Acta Cryst. (1997). A**53**, 161–167

General Expression for Probabilistic Estimation of Multiphase Structure Invariants in the Case of a Native Protein and Multiple Derivatives. Application to Estimates of the Three-Phase Structure Invariants

NING-HAI HU* AND YONG-SHENG LIU

Applied Spectroscopy Laboratory, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, People's Republic of China. E-mail: peifk@bepc2.ihep.ac.cn

(Received 15 April 1996; accepted 30 September 1996)

Abstract

Concise probabilistic formulae with definite crystallographic implications are obtained from the distribution for eight three-phase structure invariants (3PSIs) in the case of a native protein and a heavy-atom derivative [Hauptman (1982). *Acta Cryst.* A**38**, 289–294] and from the distribution for 27 3PSIs in the case of a native and two derivatives [Fortier, Weeks & Hauptman (1984). *Acta Cryst.* A**40**, 646–651]. The main results of the probabilistic formulae for the four-phase structure invariants are presented and compared with those for the 3PSIs. The analysis directly leads to a general formula of probabilistic estimation for the n-phase structure invariants in the case of a native and m derivatives. The factors affecting the estimated accuracy of the 3PSIs are examined using the diffraction data from a moderate-sized protein. A method to estimate a set of the large-modulus invariants, each corresponding to one of the eight 3PSIs, that has the largest $|\Delta|$ values and relatively large structure-factor moduli between the native and derivative is suggested, which remarkably improves the accuracy, and thus a phasing procedure making full use of all eight 3PSIs is proposed.

1. Introduction

The probabilistic theory of the three-phase structure invariants (3PSIs) that integrates the techniques of direct methods with isomorphous replacement was worked out by Hauptman (1982). The initial application (Hauptman, Potter & Weeks, 1982) confirmed the theoretical validity and promising potential of the approach. However, the mathematical complexity of the distribution makes it difficult to gain its further interpretation with crystallographic implications. Later, through some mathematical manipulations, Fortier, Weeks & Hauptman (1984*a*) obtained a useful interpretation of the distribution formula in terms of experimental parameters, the diffraction ratio and the difference in the intensities of a native protein and its heavy-atom derivative and, subsequently, they applied a similar interpretation method to the distribution formula for the case of a native and two derivatives (Fortier, Weeks & Hauptman, 1984*b*). Taking account of the resolution effects on distribution parameters, Giacovazzo, Cascarano & Zheng (1988) proposed a probabilistic formula for estimating the 3PSIs by fixing a triplet of reciprocal vectors **H**, **K**, **L** and choosing atomic coordinates to be the primitive random variables. The formula was first applied to direct solution of protein structures (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995). For the special case of a native and a heavy-atom derivative, it was shown that the formula has a concise form different from the corresponding result of Fortier *et al.* (1984*a*) and allows an easier interpretation in terms of diffraction experiments.

In this paper, we show that a concise expression can be directly obtained from Hauptman's distribution in the case of a native and a derivative, as well as from the distribution of Fortier *et al.* (1984*b*) in the case of a native and two derivatives, which is different from the formula of Giacovazzo *et al.* (1988) in approach but equally satisfactory in result. It is also shown here that the probability distribution of the four-phase structure invariants (4PSIs), which was recently derived by the present authors and a detailed account of which will be published separately, has the same property as that of the 3PSIs. Based on these results, a general formula for the multiphase invariants in the case of a native and multiple derivatives can be deduced. Finally, a phasing procedure that makes full use of the eight 3PSIs is proposed.

2. The probabilistic formulae for estimating the 3PSIs

2.1. *The case of a native protein and a heavy-atom derivative*

When the triplet of reciprocal-lattice vectors **H**, **K**, **L** satisfies $\mathbf{H}+\mathbf{K}+\mathbf{L}=\mathbf{0}$, the conditional probability distribution presented by Hauptman (1982) for eight 3PSIs,

$$\begin{aligned}
\omega_1 &= \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}, & \omega_5 &= \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}},\\
\omega_2 &= \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}}, & \omega_6 &= \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \varphi_{\mathbf{L}},\\
\omega_3 &= \varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \varphi_{\mathbf{L}}, & \omega_7 &= \psi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}},\\
\omega_4 &= \varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}}, & \omega_8 &= \psi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}},
\end{aligned} \tag{1}$$

is [for the notation in this section see Hauptman (1982) unless otherwise indicated]

$$P_i(\Omega_i|R_1, R_2, R_3, S_1, S_2, S_3) \simeq (1/K_i)\exp(A_i \cos \Omega_i), \quad i = 1, \ldots, 8. \tag{2}$$

The A_i term can be written as

$$\begin{aligned}
A_i = 2[&\beta_1 C_{1R}C_{2R}C_{3R}R_1R_2R_3 + \beta_2(C_{1R}C_{2R}C_{3S}R_1R_2S_3\\
&+ C_{1R}C_{2S}C_{3R}R_1S_2R_3 + C_{1S}C_{2R}C_{3R}S_1R_2R_3)\\
&+ \beta_3(C_{1R}C_{2S}C_{3S}R_1S_2S_3 + C_{1S}C_{2R}C_{3S}S_1R_2S_3\\
&+ C_{1S}C_{2S}C_{3R}S_1S_2R_3) + \beta_4 C_{1S}C_{2S}C_{3S}S_1S_2S_3],
\end{aligned} \tag{3}$$

where

$$C_{jR} = 1, \quad C_{jS} = I_1(x)/I_0(x) \tag{4}$$

if the jth phase of the invariant is φ,

$$C_{jR} = I_1(x)/I_0(x), \quad C_{jS} = 1 \tag{5}$$

if the jth phase of the invariant is ψ, and

$$x = 2\gamma R_j S_j, \quad j = 1, 2, 3.$$

In the case of a native and a heavy-atom derivative, the atomic content of the derivative (D) is assumed to equal the atomic content of the native protein (P) plus the heavy-atom content (H). Then, the parameters γ and β_j, $j = 1, 2, 3, 4$, are reduced to

