple formula

$$c_j = \sum_{j=1}^{n}[\lambda_{jp}(R_p - R_j) + \lambda_{j,n+p}(G_p - R_j)]/q, \qquad j \le n,$$

$$c_j = \sum_{j=1}^{n}[\lambda_{jp}(R_p - G_{j-n}) + \lambda_{j,p+n}(G_p - G_{j-n})]/q, \quad n < j \le 2n.$$

APPENDIX *C*

We introduce into (6) the c_j expressions given by (9). We examine two cases (the others can be obtained by simple generalization).

C1. One-wavelength case

The T term in (6) may be rewritten as

$$\begin{aligned} T &= \frac{2}{q}\left\{\left[\frac{\prod}{\sigma_1^{+2}}R_1 + \lambda_{12}(G_1-R_1)\right]R_{a1}\sin\varphi_{a1}^{+}\right. \\ &\quad \left. - \left[\frac{\prod}{\sigma_1^{-2}}G_1 + \lambda_{12}(R_1-G_1)\right]G_{a1}\sin\varphi_{a1}^{-}\right\} \\ &= \frac{2}{q}\left\{\prod\left[\frac{1}{\sigma_1^{+2}}R_1R_{a1}\sin\varphi_{a1}^{+} + \frac{1}{\sigma_1^{-2}}G_1G_{a1}\sin\varphi_{a1}^{-*}\right]\right. \\ &\quad \left. + \lambda_{12}(G_1-R_1)\left[R_{a1}\sin\varphi_{a1}^{+} - G_{a1}\sin\varphi_{a1}^{-*}\right]\right\} \\ &= \frac{2}{q}\left\{\prod\left[\frac{R_1}{\sigma_1^{+2}}\mathrm{Im}(E_{a1}^{+}) + \frac{G_1}{\sigma_1^{-2}}\mathrm{Im}(E_{a1}^{-*})\right]\right. \\ &\quad \left. + \lambda_{12}(G_1-R_1)\mathrm{Im}(E_{a1}^{+} - E_{a1}^{-*})\right\}, \end{aligned} \tag{18}$$

where $\mathrm{Im}(x)$ stands for 'imaginary part of x'.

The B term in (6) may be rewritten as

$$\begin{aligned} B &= \frac{2}{q}\left\{\left[\frac{\prod}{\sigma_1^{+2}}R_1 + \lambda_{12}(G_1-R_1)\right]R_{a1}\cos\varphi_{a1}^{+}\right. \\ &\quad \left. + \left[\frac{\prod}{\sigma_1^{-2}}G_1 + \lambda_{12}(R_1-G_1)\right]G_{a1}\cos\varphi_{a1}^{-}\right\} \\ &= \frac{2}{q}\left\{\prod\left[\frac{1}{\sigma_1^{+2}}R_1R_{a1}\cos\varphi_{a1}^{+} + \frac{1}{\sigma_1^{-2}}G_1G_{a1}\cos\varphi_{a1}^{-*}\right]\right. \\ &\quad \left. + \lambda_{12}(G_1-R_1)\left[R_{a1}\cos\varphi_{a1}^{+} - G_{a1}\cos\varphi_{a1}^{-*}\right]\right\} \\ &= \frac{2}{q}\left\{\prod\left[\frac{R_1}{\sigma_1^{+2}}\mathrm{Re}(E_{a1}^{+}) + \frac{G_1}{\sigma_1^{-2}}\mathrm{Re}(E_{a1}^{-*})\right]\right. \\ &\quad \left. + \lambda_{12}(G_1-R_1)\mathrm{Re}(E_{a1}^{+} - E_{a1}^{-*})\right\}, \end{aligned} \tag{19}$$

where $\mathrm{Re}(x)$ stands for 'real part of x'.

Combining (18) and (19) into (6) gives the formula recently proposed by Giacovazzo *et al.* (2003) for the SAD case [see equation (4) in that paper].

