Acta Cryst. (1984). A**40**, 646-651

On Integrating the Techniques of Direct Methods and Isomorphous Replacement. III. The Three-Phase Invariant for the Native and Two-Derivative Case*

BY SUZANNE FORTIER

Department of Chemistry, Queen's University, Kingston, Canada K7L 3N6

AND CHARLES M. WEEKS AND HERBERT HAUPTMAN

Medical Foundation of Buffalo, Inc., 73 High Street, Buffalo, NY 14203, USA

(*Received* 9 *November* 1983; *accepted* 30 *May* 1984)

Abstract

The probabilistic theory of the three-phase structure invariant for a triplet of isomorphous structures is worked out. In particular, when diffraction data are available for a native protein and two derivatives, the conditional distributions of the three-phase structure invariants, given the nine magnitudes in their first neighbourhoods, are derived for the special case that the heavy atoms of the two derivatives are located in different positions in the unit cell. The distributions have the form $P(\Omega) = (1/K)\exp(A\cos\Omega)$, where the parameters K and A are functions of the nine magnitudes in the first neighborhood. In the favorable case that the variance of a distribution happens to be small, a reliable estimate, 0 or π, of the invariant is obtained. An example shows that these distributions, which employ simultaneously the diffraction data from a triple of isomorphous structures, yield more accurate estimates for the three-phase structure invariants than are obtainable from earlier distributions, which employ diffraction data from only a pair of isomorphous structures [Hauptman (1982). *Acta. Cryst.* A**38**, 289–294]. Unique origin and enantiomorph specification in direct-methods applications to all three structures is an advantage of the present approach.

1. Introduction

In the last ten years, direct methods have been increasingly used in problems of macromolecular structure determination to supplement existing methods of isomorphous replacement, anomalous dispersion and molecular replacement. They have been shown to be a valuable addition, particularly for phase extension and refinement and for the determination of heavy-atom positions in isomorphous derivatives. More recently, a formal mathematical integration of the techniques of direct methods and isomorphous replacement has been undertaken for a pair of isomorphous structures (Hauptman, 1982). It is naturally to be anticipated that a more rigorous attempt to combine these different techniques will enhance the scope of direct methods in macromolecular structure determinations. This anticipation cannot be fully confirmed until the effects of errors in real data and of imperfect isomorphism have been satisfactorily assessed through extensive calculations and analyses. The theoretical validity of the approach, however, has been fully confirmed by initial calculations on error-free diffraction data from a moderate-size protein, cytochrome C_{550}, $M_r = 14\,500$ (Hauptman, Potter & Weeks, 1982). The calculations have shown that it is possible to estimate reliably several thousands of invariants, establishing the potential importance of integrated techniques of direct methods and isomorphous replacement. The present paper extends this recent work to triplets of isomorphous structures in the expectation that a further strengthening will result.

2. Definitions

If e_j, f_j and g_j denote atomic structure factors for a corresponding triplet of isomorphous structure in $P1$, then respective normalized structure factors $E_{\mathbf{H}}$, $F_{\mathbf{H}}$ and $G_{\mathbf{H}}$ are defined by

$$E_{\mathbf{H}} = |E_{\mathbf{H}}|\exp(i\varphi_{\mathbf{H}}) = \alpha_{200}^{-1/2}\sum_{j=1}^{N} e_j \exp(2\pi\mathbf{H}\cdot\mathbf{r}_j) \tag{2.1}$$

$$F_{\mathbf{H}} = |F_{\mathbf{H}}|\exp(i\psi_{\mathbf{H}}) = \alpha_{020}^{-1/2}\sum_{j=1}^{N} f_j \exp(2\pi i\mathbf{H}\cdot\mathbf{r}_j) \tag{2.2}$$

$$G_{\mathbf{H}} = |G_{\mathbf{H}}|\exp(i\xi_{\mathbf{H}}) = \alpha_{002}^{-1/2}\sum_{j=1}^{N} g_j \exp(2\pi\mathbf{H}\cdot\mathbf{r}_j), \tag{2.3}$$

where

$$\alpha_{mno} = \sum_{j=1}^{N} e_j^m f_j^n g_j^o, \tag{2.4}$$

* Presented at the Am. Crystallogr. Assoc. winter meeting, National Bureau of Standards, Gaithersburg, Maryland, 29 March–2 April 1982, Abstr. PC1.

0108-7673/84/060646-06$01.50

some of the e_j's, f_j's or g_j's may be zero (or negative, in the neutron diffraction case) and $\mathbf{r}_j$ is the position vector of the atom labelled j.

The mathematical formalism used throughout this paper has been introduced recently (Hauptman, 1982) and only material specific to the present work will be described in detail.