$$\begin{aligned}
\gamma &= \alpha_{20}^{1/2}\alpha_{02}^{1/2}/(\alpha_{02} - \alpha_{20}),\\
\beta_1 &= \alpha_{30}\alpha_{20}^{-3/2} - (\alpha_{03} - \alpha_{30})\alpha_{20}^{3/2}/(\alpha_{02} - \alpha_{20})^3,\\
\beta_2 &= (\alpha_{03} - \alpha_{30})\alpha_{20}\alpha_{02}^{1/2}/(\alpha_{02} - \alpha_{20})^3,\\
\beta_3 &= -(\alpha_{03} - \alpha_{30})\alpha_{20}^{1/2}\alpha_{02}/(\alpha_{02} - \alpha_{20})^3,\\
\beta_4 &= (\alpha_{03} - \alpha_{30})\alpha_{02}^{3/2}/(\alpha_{02} - \alpha_{20})^3.
\end{aligned} \tag{6}$$

Substituting (6) into (3), we get

$$\begin{aligned}
A_i = &2\sigma_{3P}\sigma_{2P}^{-3/2}C_{1R}C_{2R}C_{3R}R_1R_2R_3\\
&+ 2\sigma_{3H}\sigma_{2H}^{-3}(C_{1S}\alpha_{02}^{1/2}S_1 - C_{1R}\alpha_{20}^{1/2}R_1)\\
&\times (C_{2S}\alpha_{02}^{1/2}S_2 - C_{2R}\alpha_{20}^{1/2}R_2)(C_{3S}\alpha_{02}^{1/2}S_3 - C_{3R}\alpha_{20}^{1/2}R_3),
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
\sigma_{3P} &= \alpha_{30} = \sum_P Z_j^3, & \sigma_{2P} &= \alpha_{20} = \sum_P Z_j^2,\\
\sigma_{3H} &= \alpha_{03} - \alpha_{30} = \sum_H Z_j^3, & \sigma_{2H} &= \alpha_{02} - \alpha_{20} = \sum_H Z_j^2,
\end{aligned} \tag{8}$$

Z_j is the atomic number of the jth atom in the unit cell and the summations over P and over H state that the indices j vary over protein atoms and over heavy atoms, respectively. Obviously, $F_{jP} = \alpha_{20}^{1/2}R_j$ and $F_{jD} = \alpha_{02}^{1/2}S_j$, $j = 1, 2, 3$, are the structure-factor moduli for protein and derivative, respectively. In terms of F_P and F_D, a simplified expression for A_i is obtained:

$$\begin{aligned}
A_i &= 2\sigma_{3P}\sigma_{2P}^{-3/2}C_{1R}C_{2R}C_{3R}R_1R_2R_3\\
&\quad + 2\sigma_{3H}\sigma_{2H}^{-3}(C_{1S}F_{1D} - C_{1R}F_{1P})\\
&\quad \times (C_{2S}F_{2D} - C_{2R}F_{2P})(C_{3S}F_{3D} - C_{3R}F_{3P})\\
&= 2\sigma_{3P}\sigma_{2P}^{-3/2}C_{1R}C_{2R}C_{3R}R_1R_2R_3 + 2\sigma_{3H}\sigma_{2H}^{-3/2}\Delta_1\Delta_2\Delta_3,
\end{aligned} \tag{9}$$

where

$$\Delta_j = (C_{jS}F_{jD} - C_{jR}F_{jP})/\sigma_{2H}^{1/2}, \quad j = 1, 2, 3, \tag{10}$$

is a modified normalized structure-factor magnitude of the heavy-atom structure, as further described below. Define $\Delta_j = \Delta_{jR}$ when $C_{jR} = 1$ and $\Delta_j = \Delta_{jS}$ when $C_{jS} = 1$, $j = 1, 2, 3$. Then, for example, for $\omega_1 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$, (9) becomes

$$A_1 = 2\sigma_{3P}\sigma_{2P}^{-3/2}R_1R_2R_3 + 2\sigma_{3H}\sigma_{2H}^{-3/2}\Delta_{1R}\Delta_{2R}\Delta_{3R} \tag{11}$$

and, for $\omega_5 = \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}}$,

$$\begin{aligned}
A_5 = &2\sigma_{3P}\sigma_{2P}^{-3/2}C_{1R}C_{2R}C_{3R}R_1R_2R_3\\
&+ 2\sigma_{3H}\sigma_{2H}^{-3/2}\Delta_{1S}\Delta_{2S}\Delta_{3S}.
\end{aligned} \tag{12}$$

The $R_1R_2R_3$ term of (11) is the well known traditional Cochran (1955) distribution, which is usually negligible for protein structures. The sign of A is determined by the $\Delta_{1R}\Delta_{2R}\Delta_{3R}$ term. When $C_{jS} \simeq 1.0$, *i.e.* when $2\gamma R_jS_j$ is large, (11) is consistent with the simple algebraic rule of Karle (1983) if the distribution coefficient relevant to the content of heavy atoms, $2\sigma_{3H}\sigma_{2H}^{-3}$, is ignored. Since the formula of Fortier *et al.* (1984*a*) contains mixed terms of R and Δ besides the RRR and $\Delta\Delta\Delta$ terms, Fortier *et al.* concluded that the difference in the A values between Hauptman's distribution and Karle's simple rule is caused by these mixed terms. Our approach shows that there are no mixed terms in (11) and therefore the difference between Hauptman's distribution and Karle's simple rule is chiefly due to the distribution coefficient, which is missing in the latter, rather than the mixed terms.

