Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 26 February 2001
Accepted 9 May 2001

On integrating direct methods and isomorphous-replacement techniques: triplet estimation and treatment of errors

Carmelo Giacovazzo,[a]* Dritan Siliqi[a,b] and Lourdes García-Rodríguez[c]

[a]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana, Albania, and [c]Dipartimento Física Fundamental y Experimental, Universidad de la Laguna, c/ Astrofisíca Francisco Sánchez s/n, 38204 La Laguna, Tenerife, Spain. Correspondence e-mail: c.giacovazzo@area.ba.cnr.it

The method of joint probability distribution functions has been generalized in order to include and treat different sources of error. The probability distributions of the isomorphous pairs (E_p, E_d) and of the two triples ($E_{p\mathbf{h}}$, $E_{p\mathbf{k}}$, $E_{p\mathbf{k}}$, $E_{p\mathbf{h+k}}$, $E_{d\mathbf{h}}$, $E_{d\mathbf{k}}$, $E_{d\mathbf{h+k}}$) are obtained, on the assumption that the lack of isomorphism and the errors in measurements cumulate on the E_d variables. The conditional distributions of the two-phase and the three-phase structure invariants are derived, showing how the reliability of the probabilistic estimates depends on the errors.

1. Notation

f_j: scattering factor of the jth atom
Σ_p, Σ_d, $\Sigma_H = \sum f_j^2$, where the summation is extended to the protein atoms, to the derivative and to the heavy-atom structure
$\beta = (\Sigma_p/\Sigma_d)^{1/2}$
Z_j: atomic number of the jth atom
Σ_{3d}, Σ_{3p}, $\Sigma_{3h} = \sum f_j(\mathbf{h}_1)f_j(\mathbf{h}_2)f_j(\mathbf{h}_3)$, where the summation is extended to derivative, native protein and heavy atoms, respectively. As usual for direct-methods applications, we will approximate the ratio $(\Sigma^{3/2}/\Sigma_3)$ by $(\sigma_2^{3/2}/\sigma_3)$, where $\sigma_n = \sum Z_j^n$
$F_p = \sum_{j=1}^N f_j \exp(2\pi i\mathbf{h}\mathbf{r}_j) = |F_p|\exp(i\varphi_p)$: structure factor of the native protein
$E_p = A_p + iB_p = R_p\exp(i\varphi_p) = F_p/\Sigma_p^{1/2}$: normalized structure factor of the native protein
$F_H = \sum f_j \exp(2\pi i\mathbf{h}\mathbf{r}_j) = |F_H|\exp(i\varphi_H)$: structure factor of the heavy-atom structure
$F_d = F_p + F_H = |F_d|\exp(i\varphi_d)$: structure factor of the derivative
$E_d = A_d + iB_d = R_d\exp(i\varphi_d) = F_d/\Sigma_d^{1/2}$: E_d is the normalized structure factor of the derivative
$\delta = \varphi_d - \varphi_p$
$\Phi_p \equiv \varphi_{p1} + \varphi_{p2} + \varphi_{p3}$: triplet invariant of the native protein. The subscripts 'pi' stand for '$p\mathbf{h}_i$', under the condition that $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$. A similar notation holds for the subscripts 'di'

2. Introduction

Isomorphous-replacement techniques and direct methods were first integrated by Hauptman (1982), who initiated the probabilistic theories of two-phase and three-phase structure invariants for isomorphous pairs. In his approach, the reciprocal vectors are the primitive random variables, so that E_p and E_d, being functions of the primitive variables, are themselves random variables. In particular, the joint probability distributions

$$P(E_p, E_d), \quad (1)$$

$$P(E_{p1}, E_{p2}, E_{p3}, E_{d1}, E_{d2}, E_{d3}) \quad (2)$$

were obtained, from which the conditional distributions

$$P(\delta|R_p, R_d) \quad (3)$$

and

$$P(\Phi|R_{p1}, R_{p2}, R_{p3}, R_{d1}, R_{d2}, R_{d3}) \quad (4)$$

were respectively derived. The first application of the method (Hauptman *et al.*, 1982) on error-free data was successful, but lack of isomorphism proved to be a strong obstacle when the technique was applied to real data.