3. The probabilistic theory of the three-phase structure invariants

3.1. *The structure invariants*

For a triplet of isomorphous structures, given

$$\mathbf{H}+\mathbf{K}+\mathbf{L}=0, \tag{3.1}$$

there exist 27 three-phase structure invariants of which there are ten kinds,

$$\begin{aligned}
\omega_1&=\varphi_{\mathbf{H}}+\varphi_{\mathbf{K}}+\varphi_{\mathbf{L}}\\
\omega_2&=\varphi_{\mathbf{H}}+\varphi_{\mathbf{K}}+\psi_{\mathbf{L}}\\
\omega_3&=\varphi_{\mathbf{H}}+\varphi_{\mathbf{K}}+\xi_{\mathbf{L}}\\
\omega_4&=\varphi_{\mathbf{H}}+\psi_{\mathbf{K}}+\psi_{\mathbf{L}}\\
\omega_5&=\varphi_{\mathbf{H}}+\psi_{\mathbf{K}}+\xi_{\mathbf{L}}\\
\omega_6&=\varphi_{\mathbf{H}}+\xi_{\mathbf{K}}+\xi_{\mathbf{L}}\\
\omega_7&=\psi_{\mathbf{H}}+\psi_{\mathbf{K}}+\psi_{\mathbf{L}}\\
\omega_8&=\psi_{\mathbf{H}}+\psi_{\mathbf{K}}+\xi_{\mathbf{L}}\\
\omega_9&=\psi_{\mathbf{H}}+\xi_{\mathbf{K}}+\xi_{\mathbf{L}}\\
\omega_{10}&=\xi_{\mathbf{H}}+\xi_{\mathbf{K}}+\xi_{\mathbf{L}},
\end{aligned} \tag{3.2}$$

the first neighborhood of each of which is defined to consist of the nine magnitudes

$$|E_{\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{\mathbf{L}}|, |F_{\mathbf{H}}|, |F_{\mathbf{K}}|, |F_{\mathbf{L}}|, |G_{\mathbf{H}}|, |G_{\mathbf{K}}|, |G_{\mathbf{L}}|. \tag{3.3}$$

3.2. *The joint probability distribution of the nine structure factors* $E_{\mathbf{H}}$, $E_{\mathbf{K}}$, $E_{\mathbf{L}}$, $F_{\mathbf{H}}$, $F_{\mathbf{K}}$, $F_{\mathbf{L}}$, $G_{\mathbf{H}}$, $G_{\mathbf{K}}$, $G_{\mathbf{L}}$, *where* $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$

It is assumed that an isomorphous triplet of structures in $P1$ with atomic position vectors $\mathbf{r}_j, j=1,2,\ldots,N$ is fixed and that normalized structure factors E, F and G are defined by (2.1)–(2.4). Denote reciprocal space by S, and by $S\times S\times S$ the threefold Cartesian product, *i.e.* the collection of all ordered triples $(\mathbf{h},\mathbf{k},\mathbf{l})$ of reciprocal-lattice vectors $\mathbf{h}$, $\mathbf{k}$, $\mathbf{l}$. The primitive random variable is the ordered triple $(\mathbf{H},\mathbf{K},\mathbf{L})$ of reciprocal-lattice vectors, which is assumed to be uniformly distributed over the subset of $S\times S\times S$ defined by (3.1). Then the structure factors $E_{\mathbf{H}}$, $E_{\mathbf{K}}$, $E_{\mathbf{L}}$, $F_{\mathbf{H}}$, $F_{\mathbf{K}}$, $F_{\mathbf{L}}$, $G_{\mathbf{H}}$, $G_{\mathbf{K}}$, $G_{\mathbf{L}}$, as functions of the primitive random variables $\mathbf{H}$, $\mathbf{K}$, $\mathbf{L}$, are themselves random variables. Denote by

$$P=P(R_1, R_2, R_3, S_1, S_2, S_3, T_1, T_2, T_3; \Phi_1, \Phi_2, \Phi_3, \Psi_1, \Psi_2, \Psi_3, \Xi_1, \Xi_2, \Xi_3) \tag{3.4}$$

the joint probability distribution of the magnitudes $|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, $|E_{\mathbf{L}}|$, $|F_{\mathbf{H}}|$, $|F_{\mathbf{K}}|$, $|F_{\mathbf{L}}|$, $|G_{\mathbf{H}}|$, $|G_{\mathbf{K}}|$, $|G_{\mathbf{L}}|$ and the phases $\varphi_{\mathbf{H}}$, $\varphi_{\mathbf{K}}$, $\varphi_{\mathbf{L}}$, $\psi_{\mathbf{H}}$, $\psi_{\mathbf{K}}$, $\psi_{\mathbf{L}}$, $\xi_{\mathbf{H}}$, $\xi_{\mathbf{K}}$, $\xi_{\mathbf{L}}$ of the complex normalized structure factors $E_{\mathbf{H}}$, $E_{\mathbf{K}}$, $E_{\mathbf{L}}$, $F_{\mathbf{H}}$, $F_{\mathbf{K}}$, $F_{\mathbf{L}}$, $G_{\mathbf{H}}$, $G_{\mathbf{K}}$, $G_{\mathbf{L}}$. Then P is given by the 18-fold integral

