RESEARCH PAPERS

Acta Cryst. (1997). A**53**, 253–263

On Integrating Direct Methods and Isomorphous Replacement Techniques. The Formula P_{10}

C. GIACOVAZZO,[a]* D. SILIQI,[b] G. CASCARANO,[c] R. CALIANDRO[c] AND A. MELIDORO[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 University, Tirana, Albania, and* [c]*Istituto di Ricerca per lo Sviluppo di Metodologie Cristallografiche, CNR, c/o Dipartimento Geomineralogico, Campus Universitario, Via Orabona 4, 70125 Bari, Italy. E-mail: giacovazzo@arba.ba.cnr.it*

(*Received 5 June 1996; accepted 22 October 1996*)

Abstract

The joint probability distribution of ten pairs of isomorphous structure factors has been derived. Their indices correspond to the reflexions contained in the second phasing shell of the triplet invariant, as described by the theory of representations [Giacovazzo (1977). *Acta Cryst.* A**33**, 934–944; Giacovazzo (1980). *Acta Cryst.* A**36**, 362–373]. The conclusive formula allows the estimation of triplet invariants *via* a second representation formula, called the P_{10} formula, which is more accurate than the traditional formula of Giacovazzo, Cascarano & Zheng [*Acta Cryst.* (1988), A**44**, 45–51]. The procedure is also able to take into consideration the prior information on the heavy-atom structure, when available.

1. Symbols and notation

The notation is that used in the papers by Giacovazzo & Siliqi (1996*a*,*b*).

2. Introduction

The integration of isomorphous replacement techniques with direct methods was initiated by Hauptman (1982). The main goals of the Hauptman paper are:

(*a*) The joint probability distribution

$$P(\phi_{\mathbf{h}}, \psi_{\mathbf{k}}, R_{\mathbf{h}}, S_{\mathbf{h}}) \qquad (1)$$

was found, where

$$E_{\mathbf{h}} = R_{\mathbf{h}} \exp(i\phi_{\mathbf{h}}) = (1/\alpha_{20}^{1/2}) \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j),$$

$$G_{\mathbf{h}} = S_{\mathbf{h}} \exp(i\psi_{\mathbf{h}}) = (1/\alpha_{02}^{1/2}) \sum_{j=1}^{N} g_j \exp(2\pi i \mathbf{h}\mathbf{r}_j)$$

are the normalized structure factors of the protein and of the derivative, respectively, and

$$\alpha_{mn} = \sum_{j=1}^{N} f_j^m g_j^n.$$

From (1), the conditional distribution

$$P(\phi_{\mathbf{h}} - \psi_{\mathbf{h}} | R_{\mathbf{h}}, S_{\mathbf{h}}) \qquad (2)$$

was derived;

(*b*) The joint probability distribution function

$$P(\phi_{\mathbf{h}_1}, \phi_{\mathbf{h}_2}, \phi_{\mathbf{h}_3}, \psi_{\mathbf{h}_1}, \psi_{\mathbf{h}_2}, \psi_{\mathbf{h}_3}, R_{\mathbf{h}_1}, R_{\mathbf{h}_2}, R_{\mathbf{h}_3}, S_{\mathbf{h}_1}, S_{\mathbf{h}_2}, S_{\mathbf{h}_3}) \qquad (3)$$

was obtained, where $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$.

The Hauptman approach was revisited by Giacovazzo, Cascarano & Zheng (1988). The main goals may be described as:

(*a*) The joint probability distributions (2) and (3) were obtained by considering the atomic positions as the primitive random variables. In the Hauptman approach, the primitive random variable was the ordered triple $(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$ of the reciprocal vectors, which is assumed to be uniformly distributed over the subset of vectors satisfying the condition $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$.

(*b*) A formula for the triplet invariant estimate was derived, which holds when the derivative is obtained by addition of heavy atoms:

$$P(\Phi | R_{\mathbf{h}_1}, \ldots, S_{\mathbf{h}_3}) = [2\pi I_0 A]^{-1} \exp(A \cos \Phi), \qquad (4)$$

where

$$\Phi = \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3},$$

$$A = 2[\sigma_3/\sigma_2^{3/2}]_p R_{\mathbf{h}_1} R_{\mathbf{h}_2} R_{\mathbf{h}_3} + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_{\mathbf{h}_1} \Delta_{\mathbf{h}_2} \Delta_{\mathbf{h}_3}, \qquad (5)$$

$\sigma_n = \sum_j Z_j^n$, Z_j being the atomic number of the jth atom. The suffixes p, H and d refer to protein and heavy-atom

and derivative structure, respectively. Furthermore,

$$\Delta = (|E_d'| - |E_p'|),$$
$$E_p' = F_p/\Sigma_H^{1/2}, \quad E_d' = F_d/\Sigma_H^{1/2}.$$

E_p' and E_d' are the structure factors of the protein and of the derivative, respectively, normalized with respect to the heavy-atom structure.

Formula (4) succeeded when applied to experimental diffraction data and constitutes the main tool of a procedure aiming at phasing protein reflexions in the absence of information on the heavy-atom positions (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo & Gonzalez Platas, 1995; Giacovazzo, Siliqi, Gonzalez Platas, Hecht, Zanotti & York, 1996). A recent application of the method of joint probability distribution functions to evaluate quartet invariants from isomorphous data (Giacovazzo & Siliqi, 1996*a*,*b*) opened new perspectives. Indeed:

(*a*) The Gram–Charlier expansion of the characteristic function was used (see Giacovazzo, 1980*a*) in the calculations. Such an approach allowed the awful calculations necessary to evaluate quartet invariants *via* the exponential form of the characteristic function (as used by Hauptman for the triplet estimation). The calculations were still quite heavy but the conclusive formulas are simple to use. Such a result suggests that the derivation of even more complicated joint probability distributions could be performed *via* the same approach.

(*b*) The efficiency of the formula estimating quartet invariants indicates that the cross magnitudes of the quartet provide useful information. Why then not integrate such a mathematical approach with the theory of representations of structure factors (Giacovazzo, 1977, 1980*a*; see Hauptman, 1975, for a related principle). According to such a formulation, for each invariant or seminvariant Φ, a sequence of sets of reflexions (phasing shells) may be identified, each shell contained in the succeeding one, having the property that Φ may be estimated *via* the magnitudes constituting any shell. The sequence of the shells reflects the order of their expected efficiency (in the statistical sense) for the estimation of the Φ.