The approach was revisited by Giacovazzo *et al.* (1988): their mathematical approach used the atomic coordinates as primitive random variables and took full account of the resolution effects on the distribution parameters. When applied to the case 'native protein–heavy-atom derivative', the final formula estimating the triplet phase of the native protein assumes a very simple expression:

$$P(\Phi_p) \approx [2\pi I_0(G)]^{-1}\exp(G\cos\Phi_p), \quad (5)$$

where

$$G = 2[\sigma_3/\sigma_2^{3/2}]_p R_{p1}R_{p2}R_{p3} + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_1\Delta_2\Delta_3, \quad (6)$$

$I_0(x)$ is the modified Bessel function of order zero and

$$\Delta = (|F_d| - |F_p|)/(\Sigma_H)^{1/2}$$

is the pseudo-normalized difference (with respect to the heavy-atom structure).

A more recent series of papers (Giacovazzo *et al.*, 1994, 1995, 1996; Giacovazzo & Siliqi, 1997) made the direct-methods treatment of the single-isomorphous-replacement (SIR) case practicable and more competitive with the classical technique even when applied to real diffraction data. The aim of the proposed procedure was to phase protein data directly by application of the joint probability distribution (5) rather than by the classical two-step method (*e.g.* Blow & Crick, 1959; Terwillinger & Eisenberg, 1987), requiring first the recovery of the heavy-atom positions and then using this information for phasing protein data.

Distributions (3)–(5) present a weak point: they were obtained by assuming that there is no error in the measurements and that no lack of isomorphism occurs. This assumption is rather limiting for practical applications, mostly when the lack of isomorphism is severe. In this paper, we are interested in deriving the distributions (1)–(5) under the following hypothesis:

$$|F_d|\exp(i\varphi_d) = |F_p|\exp(i\varphi_p) + |F_H|\exp(i\varphi_H) + |\mu|\exp(i\theta), \qquad (7)$$

where $|\mu|\exp(i\theta)$ represents the cumulative error due to lack of isomorphism and to errors in measurements. We will assume, in the absence of any other prior information, that

$$\langle\mu\rangle = 0$$

while θ is a variable uniformly distributed between 0 and 2π. Equation (7) cumulates the full error on the derivative: accordingly,

$$\langle|F_d|^2\rangle = \langle|F_p|^2\rangle + \langle|F_H|^2\rangle + \langle|\mu|^2\rangle = \Sigma_d + \langle|\mu|^2\rangle. \qquad (8)$$

3. The joint probability distribution $P(E_p, E_d)$ in $P1$ and related distributions

Let us assume that: (*a*) the atomic positions of the native protein are the primitive random variables of our probabilistic approach; (*b*) the assumption (8) holds. Then the characteristic function of the distribution (3) may be written as

$$\begin{aligned} C(u_p, \mathrm{v}_p, u_d, \mathrm{v}_d) &= \langle \exp i(u_pA_p + \mathrm{v}_pB_p + u_dA_d + \mathrm{v}_dB_d)\rangle \\ &= \exp\{-\tfrac{1}{4}[u_p^2 + \mathrm{v}_p^2 + (1+\sigma_\mu^2)(u_d^2 + \mathrm{v}_d^2) \\ &\quad + 2\beta u_pu_d + 2\beta \mathrm{v}_p\mathrm{v}_d]\}, \qquad (9) \end{aligned}$$

where u_p, v_p, u_d and v_d are carrying variables associated with A_p, B_p, A_d and B_d, respectively, and $\sigma_\mu^2 = \langle|\mu|^2\rangle/\Sigma_d$. The Fourier transform of (9) gives