$$\begin{aligned}
P=(2\pi)^{-18}&R_1R_2R_3S_1S_2S_3T_1T_2T_3\\
&\times\int_{\rho_1,\rho_2,\rho_3,\sigma_1,\sigma_2,\sigma_3,\tau_1,\tau_2,\tau_3=0}^{\infty}\\
&\times\int_{\theta_1,\theta_2,\theta_3,\chi_1,\chi_2,\chi_3,\omega_1,\omega_2,\omega_3=0}^{2\pi}\rho_1\rho_2\rho_3\sigma_1\sigma_2\sigma_3\tau_1\tau_2\tau_3\\
&\times\exp\{-i[\rho_1R_1\cos(\theta_1-\Phi_1)+\rho_2R_2\cos(\theta_2-\Phi_2)\\
&+\rho_3R_3\cos(\theta_3-\Phi_3)+\sigma_1S_1\cos(\chi_1-\psi_1)\\
&+\sigma_2S_2\cos(\chi_2-\psi_2)+\sigma_3S_3\cos(\chi_3-\psi_3)\\
&+\tau_1T_1\cos(\omega_1-\Xi_1)+\tau_2T_2\cos(\omega_2-\Xi_2)\\
&+\tau_3T_3\cos(\omega_3-\Xi_3)]\}\prod_{j=1}^{N}q_j\,\mathrm{d}\rho\,\mathrm{d}\sigma\,\mathrm{d}\tau\,\mathrm{d}\theta\,\mathrm{d}\chi\,\mathrm{d}\omega,
\end{aligned} \tag{3.5}$$

where

$$\begin{aligned}
q_j=\langle\exp\{&ie_j\alpha_{200}^{-1/2}[\rho_1\cos(2\pi\mathbf{H}\cdot\mathbf{r}_j-\theta_1)\\
&+\rho_2\cos(2\pi\mathbf{K}\cdot\mathbf{r}_j-\theta_2)\\
&+\rho_3\cos(2\pi\mathbf{L}\cdot\mathbf{r}_j-\theta_3)]\\
&+if_j\alpha_{020}^{-1/2}[\sigma_1\cos(2\pi\mathbf{H}\cdot\mathbf{r}_j-\chi_1)\\
&+\sigma_2\cos(2\pi\mathbf{K}\cdot\mathbf{r}_j-\chi_2)\\
&+\sigma_3\cos(2\pi\mathbf{L}\cdot\mathbf{r}_j-\chi_3)]\\
&+ig_j\alpha_{002}^{-1/2}[\tau_1\cos(2\pi\mathbf{H}\cdot\mathbf{r}_j-\omega_1)\\
&+\tau_2\cos(2\pi\mathbf{K}\cdot\mathbf{r}_j-\omega_2)\\
&+\tau_3\cos(2\pi\mathbf{L}\cdot\mathbf{r}_j-\omega_3)]\}\rangle_{\mathbf{H}+\mathbf{K}+\mathbf{L}=0}.
\end{aligned} \tag{3.6}$$

The derivations of q_j and $\prod_{j=1}^{N}q_j$ and the evaluation of the 18-fold integral (3.5) follow the outline described by Hauptman (1982).

The final formula, the first major result of this paper, is given by (3.7)–(3.18).*

3.3. *The conditional probability distribution of the three-phase structure invariant* $\omega_1=\varphi_{\mathbf{H}}+\varphi_{\mathbf{K}}+\varphi_{\mathbf{L}}$, *given the nine magnitudes* $|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, $|E_{\mathbf{L}}|$, $|F_{\mathbf{H}}|$, $|F_{\mathbf{K}}|$, $|F_{\mathbf{L}}|$, $|G_{\mathbf{H}}|$, $|G_{\mathbf{K}}|$, $|G_{\mathbf{L}}|$ *in its first neighbourhood*

Assume again that an isomorphous triplet of structures in $P1$ is fixed and that the nine non-negative numbers R_1, R_2, R_3, S_1, S_2, S_3, T_1, T_2, T_3, instead of being variables as in (3.4) and (3.7), are also specified.