In the absence of isomorphous data, the second representation of Φ is the collection of special quintets

$$\{\Psi\} = \{\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{k}} - \phi_{\mathbf{k}}\}, \tag{6}$$

where $\mathbf{k}$ is a free vector that can span over all the reciprocal space. Since $\Psi \equiv \Phi$ for any $\mathbf{k}$, any estimate of Ψ simultaneously provides an estimate of Φ. By making explicit use of the symmetry, one may write (6) as

$$\{\Psi\} = \{\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i} - \phi_{\mathbf{kR}_i}\}, \quad i = 1, \ldots m, \tag{7}$$

where $\mathbf{k}$ varies over the symmetry-independent reflexions and the vectors $\mathbf{kR}_i$ are the m symmetry equivalents of a given $\mathbf{k}$. The second representation estimate of the triplet Φ benefits by the cross magnitudes of any quintet in (7). From a probabilistic point of view, the best estimates of Φ can be obtained *via* the study of the joint probability distribution function

$$P(E_{\mathbf{h}_1}, E_{\mathbf{h}_2}, E_{\mathbf{h}_3}, E_{\mathbf{h}_1+\mathbf{k}}, E_{\mathbf{h}_1-\mathbf{k}}, E_{\mathbf{h}_2+\mathbf{k}}, E_{\mathbf{h}_2-\mathbf{k}}, E_{\mathbf{h}_3+\mathbf{k}}, E_{\mathbf{h}_3-\mathbf{k}}).$$

Such work has been done by Cascarano, Giacovazzo, Camalli, Spagna, Burla, Nunzi & Polidori (1984). The conclusive formula was

$$P(\Phi) \simeq [2\pi I_0(G)]^{-1} \exp(G \cos \Phi), \tag{8}$$

where

$$G = C(1 + Q), \tag{9}$$

$$\begin{aligned}
C &= 2|E_{\mathbf{h}_1} E_{\mathbf{h}_2} E_{\mathbf{h}_3}|/N^{1/2} \\
Q &= \sum_{\mathbf{k}} \{T_{\mathbf{k}}/[1 + (\varepsilon_{\mathbf{h}_1}\varepsilon_{\mathbf{h}_2}\varepsilon_{\mathbf{h}_3} + B_{\mathbf{k}})]\} \\
T_{\mathbf{k}} &= \sum_{i=1}^{m} T_{\mathbf{k},i} \\
B_{\mathbf{k}} &= \sum_{i=1}^{m} B_{\mathbf{k},i} \\
T_{\mathbf{k},i} &= N^{-1}\varepsilon_{\mathbf{k}}[\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}(\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}) \\
&\quad + \varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}(\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}) \\
&\quad + \varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}(\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_2-\mathbf{kR}_i})], \\
B_{\mathbf{k},i} &= (2N)^{-1}\varepsilon_{\mathbf{h}_1}[\varepsilon_{\mathbf{k}}(\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}) \\
&\quad + \varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}] \\
&\quad + \varepsilon_{\mathbf{h}_2}[\varepsilon_{\mathbf{k}}(\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}) \\
&\quad + \varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}] \\
&\quad + \varepsilon_{\mathbf{h}_3}[\varepsilon_{\mathbf{k}}(\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}) \\
&\quad + \varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i} + \varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}].
\end{aligned} \tag{10}$$

If no $\mathbf{k}$ is used, (8) reduces to the Cochran (1955) formula. In general, G may be positive or negative; when $G < 0$, the most probable value of Φ is π.

The combination of the representation theory with isomorphous replacement techniques could allow better estimates of the invariants. In particular, more accurate triplet estimates could be obtained if the concept of representation of a triplet is extended to isomorphous data. Such an extension has been accomplished by Giacovazzo (1984): the main results are quoted below.

Let us suppose that r isomorphous data sets are available. Then,

$$\Phi = \phi_{i,\mathbf{h}_1} + \phi_{j,\mathbf{h}_2} + \phi_{s,\mathbf{h}_3} \tag{11}$$

is a triplet provided $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$. No limiting conditions hold for the indices i, j, s, which can arbitrarily vary between 1 and r.

The first representation of Φ is the collection of the triplets (11) obtained when i, j, s freely vary between 1 and r. For $r = 2$, the above definition leads to the eight combinations

$$\begin{aligned} &\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3}, \quad \psi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3}, \\ &\phi_{\mathbf{h}_1} + \psi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3}, \quad \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \psi_{\mathbf{h}_3}, \\ &\psi_{\mathbf{h}_1} + \psi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3}, \quad \psi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \psi_{\mathbf{h}_3}, \\ &\phi_{\mathbf{h}_1} + \psi_{\mathbf{h}_2} + \psi_{\mathbf{h}_3}, \quad \psi_{\mathbf{h}_1} + \psi_{\mathbf{h}_2} + \psi_{\mathbf{h}_3}. \end{aligned} \tag{12}$$

The above algebraic definition has its counterpart in the probabilistic procedure leading to the joint probability distribution function (3), which explicitly involves in its expression all the eight triplets (12). The difficulty in deriving (3) arose from the necessity of calculating the mathematical interactions among the eight types of triplets.

The second representation of Φ is the collection of special quintets

$$\Psi = \{\phi_{i,\mathbf{h}_1} + \phi_{j,\mathbf{h}_2} + \phi_{s,\mathbf{h}_3} + \phi_{p,\mathbf{k}} - \phi_{q,\mathbf{k}}\}. \tag{13}$$

Again, no limiting conditions hold for the indices i, j, s, p and q, which can arbitrarily vary between 1 and r. For $r = 2$, the second representation of Φ involves 32 special quintets, four for each triplet in (12). For example,

$$\begin{aligned} &\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{k}} - \phi_{\mathbf{k}}, \\ &\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{k}} - \psi_{\mathbf{k}}, \\ &\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \psi_{\mathbf{k}} - \phi_{\mathbf{k}}, \\ &\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \psi_{\mathbf{k}} - \psi_{\mathbf{k}}, \\ &\\ &\psi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{k}} - \phi_{\mathbf{k}}, \\ &\psi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{k}} - \psi_{\mathbf{k}}, \\ &\psi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \psi_{\mathbf{k}} - \phi_{\mathbf{k}}, \\ &\psi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \psi_{\mathbf{k}} - \psi_{\mathbf{k}} \end{aligned}$$

etc. The second representation estimate of the triplet phase Φ benefits by the prior knowledge of the cross magnitudes of the quintets (13). Accordingly, we have to study the joint probability distribution function

$$\begin{aligned} &P(E_{\mathbf{h}_1}, E_{\mathbf{h}_2}, E_{\mathbf{h}_3}, E_{\mathbf{h}_1+\mathbf{k}}, E_{\mathbf{h}_1-\mathbf{k}}, E_{\mathbf{h}_2+\mathbf{k}}, E_{\mathbf{h}_2-\mathbf{k}}, E_{\mathbf{h}_3+\mathbf{k}}, E_{\mathbf{h}_3-\mathbf{k}}, \\ &\quad G_{\mathbf{h}_1}, G_{\mathbf{h}_2}, G_{\mathbf{h}_3}, \ldots, G_{\mathbf{h}_3-\mathbf{k}}), \end{aligned} \tag{14}$$

which involves ten isomorphous pairs of structure factors. Since $\mathbf{k}$ is a free vector that can span over all the reciprocal space, one could derive an overall distribution by combining the distributions (14) obtained for single values of $\mathbf{k}$.