$$\begin{aligned} &P(A_p, A_d, B_p, B_d) \\ &= (2\pi)^{-4}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\exp(-i\overline{\mathbf{T}}\mathbf{U})\exp[-\tfrac{1}{2}(\overline{\mathbf{U}}\mathbf{K}\mathbf{U})]\,\mathrm{d}\overline{\mathbf{U}} \\ &\approx (2\pi)^{-2}(\lambda)^{-1/2}\exp(-\tfrac{1}{2}\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}), \end{aligned}$$

where

$$\overline{T} \equiv [A_p, A_d, B_p, B_d], \quad \overline{U} \equiv [u_p, u_d, \mathrm{v}_p, \mathrm{v}_d],$$

$$\mathbf{K} = \begin{vmatrix} 1/2 & \beta/2 & 0 & 0 \\ \beta/2 & (1+\sigma_\mu^2)/2 & 0 & 0 \\ 0 & 0 & 1/2 & \beta/2 \\ 0 & 0 & \beta/2 & (1+\sigma_\mu^2)/2 \end{vmatrix},$$

$$\lambda = \det(\mathbf{K}) = 2^{-4}q^2,$$

$$q = 1 + \sigma_\mu^2 - \beta^2.$$

Λ_{ij} are the elements of $\mathbf{K}^{-1}$, given by

$$\Lambda_{11} = \Lambda_{33} = 2(1+\sigma_\mu^2)/q,$$

$$\Lambda_{22} = \Lambda_{44} = 2/q,$$

$$\Lambda_{12} = \Lambda_{34} = -2\beta/q,$$

$$\Lambda_{13} = \Lambda_{14} = \Lambda_{23} = \Lambda_{24} = 0.$$

Accordingly,

$$\begin{aligned} &P(A_p, A_d, B_p, B_d) \\ &\approx [\pi^2 q]^{-1}\exp\Big\{-\frac{1}{q}[(A_p^2 + B_p^2)(1+\sigma_\mu^2) + (A_d^2 + B_d^2) \\ &\quad - 2\beta(A_pA_d + B_pB_d)]\Big\}. \end{aligned}$$

The change of variables

$$\begin{cases} A_p = R_p\cos\varphi_p \\ B_p = R_p\sin\varphi_p \end{cases} \quad \begin{cases} A_d = R_d\cos\varphi_d \\ B_d = R_d\sin\varphi_d \end{cases}$$

leads to

$$\begin{aligned} &P(R_p, R_d, \varphi_p, \varphi_d) \\ &\approx [R_pR_d/(\pi^2 q)]\exp\Big\{-\frac{1}{q}[R_p^2(1+\sigma_\mu^2) + R_d^2 \\ &\quad - 2\beta R_pR_d\cos(\varphi_d - \varphi_p)]\Big\}. \qquad (10) \end{aligned}$$

The marginal distribution

$$P(\delta|R_p, R_d) \approx [2\pi I_0(Q)]^{-1}\exp(Q\cos\delta)$$

is easily obtained, where

$$Q = 2\beta R_pR_d/q.$$

Accordingly,

$$\langle\cos\delta|R_p, R_d\rangle = I_1(Q)/I_0(Q).$$

It may be observed that the expected value of $\cos\delta$ decreases with increasing value of q (and, therefore, of $\langle|\mu|^2\rangle$). The marginal distribution of the structure-factor moduli is obtained by integrating (10) over the values of the φ_d and φ_p:

Table 1
Statistical analysis of the triplet phase error ($\langle|\Delta\varphi|\rangle$) at selected values of G.

The value of ($\langle|\Delta\varphi|\rangle$) is calculated over all the Nr triplets with G larger than THRES.