* Equations (3.7)–(3.18) have been deposited with the British Library Lending Division as Supplementary Publication No. SUP 39506(10 pp.). Copies may be obtained through The Executive Secretary, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

Suppose that the primitive random variable is the ordered triple (**H**, **K**, **L**), which is now assumed to be uniformly distributed over the subset of $S\times S\times S$ defined by (3.1) and

$$\begin{aligned} &|E_{\mathbf{H}}|=R_1, \quad |E_{\mathbf{K}}|=R_2, \quad |E_{\mathbf{L}}|=R_3,\\ &|F_{\mathbf{H}}|=S_1, \quad |F_{\mathbf{K}}|=S_2, \quad |F_{\mathbf{L}}|=S_3,\\ &|G_{\mathbf{H}}|=T_1, \quad |G_{\mathbf{K}}|=T_2, \quad |G_{\mathbf{L}}|=T_3. \end{aligned} \tag{3.19}$$

Denote by $P_1(\Omega_1|R_1, R_2, R_3, S_1, S_2, S_3, T_1, T_2, T_3)$ the conditional probability distribution of $\varphi_{\mathbf{H}}+\varphi_{\mathbf{K}}+\varphi_{\mathbf{L}}$, given (3.19). Only the case that the heavy atoms of the two derivatives are located in different positions in the unit cell is considered here. We have assumed that the normalized structure factors, the E's, F's and G's, are associated with the native protein, first derivative and second derivative, respectively.

The final formula, the second major result of this paper, is

$$P_1(\Omega_1|R_1, R_2, R_3, S_1, S_2, S_3, T_1, T_2, T_3) \simeq \frac{1}{K_1}\exp(A_1\cos\Omega_1), \tag{3.20}$$

where

$$K_1 = 2\pi I_0(A_1), \tag{3.21}$$

$$\begin{aligned} A_1 = 2\{&\beta_1 R_1R_2R_3+\beta_2[R_1R_2S_3\mu_3+R_1S_2R_3\mu_2\\ &+S_1R_2R_3\mu_1]\\ &+\beta_3[R_1R_2T_3\eta_3+R_1T_2R_3\eta_2+T_1R_2R_3\eta_1]\\ &+\beta_4[R_1S_2S_3\mu_2\mu_3+S_1R_2S_3\mu_1\mu_3+S_1S_2R_3\mu_1\mu_2]\\ &+\beta_6[R_1T_2T_3\eta_2\eta_3+T_1R_2T_3\eta_1\eta_3+T_1T_2R_3\eta_1\eta_2]\\ &+\beta_7S_1S_2S_3\mu_1\mu_2\mu_3+\beta_{10}T_1T_2T_3\eta_1\eta_2\eta_3\}, \end{aligned} \tag{3.22}$$

where

$$\mu_i = \frac{I_1(2\gamma_1R_iS_i)}{I_0(2\gamma_1R_iS_i)}, \qquad i=1,2,3 \tag{3.23}$$

$$\eta_i = \frac{I_1(2\gamma_2R_iT_i)}{I_0(2\gamma_2R_iT_i)}, \qquad i=1,2,3 \tag{3.24}$$

and

$$\gamma_1 = \frac{\alpha_{200}^{1/2}\alpha_{020}^{1/2}}{\alpha_{020}-\alpha_{200}}, \qquad \gamma_2 = \frac{\alpha_{200}^{1/2}\alpha_{002}^{1/2}}{\alpha_{002}-\alpha_{200}} \tag{3.25}$$

and I_0 and I_1 are the modified Bessel functions.

Note that in the special case that the heavy atoms of the two derivatives are located in different positions in the unit cell the coefficients β_5, β_8 and β_9 appearing in (3.7) all vanish. Moreover, the remaining β's are greatly simplified:

$$\begin{aligned} \beta_1 &\simeq -\alpha_{200}^{3/2}\left\{\frac{\alpha_{030}-\alpha_{300})}{(\alpha_{020}-\alpha_{200})^3}+\frac{(\alpha_{003}-\alpha_{300})}{(\alpha_{002}-\alpha_{200})^3}\right\}\\ \beta_2 &= \alpha_{200}\alpha_{020}^{1/2}(\alpha_{030}-\alpha_{300})(\alpha_{020}-\alpha_{200})^{-3}\\ \beta_3 &= \alpha_{200}\alpha_{002}^{1/2}(\alpha_{003}-\alpha_{300})(\alpha_{002}-\alpha_{200})^{-3}\\ \beta_4 &= -\alpha_{200}^{1/2}\alpha_{020}(\alpha_{030}-\alpha_{300})(\alpha_{020}-\alpha_{200})^{-3}\\ \beta_6 &= -\alpha_{200}^{1/2}\alpha_{002}(\alpha_{003}-\alpha_{300})(\alpha_{002}-\alpha_{200})^{-3}\\ \beta_7 &= \alpha_{020}^{3/2}(\alpha_{030}-\alpha_{300})(\alpha_{020}-\alpha_{200})^{-3}\\ \beta_{10} &= \alpha_{002}^{3/2}(\alpha_{003}-\alpha_{300})(\alpha_{002}-\alpha_{200})^{-3}. \end{aligned} \tag{3.26}$$