On the other hand, according to Karle (1989), it is not necessary to know information on the heavy atoms in order to apply the simple algebraic rule. Now it becomes clear that not requiring any knowledge concerning the heavy atoms is not an intrinsic advantage of the algebraic formula but a result of the absence of the distribution coefficient. Because $\sigma_{3H}\sigma_{2H}^{-3/2} \simeq N_H^{-1/2}$, where N_H is the statistically equivalent number of heavy atoms in the unit cell, there is an optimal amount of heavy-atom substitution, as pointed out by Fortier *et al.*

(1984*a*), which leads to sufficiently large $\sigma_{3H}\sigma_{2H}^{-3/2}$ and $|\Delta_{1R}\Delta_{2R}\Delta_{3R}|$. In this regard, the distribution coefficient is important for obtaining reliability evaluation from A values although in some instances both the probability and algebraic formulae give identical results.

Of the three kinds of quantities in (11), σ, C and F (and R), the parameters σ and C may be modified in order to obtain more accurate estimates. If the zero-angle atomic scattering factor Z_j in (8) is replaced by the scattering factor f_j, which is a function of $|\mathbf{H}|$, $|\mathbf{K}|$ or $|\mathbf{L}|$, *i.e.*

$$\begin{aligned}
\sigma_{3P} &= \sum_P f_j(\mathbf{H}) f_j(\mathbf{K}) f_j(\mathbf{L}),\\
\sigma_{2P}^{3/2} &= \left[\sum_P f_j^2(\mathbf{H}) \sum_P f_j^2(\mathbf{K}) \sum_P f_j^2(\mathbf{L})\right]^{1/2},\\
\sigma_{3H} &= \sum_H f_j(\mathbf{H}) f_j(\mathbf{K}) f_j(\mathbf{L}),\\
\sigma_{2H}^3 &= \sum_H f_j^2(\mathbf{H}) \sum_H f_j^2(\mathbf{K}) \sum_H f_j^2(\mathbf{L}),
\end{aligned} \tag{13}$$

then (11) is the same as the result of Giacovazzo *et al.* derived from a different route (Giacovazzo, Cascarano & Zheng, 1988; Giacovazzo, Siliqi & Ralph, 1994). The fact that the same result comes out from different derivation routes makes the A term more believable as a reliability measurement for probability estimation of the 3PSIs.

Since $\mathbf{F}_D = \mathbf{F}_P + \mathbf{F}_H$ and C_{jR} or C_{jS} is the expected value of $\cos(\psi_j - \varphi_j)$ (Fortier, Moore & Fraser, 1985), where $\psi_j - \varphi_j = \theta_{jPD}$ is the angle between $\mathbf{F}_P$ and $\mathbf{F}_D$, when the heavy-atom structure is known, C_{jR} or C_{jS} can be calculated according to

$$C_{jR} \text{ or } C_{jS} = (F_{jP}^2 + F_{jD}^2 - F_{jH}^2)/2F_{jP}F_{jD}. \tag{14}$$

Combining (9) and (14), we obtain a formula for A_i incorporating the heavy-atom structure information,

$$A_i \simeq 2\sigma_{3H}\sigma_{2H}^{-3/2} E'_{1H} E'_{2H} E'_{3H}, \tag{15}$$

where

$$\begin{aligned}
E'_{jH} &= -(F_{jP}^2 + F_{jH}^2 - F_{jD}^2)F_{jH}/2F_{jP}F_{jH}\sigma_{2H}^{1/2}\\
&= E_{jH}\cos\theta_{jPH}, \quad j = 1, 2, 3,
\end{aligned} \tag{16}$$

if the jth phase of the invariant is φ, and

$$\begin{aligned}
E'_{jH} &= (F_{jD}^2 + F_{jH}^2 - F_{jP}^2)F_{jH}/2F_{jD}F_{jH}\sigma_{2H}^{1/2}\\
&= E_{jH}\cos\theta_{jDH}, \quad j = 1, 2, 3,
\end{aligned} \tag{17}$$

if the jth phase of the invariant is ψ. In (16), θ_{jPH} is the angle between the structure-factor vectors of the native and heavy-atom structures, E_H the normalized structure-factor magnitude contributed from heavy atoms and thus E'_H is the projection of the normalized structure-factor vector of the heavy-atom structure on the structure-factor vector of protein structure. Similarly, in (17), θ_{jDH} is the angle between the structure-factor vectors of the derivative and heavy-atom structures, thus E'_H is the projection of the normalized structure-factor vector of the heavy-atom structure on the structure-factor vector of the derivative structure.

2.2. *The case of a native protein and two heavy-atom derivatives*

According to Fortier, Weeks & Hauptman (1984*b*), to which the notation in this section is referred except where stated, the conditional probability distribution for 27 3PSIs in the case of a native and two derivatives is given by

$$\begin{aligned}
&P_i(\Omega_i|R_1, R_2, R_3, S_1, S_2, S_3, T_1, T_2, T_3)\\
&\quad \simeq (1/K_i)\exp(A_i\cos\Omega_i), \quad i = 1, \ldots, 27.
\end{aligned} \tag{18}$$