	M-FABP (equation 6)		AZET		M-FABP (equation 15)	
THRES	Nr	$\langle\|\Delta\varphi\|\rangle$	Nr	$\langle\|\Delta\varphi\|\rangle$	Nr	$\langle\|\Delta\varphi\|\rangle$
0.4	50000	72.14	10066	52.63	21027	69.31
0.8	49393	72.04	5630	46.30	16147	68.09
1.6	14298	68.28	592	31.44	1089	62.30
2.6	1755	65.94	41	24.46	50	59.52
3.8	169	65.07	0	0.00	2	104.00

$$P(R_p, R_d) \approx [(4R_pR_d)/q]\exp[-(R_p^2 + R_d^2)/q] \times I_0[(2\beta R_p + R_d)/q].$$

The relation (10) reduces to Hauptman's results when $\langle|\mu|^2\rangle = 0$.

4. The joint probability distribution $P(E_{p1}, E_{p2}, E_{p3}, E_{d1}, E_{d2}, E_{d3})$

The characteristic function of the joint probability distribution $P(E_{p1}, E_{p2}, E_{p3}, E_{d1}, E_{d2}, E_{d3})$ is

$$\begin{aligned}
&C(u_{p1}, u_{p2}, u_{p3}, u_{d1}, u_{d2}, u_{d3}, \mathrm{v}_{p1}, \ldots, \mathrm{v}_{d3})\\
&= \langle\exp i(u_{p1}A_{p1} + u_{p2}A_{p2} + \ldots + \mathrm{v}_{d2}B_{d2} + \mathrm{v}_{d3}B_{d3})\rangle\\
&\approx \exp\bigg\{-\tfrac{1}{4}\sum_{i=1}^{3}[(u_{pi}^2 + \mathrm{v}_{pi}^2) + (1+\sigma_{\mu i}^2)(u_{di}^2 + \mathrm{v}_{di}^2)\\
&\quad + 2\beta_i(u_{pi}u_{di} + \mathrm{v}_{pi}\mathrm{v}_{di})] - (i/4)[\sigma_3/\sigma_2^{3/2}]_p\\
&\quad \times [(u_{p1}u_{p2}u_{p3} - \mathrm{v}_{p1}\mathrm{v}_{p2}u_{p3} - \mathrm{v}_{p1}u_{p2}\mathrm{v}_{p3} - u_{p1}\mathrm{v}_{p2}\mathrm{v}_{p3})\\
&\quad + \beta_1(u_{d1}u_{p2}u_{p3} - \mathrm{v}_{d1}\mathrm{v}_{p2}u_{p3} - \mathrm{v}_{d1}u_{p2}\mathrm{v}_{p3} - u_{d1}\mathrm{v}_{p2}\mathrm{v}_{p3})\\
&\quad + \beta_2(u_{p1}u_{d2}u_{p3} - \mathrm{v}_{p1}\mathrm{v}_{d2}u_{p3} - \mathrm{v}_{p1}u_{d2}\mathrm{v}_{p3} - u_{p1}\mathrm{v}_{d2}\mathrm{v}_{p3})\\
&\quad + \beta_3(u_{p1}u_{p2}u_{d3} - \mathrm{v}_{p1}\mathrm{v}_{p2}u_{d3} - \mathrm{v}_{p1}u_{p2}\mathrm{v}_{d3} - u_{p1}\mathrm{v}_{p2}\mathrm{v}_{d3})\\
&\quad + \beta_1\beta_2(u_{d1}u_{d2}u_{p3} - \mathrm{v}_{d1}\mathrm{v}_{d2}u_{p3} - \mathrm{v}_{d1}u_{d2}\mathrm{v}_{p3} - u_{d1}\mathrm{v}_{d2}\mathrm{v}_{p3})\\
&\quad + \beta_1\beta_3(u_{d1}u_{p2}u_{d3} - \mathrm{v}_{d1}\mathrm{v}_{p2}u_{d3} - \mathrm{v}_{d1}u_{p2}\mathrm{v}_{d3} - u_{d1}\mathrm{v}_{p2}\mathrm{v}_{d3})\\
&\quad + \beta_2\beta_3(u_{p1}u_{d2}u_{d3} - \mathrm{v}_{p1}\mathrm{v}_{d2}u_{d3} - \mathrm{v}_{p1}u_{d2}\mathrm{v}_{d3} - u_{p1}\mathrm{v}_{d2}\mathrm{v}_{d3})]\\
&\quad - (i/4)[\sigma_3/\sigma_2^{3/2}]_d(u_{d1}u_{d2}u_{d3} - \mathrm{v}_{d1}\mathrm{v}_{d2}u_{d3} - \mathrm{v}_{d1}u_{d2}\mathrm{v}_{d3}\\
&\quad - u_{d1}\mathrm{v}_{d2}\mathrm{v}_{d3})\bigg\}, \qquad (11)
\end{aligned}$$