3. The characteristic function of the joint probability distribution function of ten isomorphous pairs of structure factors

Let

$$C(\nu_1, \ldots, \nu_{10}, \mu_1, \ldots, \mu_{10}, \rho_1, \ldots, \rho_{10}, \gamma_1, \ldots, \gamma_{10})$$

be the characteristic function of the distribution

$$\begin{aligned} P_{10} \equiv P(&\phi_1, \ldots, \phi_{10}, \psi_1, \ldots, \psi_{10}, R_1, \ldots, R_{10}, \\ &S_1, \ldots, S_{10}), \end{aligned}$$

where

$$\begin{aligned} E_1 &= R_1 \exp(i\phi_1) = R_{\mathbf{h}_1} \exp(i\phi_{\mathbf{h}_1}) \\ E_2 &= R_2 \exp(i\phi_2) = R_{\mathbf{h}_2} \exp(i\phi_{\mathbf{h}_2}) \\ E_3 &= R_3 \exp(i\phi_3) = R_{\mathbf{h}_3} \exp(i\phi_{\mathbf{h}_3}) \\ E_4 &= R_4 \exp(i\phi_4) = R_{\mathbf{h}_4} \exp(i\phi_{\mathbf{k}}) \\ E_5 &= R_5 \exp(i\phi_5) = R_{\mathbf{h}_1+\mathbf{k}} \exp(i\phi_{\mathbf{h}_1+\mathbf{k}}) \\ E_6 &= R_6 \exp(i\phi_6) = R_{\mathbf{h}_1-\mathbf{k}} \exp(i\phi_{\mathbf{h}_1-\mathbf{k}}) \\ E_7 &= R_7 \exp(i\phi_7) = R_{\mathbf{h}_2+\mathbf{k}} \exp(i\phi_{\mathbf{h}_2+\mathbf{k}}) \\ E_8 &= R_8 \exp(i\phi_8) = R_{\mathbf{h}_2-\mathbf{k}} \exp(i\phi_{\mathbf{h}_2-\mathbf{k}}) \\ E_9 &= R_9 \exp(i\phi_9) = R_{\mathbf{h}_3+\mathbf{k}} \exp(i\phi_{\mathbf{h}_3+\mathbf{k}}) \\ E_{10} &= R_{10} \exp(i\phi_{10}) = R_{\mathbf{h}_3-\mathbf{k}} \exp(i\phi_{\mathbf{h}_3-\mathbf{k}}) \\ G_1 &= S_1 \exp(i\psi_1) = S_{\mathbf{h}_1} \exp(i\psi_{\mathbf{h}_1}) \\ G_2 &= S_2 \exp(i\psi_2) = S_{\mathbf{h}_2} \exp(i\psi_{\mathbf{h}_2}) \\ &\vdots \\ G_{10} &= S_{10} \exp(i\psi_{10}) = S_{\mathbf{h}_3-\mathbf{k}} \exp(i\psi_{\mathbf{h}_3-\mathbf{k}}). \end{aligned}$$

$\nu_i, \mu_i, \rho_i, \gamma_i$ for $i = 1, \ldots, 10$ are the carrying variables associated with ϕ_i, ψ_i, R_i, S_i for $i = 1, \ldots, 10$, respectively, and $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$. The characteristic function, expanded in a Gram–Charlier series, may be written as

$$\begin{aligned} P_{10} = &\int_0^\infty \ldots \int_0^\infty \int_0^{2\pi} \ldots \int_0^{2\pi} \prod_{i=1}^{20} \{(1/2\pi^2) R_i S_i \rho_i \gamma_i \\ &\times \exp[-\tfrac{1}{2}(\rho_i^2 + \gamma_i^2) - i2^{1/2} \rho_i R_i \cos(\phi_i - \nu_i) \\ &- i2^{1/2} \gamma_i S_i \cos(\psi_i - \mu_i) - \alpha_i \rho_i \gamma_i \cos(\nu_i - \mu_i)]\} \\ &\times \{1 - 2^{-1/2} i[\gamma_{1,2,3} \rho_1 \rho_2 \rho_3 \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\ &+ \gamma_{1,4,5} \rho_1 \rho_4 \rho_5 \cos(\nu_1 + \nu_4 - \nu_5) + \circlearrowleft \\ &+ \gamma_{1,4,6} \rho_1 \rho_4 \rho_6 \cos(\nu_1 - \nu_4 - \nu_6) + \circlearrowleft \\ &+ \gamma_{1,7,10} \rho_1 \rho_7 \rho_{10} \cos(\nu_1 + \nu_7 + \nu_{10}) + \circlearrowleft \\ &+ \gamma_{1,8,9} \rho_1 \rho_8 \rho_9 \cos(\nu_1 + \nu_8 + \nu_9) + \circlearrowleft \\ &+ \gamma_{2,4,7} \rho_2 \rho_4 \rho_7 \cos(\nu_2 + \nu_4 - \nu_7) + \circlearrowleft \\ &+ \gamma_{2,4,8} \rho_2 \rho_4 \rho_8 \cos(\nu_2 - \nu_4 - \nu_8) + \circlearrowleft \\ &+ \gamma_{2,5,10} \rho_2 \rho_5 \rho_{10} \cos(\nu_2 + \nu_5 + \nu_{10}) + \circlearrowleft \end{aligned}$$

$$
\begin{aligned}
&+ \gamma_{2,6,9}\rho_2\rho_6\rho_9 \cos(\nu_2 + \nu_6 + \nu_9) + \circlearrowleft \\
&+ \gamma_{3,4,9}\rho_3\rho_4\rho_9 \cos(\nu_3 + \nu_4 - \nu_9) + \circlearrowleft \\
&+ \gamma_{3,4,10}\rho_3\rho_4\rho_{10} \cos(\nu_3 - \nu_4 - \nu_{10}) + \circlearrowleft \\
&+ \gamma_{3,5,8}\rho_3\rho_5\rho_8 \cos(\nu_3 + \nu_5 + \nu_8) + \circlearrowleft \\
&+ \gamma_{3,6,7}\rho_3\rho_6\rho_7 \cos(\nu_3 + \nu_6 + \nu_7) + \circlearrowleft] \\
&- \tfrac{1}{4}[\gamma^2_{1,2,3}\rho^2_1\rho^2_2\rho^2_3 \cos^2(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\
&+ \gamma^2_{1,4,5}\rho^2_1\rho^2_4\rho^2_5 \cos^2(\nu_1 + \nu_4 - \nu_5) + \circlearrowleft \\
&+ \gamma^2_{1,4,6}\rho^2_1\rho^2_4\rho^2_6 \cos^2(\nu_1 - \nu_4 - \nu_6) + \circlearrowleft \\
&+ \gamma^2_{1,7,10}\rho^2_1\rho^2_7\rho^2_{10} \cos^2(\nu_1 + \nu_7 + \nu_{10}) + \circlearrowleft \\
&+ \gamma^2_{1,8,9}\rho^2_1\rho^2_8\rho^2_9 \cos^2(\nu_1 + \nu_8 + \nu_9) + \circlearrowleft \\
&+ \gamma^2_{2,4,7}\rho^2_2\rho^2_4\rho^2_7 \cos^2(\nu_2 + \nu_4 - \nu_7) + \circlearrowleft \\
&+ \gamma^2_{2,4,8}\rho^2_2\rho^2_4\rho^2_8 \cos^2(\nu_2 - \nu_4 - \nu_8) + \circlearrowleft \\
&+ \gamma^2_{2,5,10}\rho^2_2\rho^2_5\rho^2_{10} \cos^2(\nu_2 + \nu_5 + \nu_{10}) + \circlearrowleft \\
&+ \gamma^2_{2,6,9}\rho^2_2\rho^2_6\rho^2_9 \cos^2(\nu_2 + \nu_6 + \nu_9) + \circlearrowleft \\
&+ \gamma^2_{3,4,9}\rho^2_3\rho^2_4\rho^2_9 \cos^2(\nu_3 + \nu_4 - \nu_9) + \circlearrowleft \\
&+ \gamma^2_{3,4,10}\rho^2_3\rho^2_4\rho^2_{10} \cos^2(\nu_3 - \nu_4 - \nu_{10}) + \circlearrowleft \\
&+ \gamma^2_{3,5,8}\rho^2_3\rho^2_5\rho^2_8 \cos^2(\nu_3 + \nu_5 + \nu_8) + \circlearrowleft \\
&+ \gamma^2_{3,6,7}\rho^2_3\rho^2_6\rho^2_7 \cos^2(\nu_3 + \nu_6 + \nu_7) + \circlearrowleft] \\
&+ \gamma_{1,2,5,7}\rho_1\rho_2\rho_5\rho_7 \cos(\nu_1 - \nu_2 - \nu_5 + \nu_7) \\
&+ \ldots \\
&+ i2^{-3/2}[\gamma_{1,4,5}\gamma_{2,4,8}\gamma_{3,5,8}\rho_1\rho_2\rho_3\rho^2_4\rho^2_5\rho^2_8 \\
&\times \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\
&+ \gamma_{1,4,6}\gamma_{2,4,7}\gamma_{3,6,7}\rho_1\rho_2\rho_3\rho^2_4\rho^2_6\rho^2_7 \\
&\times \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\
&+ \gamma_{1,7,10}\gamma_{2,4,7}\gamma_{3,4,10}\rho_1\rho_2\rho_3\rho^2_4\rho^2_7\rho^2_{10} \\
&\times \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\
&+ \gamma_{1,8,9}\gamma_{2,4,8}\gamma_{3,4,9}\rho_1\rho_2\rho_3\rho^2_4\rho^2_8\rho^2_9 \\
&\times \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\
&+ \gamma_{1,4,5}\gamma_{2,5,10}\gamma_{3,4,10}\rho_1\rho_2\rho_3\rho^2_4\rho^2_5\rho^2_{10} \\
&\times \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft \\
&+ \gamma_{1,4,6}\gamma_{2,6,9}\gamma_{3,4,9}\rho_1\rho_2\rho_3\rho^2_4\rho^2_6\rho^2_9 \\
&\times \cos(\nu_1 + \nu_2 + \nu_3) + \circlearrowleft] \\
&+ \ldots\}, \qquad (15)
\end{aligned}
$$