A_i can be written as

$$\begin{aligned}
A_i = 2\{&\beta_1 C_{1R}C_{2R}C_{3R}R_1R_2R_3 + \beta_2[C_{1R}C_{2R}C_{3S}R_1R_2S_3\\
&+ C_{1R}C_{2S}C_{3R}R_1S_2R_3 + C_{1S}C_{2R}C_{3R}S_1R_2R_3]\\
&+ \beta_3[C_{1R}C_{2R}C_{3T}R_1R_2T_3 + C_{1R}C_{2T}C_{3R}R_1T_2R_3\\
&+ C_{1T}C_{2R}C_{3R}T_1R_2R_3] + \beta_4[C_{1R}C_{2S}C_{3S}R_1S_2S_3\\
&+ C_{1S}C_{2R}C_{3S}S_1R_2S_3 + C_{1S}C_{2S}C_{3R}S_1S_2R_3]\\
&+ \beta_6[C_{1R}C_{2T}C_{3T}R_1T_2T_3 + C_{1T}C_{2R}C_{3T}T_1R_2T_3\\
&+ C_{1T}C_{2T}C_{3R}T_1T_2R_3] + \beta_7 C_{1S}C_{2S}C_{3S}S_1S_2S_3\\
&+ \beta_{10}C_{1T}C_{2T}C_{3T}T_1T_2T_3\},
\end{aligned} \tag{19}$$

where

$$C_{jR} = 1, \quad C_{jS} = I_1(x_1)/I_0(x_1), \quad C_{jT} = I_1(x_2)/I_0(x_2) \tag{20}$$

if the jth phase of the invariant is φ,

$$\begin{aligned}
&C_{jR} = I_1(x_1)/I_0(x_1), \quad C_{jS} = 1,\\
&C_{jT} = I_1(x_1)I_1(x_2)/I_0(x_1)I_0(x_2)
\end{aligned} \tag{21}$$

if the jth phase of the invariant is ψ,

$$\begin{aligned}
&C_{jR} = I_1(x_2)/I_0(x_2), \quad C_{jS} = I_1(x_1)I_1(x_2)/I_0(x_1)I_0(x_2),\\
&C_{jT} = 1
\end{aligned} \tag{22}$$

if the jth phase of the invariant is ξ and

$$x_1 = 2\gamma_1 R_j S_j, \quad x_2 = 2\gamma_2 R_j T_j, \quad j = 1, 2, 3.$$

The γ and β parameters are defined by equations (3.25) and (3.26) of Fortier *et al.* (1984*b*). Equations (18) and (19) require that the heavy atoms of the two derivatives occupy different positions in the unit cell.

Assuming that the atomic content of the first (D_1) or second derivative (D_2) equals the atomic content of the native protein (P) plus the heavy-atom content (H_1 or H_2), respectively, we define

$$\sigma_{3H_1} = \alpha_{030} - \alpha_{300} = \sum_{H_1} Z_j^3, \quad \sigma_{2H_1} = \alpha_{020} - \alpha_{200} = \sum_{H_1} Z_j^2,$$
$$\sigma_{3H_2} = \alpha_{003} - \alpha_{300} = \sum_{H_2} Z_j^3, \quad \sigma_{2H_2} = \alpha_{002} - \alpha_{200} = \sum_{H_2} Z_j^2. \tag{23}$$

A similar approach to that in §2.1 gives

$$\begin{aligned} A_i \simeq\ & 2\sigma_{3H_1}\sigma_{2H_1}^{-3}(C_{1S}F_{1D_1} - C_{1R}F_{1P})(C_{2S}F_{2D_1} - C_{2R}F_{2P}) \\ & \times (C_{3S}F_{3D_1} - C_{3R}F_{3P}) \\ & + 2\sigma_{3H_2}\sigma_{2H_2}^{-3}(C_{1T}F_{1D_2} - C_{1R}F_{1P}) \\ & \times (C_{2T}F_{2D_2} - C_{2R}F_{2P})(C_{3T}F_{3D_2} - C_{3R}F_{3P}), \end{aligned} \tag{24}$$

where

$$F_{jP} = \alpha_{200}^{1/2}R_j, \quad F_{jD_1} = \alpha_{020}^{1/2}S_j, \quad F_{jD_2} = \alpha_{002}^{1/2}T_j, \quad j = 1, 2, 3,$$

or

$$\begin{aligned} A_i \simeq\ & 2\sigma_{3H_1}\sigma_{2H_1}^{-3/2}\Delta_{1H_1}\Delta_{2H_1}\Delta_{3H_1} \\ & + 2\sigma_{3H_2}\sigma_{2H_2}^{-3/2}\Delta_{1H_2}\Delta_{2H_2}\Delta_{3H_2}, \end{aligned} \tag{25}$$

where

$$\Delta_{jH_1} = (C_{jS}F_{jD_1} - C_{jR}F_{jP})/\sigma_{2H_1}^{1/2}$$

and

$$\Delta_{jH_2} = (C_{jT}F_{jD_2} - C_{jR}F_{jP})/\sigma_{2H_2}^{1/2}, \quad j = 1, 2, 3, \tag{26}$$

are the modified normalized structure-factor magnitudes of the first and second heavy-atom structures, respectively. For the 3PSI $\omega_1 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$,

$$\begin{aligned} \Delta_{jH_1} &= (C_{jS}F_{jD_1} - F_{jP})/\sigma_{2H_1}^{1/2}, \\ \Delta_{jH_2} &= (C_{jT}F_{jD_2} - F_{jP})/\sigma_{2H_2}^{1/2}, \quad j = 1, 2, 3. \end{aligned} \tag{27}$$

It is interesting to note that (25) consists simply of the sum of two parts corresponding to the contributions from heavy atoms in the first and second derivatives, respectively. We are aware from this result that it is possible to deduce a formula for A_i in the case of a native protein and multiple derivatives.