where $u_{p1}, u_{p2}, u_{p3}, u_{d1}, u_{d2}, u_{d3}, \mathrm{v}_{p1}, \mathrm{v}_{d3}$ are carrying variables associated with A_{p1}, A_{p2}, B_{d3} and $\beta_i = [\sum_p(\mathbf{h}_i)/\sum_d(\mathbf{h}_i)]^{1/2}$. The change of variables

$$\begin{aligned}
u_{pj} &= 2^{1/2}\rho_{pj}\cos\psi_{pj} & u_{dj} &= [2/(1+\sigma_{\mu j}^2)]^{1/2}\rho_{dj}\cos\theta_{dj}\\
\mathrm{v}_{pj} &= 2^{1/2}\rho_{pj}\sin\psi_{pj} & \mathrm{v}_{dj} &= [2/(1+\sigma_{\mu j}^2)]^{1/2}\rho_{dj}\sin\psi_{dj},\\
A_{pj} &= R_{pj}\cos\varphi_{pj}, & A_{dj} &= R_{dj}\cos\varphi_{dj},\\
B_{pj} &= R_{pj}\sin\varphi_{pj}, & B_{dj} &= R_{dj}\sin\varphi_{dj}
\end{aligned}$$

and the Fourier transform of (11) leads to

$$\begin{aligned}
&P(R_{p1}, R_{p2}, \ldots, R_{d3}, \varphi_{p1}, \ldots, \varphi_{d3})\\
&= (\pi^{-6}/t)\prod_{i=1}^{3}(R_{pi}R_{di})\\
&\quad \exp\bigg\{\sum_{i=1}^{3}[1/(1-\alpha_i^2)]\{-R_{pi}^2 - [R_{di}^2/(1+\sigma_{\mu i}^2)]\\
&\quad + 2\beta_{0i}R_{pi}R_{di}\cos(\varphi_{di} - \varphi_{pi})\}\\
&\quad + 2\beta_0R_{p1}R_{p2}R_{p3}\cos(\varphi_{p1} + \varphi_{p2} + \varphi_{p3})\\
&\quad + 2\beta_{11}R_{d1}R_{p2}R_{p3}\cos(\varphi_{d1} + \varphi_{p2} + \varphi_{p3})\\
&\quad + 2\beta_{12}R_{p1}R_{d2}R_{p3}\cos(\varphi_{p1} + \varphi_{d2} + \varphi_{p3})\\
&\quad + 2\beta_{13}R_{p1}R_{p2}R_{d3}\cos(\varphi_{p1} + \varphi_{p2} + \varphi_{d3})\\
&\quad + 2\beta_{21}R_{p1}R_{d2}R_{d3}\cos(\varphi_{p1} + \varphi_{d2} + \varphi_{d3})\\
&\quad + 2\beta_{22}R_{d1}R_{p2}R_{d3}\cos(\varphi_{d1} + \varphi_{p2} + \varphi_{d3})\\
&\quad + 2\beta_{23}R_{d1}R_{d2}R_{p3}\cos(\varphi_{d1} + \varphi_{d2} + \varphi_{p3})\\
&\quad + 2\beta_{33}R_{d1}R_{d2}R_{d3}\cos(\varphi_{d1} + \varphi_{d2} + \varphi_{d3})\\
&\quad + 2[\sigma_3/\sigma_2^{3/2}]_pR_{p1}R_{p2}R_{p3}\cos(\varphi_{p1} + \varphi_{p2} + \varphi_{p3})\bigg\}, \qquad (12)
\end{aligned}$$