where

$$\gamma_{1,4,5} = (\Sigma_1\Sigma_4\Sigma_5)^{-1/2}\sum_{j=1}^{N} f_j(\mathbf{h}_1)f_j(\mathbf{k})f_j(\mathbf{h}_1 + \mathbf{k}),$$

$$\gamma_{1,4,6} = (\Sigma_1\Sigma_4\Sigma_6)^{-1/2}\sum_{j=1}^{N} f_j(\mathbf{h}_1)f_j(\mathbf{k})f_j(\mathbf{h}_1 - \mathbf{k}),$$

$$\vdots$$

$$\gamma_{1,2,5,7} = (\Sigma_1\Sigma_2\Sigma_5\Sigma_7)^{-1/2}\sum_{j=1}^{N} f_j(\mathbf{h}_1)f_j(\mathbf{h}_2)f_j(\mathbf{h}_1 + \mathbf{k}) \times f_j(\mathbf{h}_2 + \mathbf{k})$$

and $\Sigma_i = \sum_{j=1}^{N} f_j^2$ is calculated for the ith reflexion.

The number of terms in the distribution (15) is extremely large. We have quoted [as in Giacovazzo & Siliqi (1996*a*,*b*) for the quartet distribution] only those that significantly contribute to the estimation of Φ. We note:

(*a*) We have used a curved arrow to represent the 'cyclic terms' of a prototype term [only the prototypes are quoted in (15)]. For example, the complete set of cyclic terms for $\gamma_{1,4,5}\rho_1\rho_4\rho_5\cos(\nu_1 + \nu_4 - \nu_5)$ (the prototype included) are the eight terms

$$\begin{aligned}
&\gamma_{1,4,5}\rho_1\rho_4\rho_5\cos(\nu_1 + \nu_4 - \nu_5), \quad &&\gamma_{1,4,\bar{5}}\rho_1\rho_4\gamma_5\cos(\nu_1 + \nu_4 - \mu_5), \\
&\gamma_{1,\bar{4},5}\rho_1\gamma_4\rho_5\cos(\nu_1 + \mu_4 - \nu_5), \quad &&\gamma_{1,\bar{4},\bar{5}}\rho_1\gamma_4\gamma_5\cos(\nu_1 + \mu_4 - \mu_5), \\
&\gamma_{\bar{1},4,5}\gamma_1\rho_4\rho_5\cos(\mu_1 + \nu_4 - \nu_5), \quad &&\gamma_{\bar{1},\bar{4},5}\gamma_1\gamma_4\rho_5\cos(\mu_1 + \mu_4 - \nu_5), \\
&\gamma_{\bar{1},4,\bar{5}}\gamma_1\rho_4\gamma_5\cos(\mu_1 + \nu_4 - \mu_5), \quad &&\gamma_{\bar{1},\bar{4},\bar{5}}\gamma_1\gamma_4\gamma_5\cos(\mu_1 + \mu_4 - \mu_5),
\end{aligned}$$

where

$$\gamma_{1,4,\bar{5}} = (\Sigma_1\Sigma_4\Sigma'_5)^{-1/2}\sum_{j=1}^{N} f_j(\mathbf{h}_1)f_j(\mathbf{k})g_j(\mathbf{h}_1 + \mathbf{k})$$

$$\gamma_{1,\bar{4},5} = (\Sigma_1\Sigma'_4\Sigma_5)^{-1/2}\sum_{j=1}^{N} f_j(\mathbf{h}_1)g_j(\mathbf{k})f_j(\mathbf{h}_1 + \mathbf{k})$$

$$\vdots.$$

$\Sigma'_i = \sum_{j=1}^{N} g_j^2$ is calculated for the ith reflexion. In all, the distribution involves $13 \times 8 = 104$ different triplet terms and we quote in (15) only 13 prototypes.

(*b*) the cyclic terms of $\gamma_{1,2,5,7}\rho_1\rho_2\rho_5\rho_7 \times \cos(\nu_1 - \nu_2 - \nu_5 + \nu_7)$, the prototype included, are the 16 terms

$$\begin{aligned}
&\gamma_{1,2,5,7}\rho_1\rho_2\rho_5\rho_7\cos(\nu_1 - \nu_2 - \nu_5 + \nu_7) \\
&\gamma_{1,2,5,\bar{7}}\rho_1\rho_2\rho_5\gamma_7\cos(\nu_1 - \nu_2 - \nu_5 + \mu_7) \\
&\gamma_{1,2,\bar{5},\bar{7}}\rho_1\rho_2\gamma_5\gamma_7\cos(\nu_1 - \nu_2 - \mu_5 + \mu_7) \\
&\quad\vdots \\
&\gamma_{\bar{1},\bar{2},\bar{5},\bar{7}}\gamma_1\gamma_2\gamma_5\gamma_7\cos(\mu_1 - \mu_2 - \mu_5 + \mu_7),
\end{aligned}$$

where

$$\gamma_{1,2,5,\bar{7}} = (\Sigma_1\Sigma_2\Sigma_5\Sigma'_7)^{-1/2}\sum_{j=1}^{N} f_j(\mathbf{h}_1)f_j(\mathbf{h}_2)f_j(\mathbf{h}_1 + \mathbf{k}) \times g_j(\mathbf{h}_2 + \mathbf{k})$$

etc. There are 42 prototypes, for a total of $42 \times 16 = 672$ quartet-type terms.

(*c*) There are 2^9 cyclic terms of the prototype $\gamma_{1,4,5}\gamma_{2,4,8}\gamma_{3,5,8}\rho_1\rho_2\rho_3\rho^2_4\rho^2_5\rho^2_8\cos(\nu_1 + \nu_2 + \nu_3)$, prototype included. In all, $2^9 \times 6 = 3072$ terms of order

$N^{-3/2}$ are contained in the distribution, giving contributions of order $N^{-3/2}$ for the estimation of Φ.