When the heavy-atom structures for the two derivatives are known, A_i is given by

$$\begin{aligned} A_i \simeq\ & 2\sigma_{3H_1}\sigma_{2H_1}^{-3/2}E'_{1H_1}E'_{2H_1}E'_{3H_1} \\ & + 2\sigma_{3H_2}\sigma_{2H_2}^{-3/2}E'_{1H_2}E'_{2H_2}E'_{3H_2}, \end{aligned} \tag{28}$$

where

$$\begin{aligned} E'_{jH_1} &= E_{jH_1}\cos\theta_{jPH_1}, \\ E'_{jH_2} &= E_{jH_2}\cos\theta_{jPH_2}, \quad j = 1, 2, 3, \end{aligned} \tag{29}$$

if the jth phase of the invariant is φ,

$$\begin{aligned} E'_{jH_1} &= E_{jH_1}\cos\theta_{jD_1H_1}, \\ E'_{jH_2} &= E_{jH_2}\cos\theta_{jPH_2}\cos\theta_{jPD_1}, \quad j = 1, 2, 3, \end{aligned} \tag{30}$$

if the jth phase of the invariant is ψ, and

$$\begin{aligned} E'_{jH_1} &= E_{jH_1}\cos\theta_{jPH_1}\cos\theta_{jPD_2}, \\ E'_{jH_2} &= E_{jH_2}\cos\theta_{jD_2H_2}, \quad j = 1, 2, 3, \end{aligned} \tag{31}$$

if the jth phase of the invariant is ξ.

3. The probabilistic formulae for estimating the 4PSIs

Recently, we derived the probability distribution of the 4PSIs for a pair of isomorphous structures. Only the main results are given here in order to compare the formulae with those for the 3PSIs. Details of the derivation and practical applications will be published separately.

For a pair of isomorphous structures, the conditional probability distribution of the 4PSIs $\omega_1 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}} + \varphi_{\mathbf{M}}$, where $\mathbf{H} + \mathbf{K} + \mathbf{L} + \mathbf{M} = \mathbf{0}$, given the eight structure-factor magnitudes $|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, $|E_{\mathbf{L}}|$, $|E_{\mathbf{M}}|$, $|G_{\mathbf{H}}|$, $|G_{\mathbf{K}}|$, $|G_{\mathbf{L}}|$, $|G_{\mathbf{M}}|$, is given by

$$\begin{aligned} & P_1(\Omega_1 | R_1, R_2, R_3, R_4, S_1, S_2, S_3, S_4) \\ & \quad \simeq (1/K_1)\exp(A_1\cos\Omega_1), \end{aligned} \tag{32}$$

where

$$\begin{aligned} K_1 &= 2\pi I_0(A_1), \\ A_1 &= 2[\beta_0 R_1R_2R_3R_4 - \beta_1(C_{1S}S_1R_2R_3R_4 + C_{2S}R_1S_2R_3R_4 \\ &\quad + C_{3S}R_1R_2S_3R_4 + C_{4S}R_1R_2R_3S_4) \\ &\quad + \beta_2(C_{1S}C_{2S}S_1S_2R_3R_4 + C_{1S}C_{3S}S_1R_2S_3R_4 \\ &\quad + C_{1S}C_{4S}S_1R_2R_3S_4 + C_{2S}C_{3S}R_1S_2S_3R_4 \\ &\quad + C_{2S}C_{4S}R_1S_2R_3S_4 + C_{3S}C_{4S}R_1R_2S_3S_4) \\ &\quad - \beta_3(C_{1S}C_{2S}C_{3S}S_1S_2S_3R_4 + C_{1S}C_{2S}C_{4S}S_1S_2R_3S_4 \\ &\quad + C_{1S}C_{3S}C_{4S}S_1R_2S_3S_4 + C_{2S}C_{3S}C_{4S}R_1S_2S_3S_4) \\ &\quad + \beta_4 C_{1S}C_{2S}C_{3S}C_{4S}S_1S_2S_3S_4], \end{aligned} \tag{33}$$

$$\begin{aligned} R_1 = |E_{\mathbf{H}}|, \quad R_2 = |E_{\mathbf{K}}|, \quad R_3 = |E_{\mathbf{L}}|, \quad R_4 = |E_{\mathbf{M}}|, \\ S_1 = |G_{\mathbf{H}}|, \quad S_2 = |G_{\mathbf{K}}|, \quad S_3 = |G_{\mathbf{L}}|, \quad S_4 = |G_{\mathbf{M}}| \end{aligned} \tag{34}$$

and $C_{jS} = I_1(2\gamma R_jS_j)/I_0(2\gamma R_jS_j)$ is the ratio of two modified Bessel functions.

In the case of a native protein and a heavy-atom derivative, if the atomic content of the derivative (D) equals the atomic content of the native protein (P) plus the heavy-atom content (H), the parameters γ and β_j, $j = 0, 1, 2, 3, 4$, can be greatly simplified,:

$$\begin{aligned} \gamma &= \alpha_{20}^{1/2}\alpha_{02}^{1/2}/(\alpha_{02} - \alpha_{20}), \\ \beta_0 &= \alpha_{40}/\alpha_{20}^2 + (\alpha_{04} - \alpha_{40})\alpha_{20}^2/(\alpha_{02} - \alpha_{20})^4, \\ \beta_1 &= (\alpha_{04} - \alpha_{40})\alpha_{20}^{3/2}\alpha_{02}^{1/2}/(\alpha_{02} - \alpha_{20})^4, \\ \beta_2 &= (\alpha_{04} - \alpha_{40})\alpha_{20}\alpha_{02}/(\alpha_{02} - \alpha_{20})^4, \\ \beta_3 &= (\alpha_{04} - \alpha_{40})\alpha_{20}^{1/2}\alpha_{02}^{3/2}/(\alpha_{02} - \alpha_{20})^4, \\ \beta_4 &= (\alpha_{04} - \alpha_{40})\alpha_{02}^2/(\alpha_{02} - \alpha_{20})^4, \end{aligned} \tag{35}$$

where α_{mn} is defined by equation (1.3) of Hauptman (1982). Substituting (35) into (33) and noting that $F_{jD} = \alpha_{02}^{1/2} S_j$ and $F_{jP} = \alpha_{20}^{1/2} R_j$, we have