where

$$\begin{gathered}
t = \prod_{i=1}^{3}[(1-\alpha_i^2)(1-\sigma_{\mu i}^2)^{1/2}]^{-1},\\
\alpha_i = \beta_i/(1+\sigma_{\mu i}^2)^{1/2}, \quad \beta_{0i} = \alpha_i/(1+\sigma_{\mu i}^2)^{1/2},\\
\beta_0 = -K\alpha_1\alpha_2\alpha_3, \quad \beta_{11} = K\alpha_2\alpha_3/(1+\sigma_{\mu 1}^2)^{1/2},\\
\beta_{12} = K\alpha_1\alpha_3/(1+\sigma_{\mu 2}^2)^{1/2}, \quad \beta_{13} = K\alpha_1\alpha_2/(1+\sigma_{\mu 3}^2)^{1/2},\\
\beta_{21} = -K\alpha_1/[(1+\sigma_{\mu 2}^2)(1+\sigma_{\mu 3}^2)]^{1/2},\\
\beta_{22} = -K\alpha_2/[(1+\sigma_{\mu 1}^2)(1+\sigma_{\mu 3}^2)]^{1/2},\\
\beta_{23} = -K\alpha_3/[(1+\sigma_{\mu 1}^2)(1+\sigma_{\mu 2}^2)]^{1/2},\\
\beta_{33} = K/[(1+\sigma_{\mu 1}^2)(1+\sigma_{\mu 2}^2)(1+\sigma_{\mu 3}^2)]^{1/2},\\
K = [\sigma_3/\sigma_2^{3/2}]_H\prod_{i=1}^{3}\left[\frac{\alpha_i}{1-\alpha_i^2}\frac{\sum_H(\mathbf{h}_i)}{\sum_p(\mathbf{h}_i)}\right]^{1/2}. \qquad (13)
\end{gathered}$$

From (12), the following conditional probability distribution may be derived:

$$P(\Phi_p|R_{p1}, \ldots, R_{d3}) \approx [2\pi I_0(G)]^{-1}\exp(G\cos\Phi_p),$$

where

$$\begin{aligned}
G = 2[\sigma_3/\sigma_2^{3/2}]_pR_{p1}R_{p2}R_{p3} + 2\bigg[&\beta_0R_{p1}R_{p2}R_{p3}\\
&+ \beta_{11}\frac{R_{d1}R_{p2}R_{p3}}{(1+\sigma_{\mu 1}^2)^{1/2}} + \beta_{12}\frac{R_{p1}R_{d2}R_{p3}}{(1+\sigma_{\mu 2}^2)^{1/2}}\\
&+ \beta_{13}\frac{R_{p1}R_{p2}R_{d3}}{(1+\sigma_{\mu 3}^2)^{1/2}} + \beta_{23}\frac{R_{d1}R_{d2}R_{p3}}{[(1+\sigma_{\mu 1}^2)(1+\sigma_{\mu 2}^2)]^{1/2}}\\
&+ \beta_{22}\frac{R_{d1}R_{p2}R_{d3}}{[(1+\sigma_{\mu 1}^2)(1+\sigma_{\mu 3}^2)]^{1/2}}\\
&+ \beta_{21}\frac{R_{p1}R_{d2}R_{d3}}{[(1+\sigma_{\mu 2}^2)(1+\sigma_{\mu 3}^2)]^{1/2}}\\
&+ \beta_{33}\frac{R_{d1}R_{d2}R_{d3}}{[(1+\sigma_{\mu 1}^2)(1+\sigma_{\mu 2}^2)(1+\sigma_{\mu 3}^2)]^{1/2}}\bigg].
\end{aligned}$$