There are too many terms in (15) to register the contribution of each single term. We will only give here the final results. The reader interested in the mathematical details is referred to the recent papers by Giacovazzo & Siliqi (1996*a*,*b*).

4. The triplet phase probability formula

The final expression of the conditional probability distribution function $P_{10}(\Phi|\{R, S\})$ has a simpler form when expressed in terms of pseudo-normalized (with respect to the heavy-atom structure) structure factors R'_i and S'_i, where

$$R_i = R'_i[\sigma_2]_H^{1/2}/[\sigma_2]_p^{1/2},$$
$$S_i = S'_i[\sigma_2]_H^{1/2}/[\sigma_2]_d^{1/2}.$$

We obtain

$$P_{10} \equiv P(\Phi|R'_1, R'_2, R'_3, \ldots, S'_{10}) = [2\pi I_0(A)]^{-1} \exp\{A\cos\Phi\}, \quad (16)$$

where

$$A = 2[\sigma_3/\sigma_2^{3/2}]_p R_1R_2R_3 + 2(\Delta'_1\Delta'_2\Delta'_3/N_H^{1/2}) \times \{1 + T_{\mathbf{k}}/[1 + (\langle L'_1\rangle\langle L'_2\rangle\langle L'_3\rangle + B_{\mathbf{k}})]\}, \quad (17)$$

$$N_H = [\sigma_3/\sigma_2^{3/2}]_H$$

is the equivalent number of heavy atoms in the unit cell,

$$\Delta'_j = S'_jD'_{1j} - R'_j,$$
$$D'_{1j} = I_1(2R'_jS'_j)/I_0(2R'_jS'_j),$$
$$T_{\mathbf{k}} = \sum_{i=1}^{m} T_{k,i}, \quad B_{\mathbf{k}} = \sum_{i=1}^{m} B_{k,i}.$$

I_i is the modified Bessel function of order i. Furthermore,

$$T_{k,i} = N_H^{-1}\langle L'_4\rangle[\langle L'_5\rangle\langle L'_8\rangle + \langle L'_6\rangle\langle L'_7\rangle + \langle L'_7\rangle\langle L'_{\overline{10}}\rangle + \langle L'_8\rangle\langle L'_9\rangle + \langle L'_5\rangle\langle L'_{\overline{10}}\rangle + \langle L'_6\rangle\langle L'_9\rangle]_i,$$

$$B_{k,i} = (2N_H)^{-1}[\langle L'_1\rangle\langle L'_2\rangle\langle L'_3\rangle + \langle L'_1\rangle\langle L'_4\rangle\langle L'_5\rangle + \langle L'_1\rangle\langle L'_4\rangle\langle L'_6\rangle + \langle L'_1\rangle\langle L'_7\rangle\langle L'_{\overline{10}}\rangle + \langle L'_1\rangle\langle L'_8\rangle\langle L'_9\rangle + \langle L'_2\rangle\langle L'_4\rangle\langle L'_7\rangle + \langle L'_2\rangle\langle L'_4\rangle\langle L'_8\rangle + \langle L'_2\rangle\langle L'_5\rangle\langle L'_{\overline{10}}\rangle + \langle L'_2\rangle\langle L'_6\rangle\langle L'_9\rangle + \langle L'_3\rangle\langle L'_4\rangle\langle L'_9\rangle + \langle L'_3\rangle\langle L'_4\rangle\langle L'_{\overline{10}}\rangle + \langle L'_3\rangle\langle L'_5\rangle\langle L'_8\rangle + \langle L'_3\rangle\langle L'_6\rangle\langle L'_7\rangle]_i$$

$$\langle L'_j\rangle = (S'^2_j + R'^2_j - 2R'_jS'_jD'_{1j}) - 1.$$

We observe:

(*a*) the distribution (16) is a Von Mises-type function: it is unimodal and the expected value of Φ is 0 or π according to whether A is positive or negative.

Table 1. *Code name, space group and crystallochemical data for the test structures*

Structure code	Reference	Space group	Molecular formula	Z
APP	(*a*)	$C2$	$C_{190}N_{53}O_{58}Zn$	4
CARP	(*b*)	$C2$	$C_{513}N_{131}O_{121}Ca_2S$	4
BPO	(*c*)	$P2_13$	$C_{2744}N_{712}O_{1073}$	12
E2	(*d*)	$F432$	$C_{1170}N_{310}O_{366}S_7$	96

References: (*a*) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983); (*b*) Kretsinger & Nockolds (1973); (*c*) Hecht, Sobek, Haag, Pfeifer & Van Pee (1994); (*d*) Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol (1992).

Table 2. *Relevant parameters for diffraction data of the test structures*

	Native		Derivative			
Structure code	Res. (Å)	NREFL	Heavy atom	$[\sigma_2]_H/[\sigma_2]_p$	Res. (Å)	NREFL
APP	0.99	17058	Hg	0.055	2.0	2086
CARP	1.70	5056	Hg	0.044	2.0	4687
BPO	2.35	23956	Au	0.028	2.78	15741
E2	2.65	10391	Hg	0.021	3.0	9179

(*b*) for proteins, the term $2[\sigma_3/\sigma_2^{3/2}]_pR_1R_2R_3$ is quite often negligible with respect to the second term in (17). It will be neglected in the following considerations.

(*c*) the contribution from the second phasing shell can change the value of the expected phase. According to the first representation formula, Φ is expected to be zero if $(\Delta'_1\Delta'_2\Delta'_3)$ is positive and is expected to be π if $(\Delta'_1\Delta'_2\Delta'_3)$ is negative. In the second representation formula, the term $\mathrm{CORR}_{\mathbf{k}} = T_{\mathbf{k}}/[1 + (\langle L'_1\rangle\langle L'_2\rangle\langle L'_3\rangle + B_{\mathbf{k}})]$ may be considered a correction term that modulates the first representation estimate. If $\mathrm{CORR}_{\mathbf{k}} < -1$, the second representation estimate is different by π from the first representation estimate.

(*d*) it is safe to assume $B = 0$ if $B < 0$, by analogy with the results obtained by Giacovazzo (1976, 1980*b*) for quartet estimates.

(*e*) if some of the cross terms are unknown, then suitable marginal distributions should be calculated. The final result is the following: if the pair (R'_j, S'_j) is not among the measured data then A may be updated by omitting in (17) the terms including $\langle L'_j\rangle$.

(*f*) for large values of the product $(2R'_jS'_j)$ (this occurs because R' and S' are pseudonormalized structure factors for which $\langle R'^2\rangle$ and $\langle S'^2\rangle$ are remarkably larger than unity), the value of D'_1 attains unity. Then,

$$\Delta'_i \equiv \Delta_i = S'_i - R'_i \quad \text{and} \quad \langle L'_i\rangle = \Delta_i^2 - 1.$$

In practice, we will use several indices $\mathbf{k}$ in our probabilistic approach. The contribution coming from the various special quintets so obtained may be combined in a single formula, say (16) again, in

Table 3. *E2: statistical calculations for triplet invariants (found among the 800 reflexions with the largest $|\Delta|$) relative to formulas (4) and (16)*

Calculated error-free data for native and derivative structures are used. NR is the number of phase relationships having $|A|$ [as defined by (5) and (18)] larger than ARG, % is the percentage ($\times 100$) of phase relationships whose cosine sign is correctly estimated and $\langle|\Phi|\rangle$ (°) is the average of the absolute values of the triplet phase Φ.