$$A_1 = 2\sigma_{4P}\sigma_{2P}^{-2} R_1 R_2 R_3 R_4 + 2\sigma_{4H}\sigma_{2H}^{-2}\Delta_1\Delta_2\Delta_3\Delta_4, \quad (36)$$

where

$$\Delta_j = (C_{jS}F_{jD} - F_{jP})/\sigma_{2H}^{1/2}, \quad j = 1, 2, 3, 4, \quad (37)$$

$$\sigma_{4P} = \sum_P Z_j^4, \quad \sigma_{4H} = \sum_H Z_j^4, \quad (38)$$

and σ_{2P} and σ_{2H} are defined by (8). Because $\sigma_{4P}\sigma_{2P}^{-2} \ll \sigma_{4H}\sigma_{2H}^{-2}$, the first term of (36) is negligible. Accordingly,

$$A_1 \simeq 2\sigma_{4H}\sigma_{2H}^{-2}\Delta_1\Delta_2\Delta_3\Delta_4. \quad (39)$$

Equation (39) is one of our major results. Clearly, (39) is analogous to (11) in properties: (*a*) the reliability parameter A_1 depends mainly on the contribution from heavy atoms in the derivative; (*b*) reliable estimates can be obtained even when the structure factors themselves are small provided that the $|\Delta_j|$ values are large; (*c*) since the Δ_j's are signed values, both 0 and 180° estimates are obtainable through (32).

4. General probabilistic formulae for *n*-phase structure invariants in the case of a native protein and *m* heavy-atom derivatives

For n ($n \geq 3$) reciprocal-lattice vectors $\mathbf{H}$ satisfying $\sum_{j=1}^{n} \mathbf{H}_j = 0$, in the case of a native protein (P) and m heavy-atom derivatives ($D_k, k = 1, \ldots, m$), there are Q n-phase structure invariants. Assume that the heavy atoms ($H_k, k = 1, \ldots, m$) of the m derivatives are located in different positions in the unit cell. Given the $n \times (1 + m)$ structure-factor magnitudes related to the native and the m derivatives, which are represented by a group of non-negative numbers R_{jP}, R_{jD_k}, $j = 1, \ldots, n$, $k = 1, \ldots, m$, the conditional probability distribution of the n-phase structure invariants ω_i, $i = 1, \ldots, Q$, can be directly deduced from the results introduced above:

$$P_i(\Omega_i | R_{jP}, R_{jD_k}; j = 1, \ldots, n; k = 1, \ldots, m) \simeq (1/K_i)\exp(A_i \cos\Omega_i), \quad i = 1, \ldots, Q, \quad (40)$$

where

$$K_i = 2\pi I_0(A_i),$$
$$A_i \simeq 2\sum_{k=1}^{m} \sigma_{nH_k}\sigma_{2H_k}^{-n/2} \prod_{j=1}^{n} \Delta_{jH_k}, \quad (41)$$
$$\Delta_{jH_k} = (C_{jD_k}F_{jD_k} - C_{jP}F_{jP})/\sigma_{2H_k}^{1/2}, \quad j = 1, \ldots, n; \quad k = 1, \ldots, m. \quad (42)$$

The σ parameters have similar definitions to those in (8), (23) and (38). C_{jP} and C_{jD_k}, $j = 1, \ldots, n$, are obtained by comparing the subscript P or D_k of the jth C with the jth phase of the invariant. If they both correspond to the native or both to the same derivative, then C_{jP} or $C_{jD_k} = 1, j = 1, \ldots, n$. If one corresponds to the native and the other to the derivative, then

$$C_{jP} \text{ or } C_{jD_k} = I_1(2\gamma_k R_{jP} R_{jD_k})/I_0(2\gamma_k R_{jP} R_{jD_k}), \quad j = 1, \ldots, n, \quad (43)$$

where

$$\gamma_k = \sigma_{2P}^{1/2}\sigma_{2D_k}^{1/2}/\sigma_{2H_k}. \quad (44)$$

If they correspond to the different derivatives $k1$ and $k2$, respectively, then

$$C_{jD_{k1}} \text{ or } C_{jD_{k2}} = I_1(2\gamma_{k1} R_{jP} R_{jD_{k1}}) I_1(2\gamma_{k2} R_{jP} R_{jD_{k2}}) \times I_0(2\gamma_{k1} R_{jP} R_{jD_{k1}})^{-1} I_0(2\gamma_{k2} R_{jP} R_{jD_{k2}})^{-1}, \quad j = 1, \ldots, n. \quad (45)$$

5. Test calculations

The experimental data for the protein cytochrome c_{550}, space group $P2_12_12_1$, molecular weight $\simeq$ 14 500, and its $PtCl_4^{2-}$ derivative (Timkovich & Dickerson, 1976) were used for test calculations to examine the factors affecting the accuracy of the 3PSI estimates. The number of measured independent reflections up to 2.5 Å resolution is 2993 for the native and 2807 for the derivative.

To compare the estimate results, tests 1, 2 and 3 are designed for various subsets of the reflections selected by different thresholds for R and $|\Delta|$ values. The calculations were performed using (11), *i.e.* only the 3PSI $\omega_1 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$ was estimated. The definitions of the subsets and the results for tests 1, 2 and 3 are given in Table 1. It can be seen that, for the reflections with the largest $|\Delta|$ values, test 3 gives not only a higher accuracy but more triplet relationships when the $|A_{\min}|$ value is given. In each case, the average phase errors $\langle|\phi_3 - \omega|\rangle$ decrease with increasing $|A|$ values but there is a rebound of the errors at the top of the $|A|$ values. This is probably due to the scattering effect of disordered solvent molecules on the reflections with large $|\Delta|$ values at low resolution.