According to (13), G reduces to

$$G = 2[\sigma_3/\sigma_2^{3/2}]_p R_{p1}R_{p2}R_{p3} + 2K\left(\frac{R_{d1}}{(1+\sigma_{\mu 1}^2)^{1/2}} - \alpha_1 R_{p1}\right)$$
$$\times\left(\frac{R_{d2}}{(1+\sigma_{\mu 2}^2)^{1/2}} - \alpha_2 R_{p2}\right)\left(\frac{R_{d3}}{(1+\sigma_{\mu 3}^2)^{1/2}} - \alpha_3 R_{p3}\right)$$
$$\approx 2[\sigma_3/\sigma_2^{3/2}]_p R_{p1}R_{p2}R_{p3} + 2[\sigma_3/\sigma_2^{3/2}]_H$$
$$\times\prod_{i=1}^{3}\left[\frac{1}{(1+\sigma_{\mu i}^2-\beta_i^2)}\frac{\sum_H(\mathbf{h}_i)}{\sum_d(\mathbf{h}_i)}\Delta_i\right]. \quad (14)$$

The approximation introduced into (14) consists in replacing (in the denominator) the term $[\sum_d(\mathbf{h}_i)\sum_p(\mathbf{h}_i)]^{1/2}$ by $\sum_d(\mathbf{h}_i)$, where $\Delta_i = (F_{di} - F_{pi})/(\Sigma_H)^{1/2}$ is a pseudo-normalized difference (with respect to the heavy-atom structure). Equation (14) is the desired expression estimating triplet invariant phases when errors are considered.

A simpler expression may be obtained by observing that

$$(1+\sigma_{\mu i}^2-\beta^2)^{-1}(\Sigma_H/\Sigma_d) = \Sigma_H/(\Sigma_H+\Sigma_d\sigma_\mu^2)$$
$$= (1+\langle|\mu|^2\rangle/\Sigma_H)^{-1}.$$

Then,

$$G = 2[\sigma_3/\sigma_2^{3/2}]_p R_{p1}R_{p2}R_{p3} + 2[\sigma_3/\sigma_2^{3/2}]_H$$
$$\times\frac{\Delta_1\Delta_2\Delta_3}{[1+(\sigma_{\mu 1}^2)_H][1+(\sigma_{\mu 2}^2)_H][1+(\sigma_{\mu 3}^2)_H]}, \quad (15)$$

where $(\sigma_\mu^2)_H = \langle|\mu|^2\rangle/\Sigma_H$.

Equation (15) suggests how the error influences the reliability of the triplet estimate. It may be noted that, besides the ismorphous differences ($|F_d| - |F_p|$), also the average errors $\langle|\mu|^2\rangle$ are pseudo-normalized with respect to the heavy-atom substructure. In other words, the ratio $\langle|\mu|^2\rangle/\Sigma_H$, rather than the absolute value of $\langle|\mu|^2\rangle$, is the factor responsible for the accuracy of triplet phase relationships. This agrees with common sense: even quite a small $\langle|\mu|^2\rangle$ value may be critical if the scattering power of the heavy-atom substructure is a very small percentage of the derivative scattering power. Equation (15) reduces to (6) when errors are vanishing. It may also be noted that the Cochran term remains unaffected by the error (in accordance with the assumption that the errors accumulate on the F_d's).

5. Conclusions

The method of the joint probability distributions has been applied to isomorphous pairs and isomorphous triplet invariants to treat different sources of error. Our calculations show how they influence the terms Q and G, representing the concentration parameters of the phase distributions for isomorphous pairs and for triplets, respectively.