	(4) Positive estimated triplets			(16) Positive estimated triplets			(16) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	25359	88	45	23823	90	43	581	57	100
1.2	22279	89	44	15086	92	41	69	80	131
3.2	12	100	10	882	93	39	1	100	147
4.4	0	-	-	53	100	40	0	-	-

	(4) Negative estimated triplets			(16) Positive estimated triplets			(16) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	24641	88	135	619	61	79	23098	90	137
1.2	18828	90	136	66	88	46	14134	92	139
3.2	9	100	170	0	-	-	743	92	142
4.4	0	-	-	0	-	-	0	-	-

Table 4. *BPO: statistical calculations for triplet invariants (found among the 800 reflexions with the largest $|\Delta|$) relative to formulas (4) and (16)*

Calculated error-free data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(16) Positive estimated triplets			(16) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	25187	91	42	20858	93	40	3208	19	55
1.2	25187	91	42	16667	93	39	1431	26	62
3.2	1062	98	31	6260	94	38	242	42	83
4.4	0	-	-	2450	94	38	78	53	96
9.0	0	-	-	15	100	16	1	100	125

	(4) Negative estimated triplets			(16) Positive estimated triplets			(16) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	24813	91	138	3122	19	125	20565	93	140
1.2	24813	91	138	1352	25	117	16329	93	140
3.2	905	98	148	209	43	92	5979	94	142
4.4	0	-	-	66	64	69	2318	94	143
9.0	0	-	-	2	100	4	14	100	152

which (17) is replaced by

$$A = 2[\sigma_3/\sigma_2^{3/2}]_p R_1 R_2 R_3 + 2(\Delta_1'\Delta_2'\Delta_3'/N_H^{1/2}) \times \left\{1 + \sum_{\mathbf{k}} \mathrm{CORR}_{\mathbf{k}}\right\}. \quad (18)$$

5. Check of the probabilistic formula (16) by calculated diffraction data

We first apply (4) and (16) to calculated data in order to check their relative efficiency in the case of perfect isomorphism and in the absence of experimental errors in measurements. Structure factors are calculated from refined positional and vibrational parameters of the test structures quoted in Table 1. The relevant parameters for the experimental data of the test structures are shown in Table 2; calculated data will reflect the same parameters.

For each test structure, the normalized structure factors up to derivative resolution are arranged in decreasing order of $|\Delta|$: the code number of the ith reflexion [CODE(i)] is just such an order number. Millions of triplets can be calculated among the reflexions: the estimate of each triplet *via* (16) requires the exploration of the reciprocal space *via* a ten-node figure, which sweeps out by letting $\mathbf{k}$ freely vary over the set of reciprocal vectors. It is clear, from the above observations, that the use of (16) may be very time consuming even for very fast computers. We therefore decided to check the relative efficiency of (4) and (16) by introducing two restrictions in the calculations: (*a*) triplets are only found among the reflexions with the smallest value of CODE (*e.g.* those with the largest

Table 5. *E2: statistical calculations for triplet invariants (found among the 800 reflexions with the largest $|\Delta|$) relative to formulas (4) and (19)*

Calculated error-free data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	25359	88	45	24004	90	43	675	62	105
1.2	22279	89	44	15750	93	39	160	86	138
3.2	12	100	10	532	100	26	16	100	166
4.4	0	-	-	20	100	20	6	100	164

	(4) Negative estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	24641	88	135	647	67	71	23255	90	137
1.2	18828	90	136	153	86	44	14788	93	141
3.2	9	100	170	14	86	32	425	99	153
4.4	0	-	-	4	75	49	16	100	161

Table 6. *BPO: statistical calculations for triplet invariants (found among the 800 reflexions with the largest $|\Delta|$) relative to formulas (4) and (19)*

Calculated error-free data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	25187	91	42	24600	92	41	388	58	97
1.2	25187	91	42	22018	93	39	137	78	115
3.2	1062	98	31	3956	98	33	29	97	139
4.4	0	-	-	351	99	26	7	100	150
9.0	0	-	-	0	-	-	0	-	-

	(4) Negative estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	24813	91	138	397	54	81	24204	92	139
1.2	24813	91	138	154	70	63	21604	93	140
3.2	905	98	148	29	100	27	3625	98	148
4.4	0	-	-	10	100	14	282	99	154
9.0	0	-	-	0	-	-	0	-	-

value of $|\Delta|$) instead of among the full set of data. However, the number of triplets involved in the calculations is statistically significant; (*b*) **k** varies over a very limited subset of reflexions, the 20 reflexions with codes between 1 and 20. The hope is that even a small number of special quintets can provide useful supplementary information for the triplet estimation (in fact, N_H is a small number).

We show in Tables 3 and 4 the results of our tests for E3 and BPO. NR is the number of triplets with $|A|$ [as defined by (5) or (18)] larger than ARG, % is the percentage ($\times 100$) of the phase relationships whose cosine sign is correctly determined, and $\langle|\Phi|\rangle$ is the average (in degrees) of the absolute values of the triplet phases Φ. The key for correct reading of the tables is: the triplets estimated positive by (4) are submitted to (16), which splits them into positive and negative estimated triplets. Similarly, triplets estimated negative by (4) are submitted to (16), which splits them again into positive and negative estimated triplets. We observe:

(*a*) the efficiency of (16) for E2 is satisfactory: a non-negligible number of triplets wrongly estimated by (4) are recognized and correctly evaluated by (16);

(*b*) (16) is less efficient for BPO: triplets wrongly estimated by (4) are correctly estimated by (16) only for large values of $|A|$.

The above behaviour is not unexpected. In two recent papers (Giacovazzo & Siliqi, 1996*a*,*b*), a theory for the estimation of quartet invariants has been described, exploiting the prior information provided by isomorphous data. In the conclusive formula, estimating the quartet phase given seven pairs of isomorphous reflexions, the cross terms influence the estimates *via*

Table 7. *E2: statistical calculations for triplet invariants (found among the 855 reflexions with the largest $|\Delta|$) relative to formulas (4) and (19)*

Observed data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	25058	72	65	19537	79	57	2967	62	104
1.2	4281	81	54	8088	85	50	599	74	119
3.2	0	-	-	239	95	36	21	91	146
4.4	0	-	-	30	100	23	1	100	159

	(4) Negative estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	24942	71	114	2961	64	74	19234	78	122
1.2	3234	81	126	531	75	62	7161	85	131
3.2	0	-	-	27	85	56	207	94	143
4.4	0	-	-	7	86	55	17	100	157

Table 8. *BPO: statistical calculations for triplet invariants (found among the 1500 reflexions with the largest $|\Delta|$) relative to formulas (4) and (19)*

Observed data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	25195	68	69	20107	72	65	2785	52	92
1.2	8680	72	64	10145	77	59	676	58	100
3.2	0	-	-	531	84	50	30	40	98
4.4	0	-	-	70	90	44	2	50	116