3.4. *The optimal case*

Let us assume that the atomic content of the first or second derivative equals the atomic content of the native protein (P) plus the heavy-atom content (H_1) or (H_2), respectively. For abbreviation define

$$\sum_{j\in P} Z_j^2 = \sum Z_p^2, \quad \sum_{k\in H_1} Z_k^2 = \sum Z_{H_1}^2, \quad \sum_{l\in H_2} Z_l^2 = \sum Z_{H_2}^2.$$

Then,

$$\begin{aligned} \alpha_{200} &= \sum Z_p^2\\ \alpha_{020} &= \sum Z_p^2+\sum Z_{H_1}^2\\ \alpha_{002} &= \sum Z_p^2+\sum Z_{H_2}^2 \end{aligned} \tag{3.27}$$

and

$$\begin{aligned} 2\gamma_1 &\simeq (1+2\sum Z_p^2/\sum Z_{H_1}^2)\\ 2\gamma_2 &\simeq (1+2\sum Z_p^2/\sum Z_{H_2}^2), \end{aligned} \tag{3.28}$$

where the Z_p are the atomic numbers (zero-angle scattering factors) of the atoms in the native protein, the Z_{H_1} the atomic numbers of the heavy atoms in the first derivative and the Z_{H_2} those of the heavy atoms in the second derivative.

When the μ_i's and η_i's approach 1, *i.e.* when $2\gamma_1R_iS_i$ and $2\gamma_2R_iT_i$ are large, we have

$$\begin{aligned} A_1 \simeq\ &[\sum Z_{H_1}^3/4(\sum Z_p^2)^{3/2}]\{R_1R_2R_3\\ &+2\gamma_1(R_1R_2\Delta_{13}+R_1\Delta_{12}R_3+\Delta_{11}R_2R_3)\\ &+4\gamma_1^2(R_1\Delta_{12}\Delta_{13}+\Delta_{11}R_2\Delta_{13}+\Delta_{11}\Delta_{12}R_3)\\ &+8\gamma_1^3\Delta_{11}\Delta_{12}\Delta_{13}\}\\ &+[\sum Z_{H_2}^3/4(\sum Z_p^2)^{3/2}\{R_1R_2R_3\\ &+2\gamma_2(R_1R_2\Delta_{23}+R_1\Delta_{22}R_3+\Delta_{21}R_2R_3)\\ &+4\gamma_2^2(R_1\Delta_{22}\Delta_{23}+\Delta_{21}R_2\Delta_{23}+\Delta_{21}\Delta_{22}R_3)\\ &+8\gamma_2^3\Delta_{21}\Delta_{22}\Delta_{23}\}, \end{aligned} \tag{3.29}$$

where

$$\begin{aligned} \Delta_{11} &= S_1-R_1, \qquad \Delta_{21} = T_1-R_1,\\ \Delta_{12} &= S_2-R_2, \qquad \Delta_{22} = T_2-R_2,\\ \Delta_{13} &= S_3-R_3, \qquad \Delta_{23} = T_3-R_3. \end{aligned} \tag{3.30}$$

The coefficients $2\gamma_1$ and $2\gamma_2$ are calculated from their respective diffraction ratios; the diffraction ratios are a measure of the average change in intensity due to the addition of heavy atoms and are estimated at low resolution as $(2\sum Z_{H_1}^2/\sum Z_p^2)^{1/2}$ and $(2\sum Z_{H_2}^2/\sum Z_p^2)^{1/2}$, respectively (Crick & Magdoff, 1956):

$$2\gamma_i \simeq 1+4\times(\text{Diffraction ratio}_i)^{-2},\ i=1,\ 2.$$

In analogy with the distributions for a pair of isomorphous structures (Fortier, Weeks & Hauptman, 1984), the predominant terms of the distribution are the $\Delta\Delta\Delta$ terms and, to a lesser extent, the $R\Delta\Delta$ terms.

The distribution is capable of yielding extremely reliable estimates, particularly in those cases when both the Δ_{ij}'s and the 2γ coefficients are large. There is clearly an optimal amount of heavy-atom substitution that leads to both sufficiently large differences in the intensities, and consequently in the normalized structure factors, and sufficiently large 2γ coefficients. For any given problem, the optimal amount of heavy-atom addition can be computed easily. In the optimal case, since the $\Delta_{i1}\Delta_{i2}\Delta_{i3}$ terms are the predominant terms of the distribution, it follows that reliable estimates (*i.e.* large A values) are obtainable, even when the normalized structure factors themselves are small, provided that the differences between the normalized structure factors of the native protein and the two derivatives are large. Furthermore, since the Δ_{ij}'s are signed values, both 0 and 180° estimates are obtainable. A particularly favorable situation occurs when $\Delta_{11}\Delta_{12}\Delta_{13}$ and $\Delta_{21}\Delta_{22}\Delta_{23}$ have the same sign.