Test 4 was done, as shown in Table 2, using the same reflections as those for test 3 in order to judge whether the effectiveness of (11) can be enhanced by substituting (13) for (8) to calculate the σ parameters. For the convenience of comparison, the results of test 3 were reaccumulated in Table 2 according to various $|A_{\min}|$ values, which were chosen so as to allow the N_3 values to approximately equal those of test 4. Comparison of test 3 with test 4 suggests that the substitution of (13) for (8), *i.e.* atomic number Z_j is replaced by scattering factor $f_j(\mathbf{H})$ in the σ parameters, only results in an overall rise of the $|A|$ values and has little effect on the estimate accuracy. Such an effect is quite different to

Table 1. *Statistical results of the 3PSIs estimated via (11) from the experimental data of cytochrome c_{550}*

N_3 is the number of the 3PSIs having $|A| > |A_{min}|$, % is the percentage of the 3PSIs whose cosine signs are correctly estimated, ϕ_3 (°) is the true value of the 3PSI and ω (0 or 180°) is its estimated value. The subsets of reflections used in the calculations are defined as follows. Test 1: 619 reflections with $R > 1.49$; test 2: 631 reflections with $R > 1.20$ and $|\Delta_R| > 0.35$; test 3: 592 reflections with $|\Delta_R| > 0.70$.

Test 1				Test 2			
$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$	$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$
0.0	51209	54.4	84.7	0.0	51207	63.6	74.0
0.5	7375	70.0	66.3	0.5	18638	71.2	64.8
1.0	1903	79.3	52.2	1.0	5760	78.3	54.4
1.5	728	88.0	39.6	1.5	2324	85.2	44.7
2.0	309	91.9	34.6	2.0	1053	87.6	39.2
3.0	83	86.7	41.2	3.0	301	88.4	34.4
4.0	24	100.0	15.5	4.0	109	92.7	28.7
5.0	7	100.0	0.0	5.0	39	87.2	34.8

Test 3			
$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$
0.0	51726	77.0	55.5
0.5	49212	77.6	54.6
1.0	29024	82.1	47.7
1.5	14600	86.4	40.2
2.0	7686	87.4	36.9
3.0	2718	85.2	37.8
4.0	1152	76.6	48.9
5.0	591	69.9	57.3

that in the small-molecule case observed by Giacovazzo, Cascarano & Zheng (1988). Calculations similar to tests 1 to 4 were also carried out on the error-free diffraction data of cytochrome c_{550} and its $PtCl_4^{2-}$ derivative, which were obtained from the known atomic coordinates to a resolution of 2.5 Å (total 4159 structure factors). The results confirm those from the experimental data but, as expected, they have higher accuracy and no error rebound at the top $|A|$ value.

The role of the reflections with large $|\Delta|$ values, as shown in Table 1, has already been emphasized by Giacovazzo, Siliqi & Ralph (1994) for the direct crystal structure solution of proteins. In their successful phasing procedure (Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995), a small set of reflections with large $|\Delta|$ and R values, just like the subset in test 2, was first phased and then used as seeds for subsequent phase expansion. The condition 'large $|\Delta|$' selects the reflections whose phase values may be reliably estimated and the condition 'large R' is used in order to guarantee a valuable contribution to Fourier synthesis once the reflection is phased. However, the limitations of their procedure are:

(*a*) the seed set does not include all of the reflections with the largest $|\Delta|$ values owing to the restriction of the R threshold;

(*b*) not all types of the invariant but only the 'pure' invariant $\omega_1 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$ is used in the phasing procedure.

According to Table 1, the subset of reflections in test 3 seems to be more advisable than that in test 2 as seed set for the sake of accuracy. But the problem is that test 3 may include some reflections with small R values because only the $|\Delta|$ value is considered as the selecting condition and these weak reflections usually have large phase errors, which propagate easily to the other reflections during phase expansion. In order to solve this problem, we consider ways for making full use of all eight 3PSIs. We note the fact that for the reflection having larger $|\Delta|$ value, even if the R value is rather small, the S value can be large and *vice versa*. Accordingly, the phases, φ or ψ, associated with the larger structure-factor magnitudes can be used to constitute the triplet relationship corresponding to one of the eight 3PSIs in (1). For example, if $R_1 < S_1$, $R_2 > S_2$ and $R_3 < S_3$, then the 3PSI $\omega_7 = \psi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}}$ is actively used in the phasing process and the corresponding A_7 value is calculated *via* (9),

$$A_7 = 2\sigma_{3P}\sigma_{2P}^{-3/2}C_{1R}C_{3R}R_1R_2R_3 + 2\sigma_{3H}\sigma_{2H}^{-3/2}\Delta_{1S}\Delta_{2R}\Delta_{3S}.$$

This enables us to obtain a set of the most reliable 3PSIs among the reflections with the largest $|\Delta|$ values. Such invariants are here called 'large-modulus invariants'.

Table 2. *A comparison between the estimated results of the 3PSIs with the σ parameters calculated via (8) (test 3) and via (13) (test 4) from the experimental data of cytochrome c_{550}*

Test 3				Test 4			
$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$	$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$
0.00	51726	77.0	55.5	0.0	51726	77.0	55.5
0.65	44308	78.8	53.0	1.0	44402	78.8	53.0
0.99	29416	82.1	47.8	1.5	29463	82.2	47.6
1.33	18392	85.0	42.5	2.0	18312	84.6	43.1
1.68	11431	87.1	38.6	2.5	11479	86.0	40.3
2.03	7454	87.2	36.9	3.0	7482	86.2	38.7
2.74	3477	86.3	36.7	4.0	3484	85.0	39.2
3.42	1853	81.5	41.6	5.0	1845	80.3	44.6

Table 3 lists the results for the pure invariants (test 5) and the large-modulus invariants (tests 6 and 7) estimated from the error-free data. Indeed, the comparison of test 6 with test 5 indicates that a remarkable increase in accuracy can be achieved by using the large-modulus invariants. In the calculations of A values for the large-modulus invariants, the parameters C_{jR} and C_{jS} may no longer be negligible since some of the $2\gamma R_jS_j$ may happen to be small. In test 6, C_{jR} and C_{jS} were calculated from (4) and (5) while they were assigned to have a value of 1.0 in test 7. It is observed from the comparison of test 6 with test 7 that the number of invariants (N_3) for test 7, where C_{jR} and C_{jS} are ignored, is smaller than that for test 6 at the same accuracy level, especially for those with $|A_{min}| > 2.0$. Therefore, the use of C_{jR} and C_{jS} is advisable for the large-modulus invariants.