	(4) Negative estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	24805	68	110	2739	51	89	19688	71	115
1.2	6919	72	115	581	58	82	9485	76	120
3.2	0	-	-	27	74	61	437	80	126
4.4	0	-	-	8	75	67	45	78	122

the same functions $\langle L_i' \rangle$ occurring in (16). It was noted that a component of $\langle L_i' \rangle$, say $(S_i'^2 + R_i'^2 - 2R_i'S_i'D_{1i}')$, is nothing but the expected squared magnitude of the normalized structure factor of the heavy-atom structure. The lack of information on $|E_H|_i$ (not directly available from the moduli R_i' and S_i') brings an ambiguity into the probabilistic approach that may be eliminated when $|E_H|_i$ is known. The reader is referred to the detailed analysis of the problem described by Giacovazzo & Siliqi (1996*b*) for the quartet case. We observe here that the information on the heavy-atom structure is not a necessary requisite for determining protein phases *via* triplet invariants (see Giacovazzo, Siliqi, Gonzalez Platas, Hecht, Zanotti & York, 1996). However, once protein phases are available, the difference Fourier synthesis with coefficients $(|F_d| - |F_p|)\exp(i\phi_p)$ straightforwardly reveals the heavy-atom positions: then, the set $\{F_H\}$ becomes available and may be used for improving triplet or quartet estimates. Consequently, protein phases can be improved too. It is therefore of non-negligible interest to check how the information on the heavy-atom structure modifies the probabilistic formula (16). The results obtained for the quartet invariants suggest that (16) has to be replaced by

$$P_{10} \equiv P(\Phi|R_1', R_2', R_3', \ldots, S_{10}', R_{H1}, R_{H2} \ldots, R_{H10}) = [2\pi I_0(A_H)]^{-1} \exp\{A_H \cos \Phi\}, \quad (19)$$

where

$$A_H = 2[\sigma_3/\sigma_2^{3/2}]_p R_1 R_2 R_3 + 2(\Delta_1'\Delta_2'\Delta_3'/N_H^{1/2}) \times \left\{1 + \sum_{\mathbf{k}} [\mathrm{CORR}_{\mathbf{k}}]_H\right\}, \quad (20)$$

where $[\mathrm{CORR}_{\mathbf{k}}]_H$ is obtained from $\mathrm{CORR}_{\mathbf{k}}$ by replacing $\langle L_j' \rangle$ by $\varepsilon_{Hj} = |E_H|_j^2 - 1$ [that is, the expected value of $(|E_H|^2 - 1)$ by its known (on the basis of the heavy-atom model structure) value]. The efficiency of

Table 9. *APP: statistical calculations for triplet invariants (found among the 400 reflexions with the largest $|\Delta|$) relative to formulas (4) and (19)*

Calculated error-free data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	7092	91	42	6817	91	42	10	0	29
1.2	6896	91	42	5408	91	43	0	-	-
2.6	882	99	28	2012	95	36	0	-	-
4.4	0	-	-	80	100	21	0	-	-

	(4) Negative estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	5333	91	137	7	0	143	5159	91	137
1.2	4529	92	139	0	-	-	3758	91	138
2.6	278	100	161	0	-	-	816	98	147
4.4	0	-	-	0	-	-	6	100	152

Table 10. *APP: statistical calculations for triplet invariants (found among the 400 reflexions with the largest $|\Delta|$) relative to formulas (4) and (19)*

Observed data for native and derivative structures are used. See Table 3 for additional details on the symbols.

	(4) Positive estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	5703	82	53	5160	84	51	338	47	86
1.2	5567	83	52	4370	84	51	126	83	128
2.6	3878	86	48	3132	84	51	56	91	141
4.4	1697	90	40	1860	87	48	16	94	145

	(4) Negative estimated triplets			(19) Positive estimated triplets			(19) Negative estimated triplets		
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)	NR	%	$\langle\|\Phi\|\rangle$ (°)
0.2	4654	75	119	278	61	73	4179	78	122
1.2	4348	77	121	116	89	40	3412	77	121
2.6	2662	82	126	31	100	30	2318	79	126
4.4	1004	89	134	8	100	26	1260	85	129

the new formula may be deduced from Tables 5 and 6. The results are now satisfactory also for BPO.

6. Tests on experimental data

We focus our attention on distribution (19) since it seems to have the potentiality of substantially improving triplet estimates obtained *via* (4) even when experimental data are used, that is in the presence of imperfect isomorphism and of experimental errors in measurements. The results of our tests on the experimental data of E2 and BPO are shown in Tables 7 and 8. Distribution (19) proves to be an efficient tool for recognizing triplets wrongly estimated by (4).

7. A special case

The distribution (19) unexpectedly failed when our calculations were applied to APP: we show in Table 9 our statistical tests on calculated data (experimental data behave similarly). A negligible number of triplets estimated positive by (4) are estimated negative by (19) and *vice versa*: their reliability parameter is very small, *i.e.* ($|A_H|$ is close to 0.1) and they are substantially wrongly estimated. The reason is not perfectly understood: probably, it concerns the centrosymmetrical nature of the heavy-atom structure and the quite small value of N_H (when $N_H = 2$, the Wilson distribution is strongly violated). In order to check if our guess stands, we simulated a different isomorphous derivative of APP, where the heavy-atom structure is constituted by two symmetry-independent Hg atoms, one more than the real Hg derivative. The statistical results on the triplet reliability are shown in Table 10: they are quite satisfactory, proving the general efficiency of our probabilistic approach. Similar tests were made for CARP, where the same situation occurring for APP may be found. The results, not shown for brevity, fit quite well those obtained for APP.

8. The use of the heavy-atom structure information

We provided two types of P_{10} formula, the first [(16)] working in the absence of any prior information on the heavy-atom structure, the second [(19)] using the heavy-atom structure as prior. A technique for improving triplet estimates by introducing such a prior was suggested by Fortier, Moore & Fraser (1985) and, in another context, by Klop, Krabbendam & Kroon (1987). The Fortier *et al.* method may be summarized as: once the heavy atoms have been located, the cosine moduli of the doublet invariants $\delta_i = \psi_i - \phi_i$, $i = 1, 2, 3$, can be estimated *via* the Carnot relation

$$\cos\delta = (|F_d|^2 - |F_p|^2 - |F_H|^2)/2|F_dF_p|. \quad (21)$$

Accordingly, the distribution

$$P(\phi_1, \phi_2, \phi_3, \delta_1, \delta_2, \delta_3 | \{R'_i, S'_i, i = 1, 2, 3\}) \quad (22)$$

can be derived by a simple change of variable from the distribution

$$P(\phi_1, \phi_2, \phi_3, \psi_1, \psi_2, \psi_3 | \{R'_i, S'_i, i = 1, 2, 3\}).$$

We have

$$\begin{aligned} P(\phi_1, \phi_2, \phi_3, \delta_1, \delta_2, \delta_3 | \{R'_i, S'_i, i = 1, 2, 3\}) \\ \cong (1/L)\exp\Big\{\sum_{i=1}^{3} 2R'_iS'_i\cos\delta_i + 2[\sigma_3/\sigma_2^{3/2}]_pR'_1R'_2R'_3\cos\Phi \\ + 2[\sigma_3/\sigma_2^{3/2}]_H[-R'_1R'_2R'_3\cos\Phi + S'_1R'_2R'_3\cos(\Phi+\delta_1) \\ - R'_1S'_2R'_3\cos(\Phi+\delta_2) + R'_1R'_2S'_3\cos(\Phi+\delta_3) \\ + R'_1S'_2S'_3\cos(\Phi+\delta_2+\delta_3) - S'_1R'_2S'_3\cos(\Phi+\delta_1+\delta_3) \\ - S'_1S'_2R'_3\cos(\Phi+\delta_1+\delta_3) \\ + S'_1S'_2S'_3\cos(\Phi+\delta_1+\delta_2+\delta_3)]\Big\}. \end{aligned} \quad (23)$$

In the Fortier *et al.* method, the signs of the δ_i's were supposed unknown: then, from (4), the conditional distribution

$$P(\Phi | \{R'_i, S'_i, |\delta_i|, i = 1, 2, 3\}) \quad (24)$$

may be obtained as a weighted sum of the eight distributions corresponding to the eight sign combinations of the doublet invariants δ_1, δ_2 and δ_3. From the eight sign combinations, four cosine-invariant estimates can be obtained: the final cosine invariant is then obtained as their weighted average. Fortier *et al.* recognized: (*a*) with the cosine invariants in hand, the tangent refinement is no longer the best tool available for the determination of the individual phases; (*b*) a least-squares analysis of cosine invariants (Karle & Hauptman, 1957; Hauptman, 1972) could be better used to evaluate the individual phases.