3.5. *The conditional probability distributions of the remaining nine kinds of structure invariants,* $\omega_2, \omega_3, \ldots, \omega_{10}$

In general,

$$P_j(\Omega_j|R_1,\ R_2,\ R_3,\ S_1,\ S_2,\ S_3,\ T_1,\ T_2,\ T_3) \simeq \frac{1}{K_j}\exp(A_j\cos\Omega_j), \tag{3.31}$$

where

$$K_j = 2\pi I_0(A_j), \qquad j=1,2,\ldots,10 \tag{3.32}$$

and

$$\begin{aligned} A_j = 2\{&\beta_1\tau_1 R_1R_2R_3 \\ &+\beta_2[\tau_{21}R_1R_2S_3+\tau_{22}R_1S_2R_3+\tau_{23}S_1R_2R_3] \\ &+\beta_3[\tau_{31}R_1R_2T_3+\tau_{32}R_1T_2R_3+\tau_{33}T_1R_2R_3] \\ &+\beta_4[\tau_{41}R_1S_2S_3+\tau_{42}S_1R_2S_3+\tau_{43}S_1S_2R_3] \\ &+\beta_6[\tau_{61}R_1T_2T_3+\tau_{62}T_1R_2T_3+\tau_{63}T_1T_2R_3] \\ &+\beta_7\tau_7S_1S_2S_3+\beta_{10}\tau_{10}T_1T_2T_3\}, \\ j=1,2,\ldots,10. \end{aligned} \tag{3.33}$$

where $\tau = C_1C_2C_3$ and C_i, $i=1,2,3$, is obtained by comparing the ith structure factor associated with the coefficient of τ with the ith structure factor associated with the invariant. If these are of the same type, *i.e.* both R or both S or both T, then $C_i = 1{\cdot}0$, $i=1,2,3$. If one of these is of type R and the other of type S, then $C_i=\mu_i$ [(3.23)], $i=1,2,3$. If one of these is of type R and the other of type T, then $C_i=\eta_i$ [(3.24)], $i=1,2,3$. If one of these is type S and the other of type T, then $C_i=\mu_i\eta_i$, $i=1,2,3$.

As an example, let us consider the invariant $\Omega_5 = \varphi_{\mathbf{H}}+\psi_{\mathbf{K}}+\xi_{\mathbf{L}}$. Its associated normalized structure factors are R_1, S_2 and T_3 and

$$\begin{aligned} \tau_1 &= 1{\cdot}0\times\mu_2+\eta_3 \\ \tau_{21} &= 1{\cdot}0\times\mu_2\times\mu_3\eta_3 \\ \tau_{22} &= 1{\cdot}0\times1{\cdot}0\times\eta_3 \\ \tau_{23} &= \mu_1\times\mu_2\times\eta_3 \\ \tau_{31} &= 1{\cdot}0\times\mu_2\times1{\cdot}0 \\ \tau_{32} &= 1{\cdot}0\times\mu_2\eta_2\times\eta_3 \\ \tau_{33} &= \eta_1\times\mu_2\times\eta_3 \\ \tau_{41} &= 1{\cdot}0\times1{\cdot}0\times\mu_3\eta_3 \\ \tau_{42} &= \mu_1\times\mu_2\times\mu_3\eta_3 \\ \tau_{43} &= \mu_1\times1{\cdot}0\times\eta_3 \\ \tau_{61} &= 1{\cdot}0\times\mu_2\eta_2\times1{\cdot}0 \\ \tau_{62} &= \eta_1\times\mu_2\times1{\cdot}0 \\ \tau_{63} &= \eta_1\times\mu_2\eta_2\times\eta_3 \\ \tau_7 &= \mu_1\times1{\cdot}0\times\mu_3\eta_3 \\ \tau_{10} &= \eta_1\times\mu_2\eta_2\times1{\cdot}0. \end{aligned}$$

3.6. *Advantages in phasing the native protein and the derivatives simultaneously*