The assumptions on which our theoretical results are based cannot be strictly verified at this stage. Indeed, we assumed in (7) and (8) that μ is the so-called 'closure error vector', that is,

$$\mu = F_d - (F_p + F_H),$$

cumulating errors due to lack of isomorphism, errors in the heavy-atom structure model and errors in the measurements. The theory so far developed does not exploit F_H as prior information: to take this supplementary information into account, we should have studied the distribution

$$P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H) \quad (16)$$

rather than $P(\mathbf{E}_p, \mathbf{E}_d)$, where $\mathbf{E}_p$, $\mathbf{E}_d$ and $\mathbf{E}_H$ may represent triplets of reflections, *e.g.*

$$\mathbf{E}_p \equiv (E_{p\mathbf{h}_1}, E_{p\mathbf{h}_2}, E_{p\mathbf{h}_3}), \quad \mathbf{E}_d \equiv (E_{d\mathbf{h}_1}, E_{d\mathbf{h}_2}, E_{d\mathbf{h}_3}),$$
$$\mathbf{E}_H \equiv (E_{H\mathbf{h}_1}, E_{H\mathbf{h}_2}, E_{H\mathbf{h}_3})$$

with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$. Accordingly, at this stage of the theory, the measurement errors are the only (and often minor) component of μ. What one can expect from (15) is not an improvement of the triplet phase estimates (indeed the use of the σ factors does not change the signs of the Δ's) but only an assessment of their reliability on a more reasonable footing. Actually, (6) strongly overestimates the reliability of the triplet phases involving isomorphous differences. In Table 1, we show (first three columns) a statistical analysis of the average phase error *versus* selected G values [calculated according to (6)] for the protein M-FABP (Zanotti *et al.*, 1992). In columns 4 and 5 of the same table, we show a similar analysis for the triplets of a typical small crystal structure (AZET; Colens *et al.*, 1974); only the classical Cochran contribution was used to estimate triplet reliability. It is evident that, at the same THRES value, the triplet phase errors for AZET are markedly lower. The use of (15) (see the last two columns of Table 1) still overestimates M-FABP triplet reliability: in order to compensate for the lack of isomorphism, the cumulative error μ^2 in (15) was assumed to be ten times the measurement error.

We conclude with the following remarks:

(*a*) Equation (15) is the first example of a reliability parameter, estimating triplet invariants from isomorphous differences, able to incorporate errors of different natures.

(*b*) The theoretical results so obtained are preliminary to the more ambitious task of calculating (16). The scenario in which (16) should be applied is the following: (15) provides triplet estimates that, involved in a tangent procedure, should be able to generate useful electron-density maps without any information on the heavy-atom positions. However, once phases are available, a difference Fourier synthesis with coefficient $(F_d - F_p)\exp(i\varphi_p)$ may automatically provide approximate heavy-atom structure parameters that can be refined by standard techniques. Then the subsequent use of (16) would be able to improve the triplet phase estimates and therefore lead, *via* a tangent approach, to improved protein electron density. The process should be completely automated.

References

Blow, D. M & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Colens, A., Declercq, J. P, Germain, G., Putzeys, J. P & Van Meerssche, M. (1974). *Cryst. Struct. Commun.* **3** 119–122.

Giacovazzo, C., Cascarano, G. & Zheng, C.-D. (1988). *Acta Cryst.* A**44**, 45–51.
Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.
Giacovazzo, C., Siliqi, D., González-Platas, J., Hecht, H., Zanotti, G. & York, B. (1996). *Acta Cryst.* D**52**, 813–825.
Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* A**50**, 609–621.
Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst.* A**51**, 177–188.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Hauptman, H., Potter, S. & Weeks, C. M. (1982). *Acta Cryst.* A**38**, 294–300.
Terwilliger, T. C. & Eisenberg, D. (1987). *Acta Cryst*. A**43**, 1–5, 6–13.
Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.