The above considerations indicate that the Fortier *et al.* formula provides estimates that should be compared with the expected cosine, while (19), like any formula that has to be used in a tangent-refinement process, generates phase estimates. Owing to such a difference, the relative efficiency of the two formulas may be evaluated by taking into consideration triplets constituted by symmetry-restricted phases. Indeed, the estimate 0 or π for the triplet phase obtained *via* (19) may be directly compared with the sign of the triplet cosine obtained *via* the Fortier *et al.* formula.

Table 11. *NRP(i) and NRN(i) are the number of triplets evaluated positive or negative by the relation (i), respectively; NER(i) is the number of wrong estimates*

E2			
NRP(4)	NRP(19)	NRN(19)	NRP(21)
[NER(4)]	[NER(19)]	[NER(19)]	[NER(21)]
1427	1172	255	1427
(212)	(18)	(61)	(212)
NRN(4)	NRN(19)	NRP(19)	NRN(21)
[NER(4)]	[NER(19)]	[NER(19)]	[NER(21)]
1457	1190	266	1457
(220)	(14)	(61)	(220)
BPO			
NRP(4)	NRP(19)	NRN(19)	NRP(21)
[NER(4)]	[NER(19)]	[NER(19)]	[NER(21)]
324	316	8	324
(7)	(0)	(0)	(7)
NRN(4)	NRN(19)	NRP(19)	NRN(21)
[NER(4)]	[NER(19)]	[NER(19)]	[NER(21)]
410	399	11	410
(10)	(0)	(1)	(10)

Our calculations are summarized in Table 11. For BPO and E2, we calculated, among the 1600 reflexions with highest value of $|\Delta|$, 734 and 2884 restricted triplet phase invariants, respectively. NRP(*i*) is the number of triplets evaluated positive by the relation (*i*), NER(*i*) is the number of wrong estimates, NRN(*i*) is the number of triplets evaluated negative by relation (*i*).

Table 11 clearly suggests that (24) is unable to change the triplet sign estimates provided by (4). The reason is readily understood: the estimates of $\cos\delta$, obtained *via* (21), do not contain enough supplementary information with respect to that originally exploited by (4). Such supplementary information is provided by the cross terms of the quintet invariants constituting the second representation of the triplet invariant. A last observation deserves to be made: the estimates provided by (19) are more biased toward Φ_H than the estimates *via* (4) or (16). Such behaviour is the natural consequence of the *prior* information (*i.e.* heavy-atom structure) used in (19). The amount of such bias is related to the percentage of the cosine triplet invariants that change sign when (4) and (16) are replaced by (19).

9. Conclusions

This paper shows that what a few years ago seemed a formidable task, that is the estimation of the triplet

invariants *via* their second representation formula applied to isomorphous data, may be accomplished. The approach is rather complicated but conclusive formulas are effective and sufficiently robust against loss of isomorphism and errors in measurements, so that they can be applied to practical cases. The method reveals an interesting feature: cross-vector contribution is of order $N_H^{-3/2}$. Since N_H is usually a small value, the cross terms of the special quintets may substantially modify the triplet estimates provided by the first representation formula of Giacovazzo, Cascarano & Zheng (1988). A undesirable feature of the method is the following: the cross-term contribution depends on $\varepsilon_H = (R_H^2 - 1)$, but this value is not available from experimental data. In this case, the mathematical approach naturally requires the use of the expected value of ε_H, say $\langle L \rangle$. However, when the heavy-atom structure is available, ε_H may be used, and this information improves the efficiency of the formula.

The formulas derived in this paper exploit both the basis and the cross magnitudes in the second representation of the triplet invariants. They can contribute, together with the results recently obtained for the quartet invariants (Giacovazzo & Siliqi, 1996*a*,*b*; Kyriakidis, Peschar & Schenk, 1996), to enlighten three highly relevant problems: (*a*) the evaluation of the information really provided by the cross terms for estimation of the structure invariants when isomorphous data are available; (*b*) the nature and the amount of the information provided by the prior knowledge of the heavy-atom structure; (*c*) the relation between the entire armory of direct methods and the traditional single isomorphous replacement (SIR) approach, when the heavy-atom structure is or is not available. All such themes will be discussed in a future paper.

References

Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* A**40**, 278–283.

Cochran, W. (1955). *Acta Cryst.* **8**, 91–99.

Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.

Giacovazzo, C. (1976). *Acta Cryst.* A**32**, 473.

Giacovazzo, C. (1977). *Acta Cryst.* A**33**, 934–944.

Giacovazzo, C. (1980*a*). *Acta Cryst.* A**36**, 362–373.

Giacovazzo, C. (1980*b*). *Direct Methods in Crystallography*. London: Academic Press.

Giacovazzo, C. (1984). International School of Crystallography, Lecture Notes from Direct Methods of Solving Crystal Structures, Erice, Italy.

Giacovazzo, C., Cascarano, G. & Zheng, C. (1988). *Acta Cryst.* A**44**, 45–51.

Giacovazzo, C. & Gonzalez Platas, J. (1995). *Acta Cryst.* A**51**, 398–404.

Giacovazzo, C. & Siliqi, D. (1996*a*). *Acta Cryst.* A**52**, 133–142.

Giacovazzo, C. & Siliqi, D. (1996*b*). *Acta Cryst.* A**52**, 143–151.

Giacovazzo, C., Siliqi, D. & Gonzalez Platas, J. (1995). *Acta Cryst.* A**51**, 811–820.

Giacovazzo, C., Siliqi, D., Gonzalez Platas, J., Hecht, H. J., Zanotti, G. & York, B. (1996). *Acta Cryst.* D**52**, 813–825.

Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* A**50**, 503–510.

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.

Glover, I., Haneef, I., Pitts, J., Woods, S., Moss, D., Tickle, I. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293–304.

Hauptman, H. (1972). *Crystal Structure Determination: the Role of the Cosine Seminvariants*. New York/London: Plenum Press.

Hauptman, H. (1975). *Acta Cryst.* A**31**, 680–687.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.

Hecht, H., Sobek, H., Haag, T., Pfeifer, O. & Van Pee, K. H. (1994). *Nature (London) Struct. Biol.* **1**, 532–537.

Karle, J. & Hauptman, H. (1957). *Acta Cryst.* **10**, 515–524.

Klop, E. A., Krabbendam, H. & Kroon, J. (1987). *Acta Cryst.* A**43**, 810–820.

Kretsinger, R. H. & Nockolds, C. E. (1973). *J. Biol. Chem.* **248**, 3313–3326.

Kyriakidis, C. E., Peschar, R. & Schenk, H. (1996). *Acta Cryst.* A**52**, 77–87.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., De Kok, A. & Hol, W. G. J. (1992). *Science*, **255**, 1544–1550.