C2. Two-wavelength case

The T term in (6) may be rewritten as

$$\begin{aligned} &\frac{2}{q}\left\{\left[\frac{\prod}{\sigma_1^{+2}}R_1 + \lambda_{12}(R_2-R_1)\right.\right. \\ &\qquad \left.\left. + \lambda_{1,3}(G_1-R_1) + \lambda_{14}(G_2-R_1)\right]R_{a1}\sin\varphi_{a1}^{+}\right. \\ &+\left[\frac{\prod}{\sigma_2^{+2}}R_2 + \lambda_{21}(R_1-R_2)\right. \\ &\qquad \left. + \lambda_{23}(G_1-G_2) + \lambda_{24}(G_2-R_2)\right]R_{a2}\sin\varphi_{a2}^{+} \\ &-\left[\frac{\prod}{\sigma_1^{-2}}G_1 + \lambda_{31}(R_1-G_1)\right. \\ &\qquad \left. + \lambda_{32}(R_2-G_1) + \lambda_{34}(G_2-G_1)\right]G_{a1}\sin\varphi_{a1}^{-} \\ &-\left[\frac{\prod}{\sigma_2^{-2}}G_2 + \lambda_{41}(R_1-G_2)\right. \\ &\qquad \left. + \lambda_{42}(R_2-G_2) + \lambda_{43}(G_1-G_2)\right]G_{a2}\sin\varphi_{a2}^{-}\Big\} \\ &= \frac{2}{q}\left\{\prod\left[\frac{1}{\sigma_1^{+2}}R_1\mathrm{Im}(E_{a1}^{+}) + \frac{1}{\sigma_2^{+2}}R_2\mathrm{Im}(E_{a2}^{+})\right.\right. \\ &\qquad \left.\left. + \frac{1}{\sigma_1^{-2}}G_1\mathrm{Im}(E_{a1}^{-*}) + \frac{1}{\sigma_2^{-2}}G_2\mathrm{Im}(E_{a2}^{-*})\right]\right. \\ &- \lambda_{12}(R_1-R_2)\mathrm{Im}(E_{a1}^{+} - E_{a2}^{+}) - \lambda_{13}(R_1-G_1)\mathrm{Im}(E_{a1}^{+} - E_{a1}^{-*}) \\ &- \lambda_{14}(R_1-G_2)\mathrm{Im}(E_{a1}^{+} - E_{a2}^{-*}) - \lambda_{23}(R_2-G_1)\mathrm{Im}(E_{a2}^{+} - E_{a1}^{-*}) \\ &\left. - \lambda_{24}(R_2-G_2)\mathrm{Im}(E_{a2}^{+} - E_{a2}^{-*}) - \lambda_{34}(G_1-G_2)\mathrm{Im}(E_{a1}^{-*} - E_{a2}^{-*})\right\} \end{aligned} \tag{20}$$

Accordingly, the B term in (6) may be rewritten as

$$\begin{aligned} &\frac{2}{q}\left\{\prod\left[\frac{1}{\sigma_1^{+2}}R_1\mathrm{Re}(E_{a1}^{+}) + \frac{1}{\sigma_2^{+2}}R_2\mathrm{Re}(E_{a2}^{+})\right.\right. \\ &\left. + \frac{1}{\sigma_1^{-2}}G_1\mathrm{Re}(E_{a1}^{-*}) + \frac{1}{\sigma_2^{-2}}G_2\mathrm{Re}(E_{a2}^{-*})\right] \\ &- \lambda_{12}(R_1-R_2)\mathrm{Re}(E_{a1}^{+} - E_{a2}^{+}) - \lambda_{13}(R_1-G_1)\mathrm{Re}(E_{a1}^{+} - E_{a1}^{-*}) \\ &- \lambda_{14}(R_1-G_2)\mathrm{Re}(E_{a1}^{+} - E_{a2}^{-*}) - \lambda_{23}(R_2-G_1)\mathrm{Re}(E_{a2}^{+} - E_{a1}^{-*}) \\ &\left. - \lambda_{24}(R_2-G_2)\mathrm{Re}(E_{a2}^{+} - E_{a2}^{-*}) - \lambda_{34}(G_1-G_2)\mathrm{Re}(E_{a1}^{-*} - E_{a2}^{-*})\right\}. \end{aligned} \tag{21}$$

The expressions (20) and (21) reveal the role of the Sim-like terms and the form of the contribution arising from anomalous and dispersive differences.