For a triplet of reciprocal-lattice vectors **H**, **K**, **L** satisfying $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$ there exist 27 three-phase structure invariants (3.2). Estimates of each of the 27 invariants are obtained from the evaluation of (3.31) and, in particular, from the computation of their respective A values [(3·33)]. In most cases, the estimates (0 or 180°) are the same for all 27 invariants belonging to the same family of reciprocal-lattice vectors **H**, **K** and **L**. The A values, however, may differ significantly. In the special case that all the $2\gamma_1R_iS_i$ and $2\gamma_2R_iT_i$ are large, the 27 invariants have the same A values. If some of the $2\gamma_1R_iS_i$ or $2\gamma_2R_iT_i$ are small, the A values for the 27 invariants differ. For example, if $2\gamma_1R_1S_1$ is small, all the invariants containing the phase $\varphi_{\mathbf{H}}$ have the same A values while the remaining invariants have a common A value, different from the former. For a given **H**, **K**, **L** satisfying $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$, the family of 27 invariants and their estimates constitute a redundant system of linear

Table 1. *Heavy-atom content of the two cytochrome c_{550} derivatives used in the calculations*

Derivative	Abbreviation	Effective Z(e)	Occupancy
$PtCl_4^{2-}$	1Pt	84·24	1·08
UO_2^{2+}	2U	80·04 [U(1)]	0·87
		7·36 [U(2)]	0·08

equations. In this system, owing to the relation $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$, only two phases, *e.g.* $\varphi_{\mathbf{H}}$ and $\psi_{\mathbf{K}}$, are linearly independent and are therefore suitable for origin specification; their values may be specified arbitrarily. The remaining seven phases, again because of the relation $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$, are linearly dependent on the pair $\Phi_{\mathbf{H}}$, $\Psi_{\mathbf{K}}$ and, once an enantiomorph has been fixed, *e.g.* by specifying arbitrarily the sign of a suitable structure invariant, are uniquely determined by the observed magnitudes $|E|, |F|, |G|$ and the specified values of the origin fixing pair $\varphi_{\mathbf{H}}$, $\psi_{\mathbf{K}}$. Thus, the simultaneous use of all of the invariants, in convergence mapping and tangent refinement, automatically ensures common origin and enantiomorph definition in the native and the derivatives. Definition of the origin and enantiomorph is done in the usual manner.

3.7. *Test calculations*

The three-phase invariant probability distributions for the native and one derivative case (Hauptman, 1982) and for the native and two derivative case [(3.31)] were used to estimate the values of these invariants for cytochrome c_{550}. Cytochrome c_{550} from *Paracoccus denitrificans* is a moderate-size protein, $M_r = 14\,500$, which crystallizes in space group $P2_12_12_1$ with four molecules in the unit cell (Timkovich & Dickerson, 1973, 1976). Coordinates were obtained from the Protein Data Bank (Bernstein *et al.*, 1977) and used to calculate structure factors and normalized structure factors for the native protein and the two derivatives described in Table 1. Fixed spherical atoms were assumed. The computed normalized structure-factor amplitudes were than used to generate three-phase structure invariants and to estimate their A values.

The estimated invariants were then sorted in decreasing order according to $|A|$, and the desired numbers of highest ranking invariants were retained. The results of the calculations are shown in Tables 2 and 3. The invariants in Tables 2 and 3 were evaluated using probability distributions for a pair of isomorphous structures and for a triplet of isomorphous structures, respectively. More specifically, Hauptman's (1982) formulas were applied to the three isomorphous pairs defined by the isomorphous triplet of structures, and the results summarized in Table 2. In Table 3 are summarized the results obtained by applying (3.31), the formulas specific for the isomorphous triplet. Comparison of Tables 2 and 3 clearly shows that, as anticipated, diffraction data from a triplet of isomorphous structures yield better estimates for the three-phase structure invariants *via* (3.31) than are obtainable by means of the earlier formulas appropriate to isomorphous pairs.

Table 2. *Average magnitude of the error in estimated values* (0 *or* 180°) *of* 128 000 *three-phase invariants for cytochrome c_{550} and the derivatives* 1Pt *and* 2U

The full 4 Å sets of phases φ, ψ and ξ (*i.e.* 1076 φ's, 1076 ψ's and 1076 ξ's) were used, two at a time, to generate three sets of three-phase invariants. The invariant sets were merged and the 128 000 three-phase structure invariants corresponding to the 128 000 largest $|A_i|$ values were used to construct this table.

Number in group	Average value of $\|A_i\|$	Average |error| (°)	% of invariants with |error|≥90°
1000	6·29	12·7	0·80
5000	5.34	13.4	0.32
10 000	4·81	15·4	0·44
15 000	4·47	17·1	0·50
20 000	4·23	18·7	0·60
25 000	4·03	19·9	0·55
50 000	3·42	24·4	1·06
75 000	3·05	27·4	1·90
100 000	2·79	29·4	2·71
128 000	2·57	31·6	3·60

Table 3. *Average magnitude of the error in estimated values* (0 *or* 180°) *of* 128 000 *three-phase invariants for cytochrome c_{550} and the derivatives* 1Pt *and* 2U [*using equation* 3.31)]

The full 4 Å sets of phase φ, ψ and ξ (*i.e.* 1076 φ's, 1076 ψ's and 1076 ξ's) were used to generate the three-phase invariants. The 128 000 three-phase invariants corresponding to the 128 000 largest $|A_i|$ values were used to construct this table.