Table 3. *A comparison between the estimated results of the pure invariants via (11) (test 5) and the large-modulus invariants via (9) (tests 6 and 7) from the error-free data of cytochrome c_{550}*

In test 5, 601 reflections with $R > 1.0$ and $|\Delta_R| > 0.6$ were used. In test 6, 679 reflections with $|\Delta_R|$ or $|\Delta_S| > 1.0$ were used and C_{jR} or C_{jS} was calculated *via* (4) or (5). In test 7, the same reflections were used as those in test 6 but C_{jR} or $C_{jS} = 1.0$.

Test 5				Test 6			
$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$	$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$
0.0	40986	86.6	45.1	0.0	94551	99.9	16.6
1.0	21546	94.6	33.6	1.0	94545	99.9	16.6
2.0	7237	98.0	23.2	2.0	65064	100.0	13.6
3.0	2759	100.0	15.3	3.0	24206	100.0	9.2
4.0	1025	100.0	10.0	4.0	7758	100.0	5.5
5.0	369	100.0	6.3	5.0	2464	100.0	3.0

Test 7			
$\|A_{min}\|$	N_3	%	$\langle\|\phi_3 - \omega\|\rangle$
0.0	94551	99.9	16.6
1.0	93566	99.9	16.4
2.0	55920	100.0	13.1
3.0	19257	100.0	9.1
4.0	6026	100.0	5.2
5.0	1822	100.0	3.2

6. Concluding remarks

Through simple mathematical manipulations, we have simplified the probabilistic formulae for eight 3PSIs in the case of a native protein and a heavy-atom derivative (Hauptman, 1982) and for 27 3PSIs in the case of a native and two derivatives (Fortier, Weeks & Hauptman, 1984*b*). The probabilistic formula for the 4PSIs, when simplified with a similar approach, is comparable in its properties with that for the 3PSIs. The analysis directly leads to a general expression of probabilistic estimation for the n-phase structure invariants in the case of a native and m derivatives.

A method to estimate the large-modulus invariants is proposed, which remarkably improves the accuracy. The advantage of the method is to make use of the information concerning both the magnitudes and the phases of the structure factors of the derivative while only the magnitude information is utilized when the pure invariant $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$ is involved alone. Moreover, since only the 3PSI associated with three large structure-factor moduli, rather than all eight 3PSIs, is calculated for each triplet of **H**, **K** and **L**, the method is not time consuming for computation. The limitation of the method is that the reflection set required by the large-modulus invariants is a mixture of the reflections from the native and derivative and may not contain enough native reflections to produce an interpretable electron-density map for the protein. So we suggest a phasing procedure in two steps.

(i) A small set of reflections with the largest $|\Delta_R|$ or $|\Delta_S|$ values is phased by a tangent multisolution process using the large-modulus invariants. The phases to be assigned could be either φ or ψ depnding on whether R or S is large. This requires a common origin and enantiomorph definition for the native and derivative. In addition, many reliable seminvariant phases could be obtained by a modified Σ_1 formula for the case of isomorphous replacement (Hu & Liu, 1995; Liu & Hu, 1996).

(ii) The phases obtained above are used as seeds for further phase expansion to determine the other phases $\varphi_{\mathbf{H}}$ by constituting the triplet sets of the 3PSIs: $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$, $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}}$, $\varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \varphi_{\mathbf{L}}$, $\varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}}$, where the reflections **K** and **L** with the largest $|\Delta|$ values have been phased in (i) and the reflection **H** from the native protein has a sufficiently large R value for a useful contribution to the Fourier map.

There may still be some reflections with $|\Delta| \simeq 0$ but large R values that cannot be phased by this procedure. These reflections are not negligible especially for large protein structures. In this case, the diffraction data from two or more heavy-atom derivatives are necessary and (25) or (41) should play a role in the phasing process.

This work was supported by the National Natural Science Foundation of China (no. 29573127).

References

Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.
Fortier, S., Weeks, C. M. & Hauptman, H. (1984*a*). *Acta Cryst.* A**40**, 544–548.
Fortier, S., Weeks, C. M. & Hauptman, H. (1984*b*). *Acta Cryst.* A**40**, 646–651.
Giacovazzo, C., Cascarano, G. & Zheng, C.-D. (1988). *Acta Cryst.* A**44**, 45–51.
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.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Hauptman, H., Potter, S. & Weeks, C. M. (1982). *Acta Cryst.* A**38**, 294–300.
Hu, N.-H. & Liu, Y.-S. (1995). *Acta Cryst.* A**51**, 520–524.
Karle, J. (1983). *Acta Cryst.* A**39**, 800–805.
Karle, J. (1989). *Acta Cryst.* A**45**, 765–781.
Liu, Y.-S. & Hu, N.-H. (1996). *Acta Cryst.* A**52**, 56–61.
Timkovich, R. & Dickerson, R. E. (1976). *J. Biol. Chem.* **251**, 4033–4046.