The above algebraic expressions may be easily generalized to the n-wavelength case for any value of n.

We are indebted to S. K. Burley, Z. Dauter, K. Djinovic, R. Hilgenfeld, A. González, D. Matak, M. Walsh and C. Weeks who kindly provided us with experimental data.

References

Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J. & Hilgenfeld, R. (2002). *EMBO J.* **21**, 3213–3224.

Betzel, C., Dauter, Z., Dauter, M., Ingelman, M., Papendorf, G., Wilson, K. S. & Branner, S. (1988). *J. Mol. Biol.* **204**, 803–804.

Blessing, R. H. (1995). *Acta Cryst.* A**51**, 33–38.

Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.

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.

Bonanno, J. B., Edo, C., Eswar, N., Pieper, U., Romanowski, M. J., Ilyin, V., Gerchman, S. E., Kycia, H., Studier, F. M., Sali, A. & Burley, S. K. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 12896–12901.

Burla, M. C., Carrozzini, D., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* D**59**, 662–669.

Burla, M. C., Carrozzini, D., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.

Carrell, H. L., Glusker, J. P., Burger, V., Manfre, F., Tritsch, D. & Biellmann, J. F. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 4440–4444.

Chen, C. C. H., Zhang, H., Kim, A. D., Howard, A., Sheldrick, G. M., Mariano-Dunnaway, D. & Herzberg, O. (2002). *Biochemistry*, **41**, 13162.

Chiadmi, M., Kahn, R., de La Fortelle, E. & Fourme, R. (1993). *Acta Cryst.* D**49**, 522–529.

Cowtan, K. (1994). Jnt CCP4 ESF-EACBM Newslett. Protein Crystallogr. **31**, 34–38.

Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.

Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.

Dauter, Z., Li, M. & Wlodawer, A. (2001). *Acta Cryst.* D**57**, 239–249.

Dauter, Z., Wilson, K. S., Sieker, L. C., Meyer, J. & Moulis, J.-M. (1997). *Biochemistry*, **36**, 14493–14502.

Giacovazzo, C., Ladisa, M. & Siliqi, D. (2003). *Acta Cryst.* A**59**, 262–265.

Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.

Giacovazzo, C. & Siliqi, D. (2001*a*). *Acta Cryst.* A**57**, 40–46.

Giacovazzo, C. & Siliqi, D. (2001*b*). *Acta Cryst.* A**57**, 700–707.

González, A. (2003). *Acta Cryst.* D**59**, 315–322.

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

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Li, J., Derewenda, U., Dauter, Z., Smith, S. & Derewenda, Z. S. (2000). *Nature Struct. Biol.* **7**, 555–559.

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

Miller, R., Gallo, S., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.

North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.

Otwinowski, Z. (1991). *Proceedings of the Daresbury Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Pähler, A., Smith, J. L. & Hendrickson, W. A. (1990). *Acta Cryst.* A**46**, 537–540.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Sheldrick, G. M. & Gould, R. G. (1995). *Acta Cryst.* B**51**, 423–431.

Silinski, P., Allingham, M. J. & Fitzgerald, M. C. (2001). *Biochemistry*, **40**, 14493–14502.

Sim, G. A. (1959*a*). *Acta Cryst.* **12**, 813–815.

Sim, G. A. (1959*b*). *Acta Cryst.* **13**, 511–512.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.

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

Terwilliger, T. C., Kim, S.-H. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 1–5.

Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.

Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure*, **8**, 505–514.