Number in group	Average value of $\|A_i\|$	Average |error| (°)	% of invariants with |error|≥90°
1000	9·43	9·5	0·00
5000	7·65	13·3	0·00
10 000	6·86	14·7	0·12
15 000	6·40	15·7	0·19
20 000	6·07	16·3	0·20
25 000	5·81	17·0	0·25
50 000	5·01	19·7	0·63
75 000	4·54	21·7	0·79
100 000	4·21	23·3	1·00
128 000	3·93	24·8	1·22

4. Concluding remarks

Recent advances in direct methods have been integrated with the method of isomorphous replacement, and the probabilistic theory of the three-phase structure invariants for a triplet of isomorphous structures has been worked out. In particular, the conditional probability distribution of the three-phase structure invariant, assuming as known the nine magnitudes in its first neighbourhood, has been derived for the case of a native protein and two heavy-atom

derivatives where the heavy atoms of the derivatives occupy different positions. The distribution yields a reliable estimate (0 or π) for the invariant in the favorable case that the variance of the distribution is small. An example shows the improvement in estimates of the three-phase structure invariants which results from the ability now to exploit simultaneously the diffraction data from a triple of isomorphous structures, at least in the special case of a native protein and two heavy-atom derivatives in which the heavy atoms of the derivatives are located in different positions in the unit cell. Particularly noteworthy is the ease of unique origin and enantiomorph specification in direct-methods applications to all three structures.

It would be premature to assess, at this point, the role that the distributions will play in actual macromolecular structure determinations, or to compare the present technique with the standard multiple isomorphous replacement technique. As mentioned earlier, several questions remain to be answered, principally concerning the effects of errors in the diffraction data and of imperfect isomorphism. These questions are the subject of a present study and the results will be presented at a later date.

It should be stated in conclusion that, in view of the available evidence, the integrated direct methods–isomorphous replacement probability distributions constitute a sound theoretical basis for macromolecular phase determination.

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (SF), a Queen's University Research Award (SF); grant No. CHE-8203930 from the National Science Foundation (CMW and HH) and a grant from the James H. Cummings Foundation (CMW and HH).

References

BERNSTEIN, F. C., KOETZLER, T. F., WILLIAMS, G. J. B., MEYER, E. F. JR, BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.

CRICK, F. H. C. & MAGDOFF, B. (1956). *Acta Cryst.* **9**, 901–908.

FORTIER, S., WEEKS, C. & HAUPTMAN, H. (1984). *Acta Cryst.* A**40**, 544–548.

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

HAUPTMAN, H., POTTER, S. & WEEKS, C. (1982). *Acta Cryst.* A**38**, 294–300.

TIMKOVICH, R. & DICKERSON, R. E. (1973). *J. Mol. Biol.* **79**, 39–56.

TIMKOVICH, R. & DICKERSON, R. E. (1976). *J. Biol. Chem.* **251**, 4033–4046.

Acta Cryst. (1984). A**40**, 651-660

Exact Random-Walk Models in Crystallographic Statistics. I. Space Groups $P\bar{1}$ and $P1$

BY URI SHMUELI

Department of Chemistry, Tel Aviv University, 69 978 *Tel Aviv, Israel*

GEORGE H. WEISS AND JAMES E. KIEFER

National Institutes of Health, Bethesda, Maryland 20205, *USA*

AND ARTHUR J. C. WILSON

Crystallographic Data Centre, University Chemical Laboratory, Cambridge CB2 1EW, England

(*Received* 31 *October* 1983; *accepted* 5 *June* 1984)

Abstract

Probability density functions that are exact solutions to classical random-walk problems have been adapted to represent distributions of the magnitude of the normalized structure factor, for the space groups $P\bar{1}$ and $P1$. The functions are given by readily summable Fourier and Fourier–Bessel series, and account explicitly for the atomic composition of the asymmetric unit. These new probability density functions have been extensively tested by comparison with simulated histograms of $|E|$, for a wide range of atomic compositions. The most heterogeneous compositions examined are $C_{14}U$ and $C_{29}U$, for $P\bar{1}$ and $P1$, respectively. Very good agreement between the simulated and theoretical distributions has been obtained in all these tests, over the entire (useful) range $0<|E|<3$. A distribution of $|E|$ values, recalculated from published data on a triclinic platinum complex with chloroorganic ligands, has also been

0108-7673/84/060651-10$01.50