Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 16 March 2012
Accepted 18 May 2012

Estimates of triplet invariants given a model structure

Maria Cristina Burla,[a] Benedetta Carrozzini,[b] Giovanni Luca Cascarano,[b] Giuliana Comunale,[c] Carmelo Giacovazzo,[b]* Annamaria Mazzone[b] and Giampiero Polidori[b]

[a]Department of Earth Sciences, University of Perugia, 06100 Perugia, Italy, [b]Institute of Crystallography – CNR, Via G. Amendola, 122/O 70126 Bari, Italy, and [c]Dipartimento di Chimica, Università della Basilicata, 85100 Potenza, Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

The triplet structure invariant is estimated *via* the method of joint probability distribution functions when a model structure is available. The six-variate probability distribution function $P(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}}, E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}})$ is studied under the condition that imperfect isomorphism between the target and model structures exist. The results are compared with those available in the literature, which were obtained under the condition of perfect isomorphism. It is shown that the new formalism is more suitable for real cases, where perfect isomorphism is very rare.

1. Symbols and abbreviations

N: number of atoms in the unit cell for the target structure (this is the crystal structure we want to solve).
N_p: number of atoms in the unit cell for the model structure (this corresponds to the structural model available at a given step of the phasing process). Usually $N_p \leq N$.
f_j, $j = 1, \ldots, N$: atomic scattering factors for the target structure (temperature factor included).
$\sum_{N1}$, $\sum_{N2}$, $\sum_{N3} = \sum_{j=1}^{N} f_j^2(\mathbf{h})$, $\sum_{j=1}^{N} f_j^2(\mathbf{k})$, $\sum_{j=1}^{N} f_j^2(\mathbf{h}+\mathbf{k})$, respectively.
$\sum_{p1}$, $\sum_{p2}$, $\sum_{p3} = \sum_{j=1}^{Np} f_j^2(\mathbf{h})$, $\sum_{j=1}^{Np} f_j^2(\mathbf{k})$, $\sum_{j=1}^{Np} f_j^2(\mathbf{h}+\mathbf{k})$, respectively.
$t(\mathbf{h},\mathbf{k}) = [\sum_{j=1}^{N} f_j(\mathbf{h}) f_j(\mathbf{k}) f_j(\mathbf{h}+\mathbf{k})]/(\sum_{N1}\sum_{N2}\sum_{N3})^{1/2}$. In practical cases $t(\mathbf{h},\mathbf{k})$ is approximated by $1/(N_{\rm eq})^{1/2} = (\sum_{j=1}^{N} Z_j^3)/(\sum_{j=1}^{N} Z_j^2)^{3/2}$, where Z_j is the atomic number of the jth atom.
$t_p(\mathbf{h},\mathbf{k}) = [\sum_{j=1}^{Np} f_j(\mathbf{h}) f_j(\mathbf{k}) f_j(\mathbf{h}+\mathbf{k})]/(\sum_{p1}\sum_{p2}\sum_{p3})^{1/2}$, usually approximated by $1/(N_{p,\rm eq})^{1/2} = (\sum_{j=1}^{Np} Z_j^3)/(\sum_{j=1}^{Np} Z_j^2)^{3/2}$.
$F = \sum_{j=1}^{N} f_j \exp[2\pi i \mathbf{h}(\mathbf{r}_{pj} + \Delta\mathbf{r}_j)] = |F| \exp(i\varphi)$: structure factor of the target structure.
$\Delta\mathbf{r}_j$ is the misfit between the atomic position $\mathbf{r}_j$ in the target and the corresponding position in the model structure.
$F_p = \sum_{j=1}^{Np} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_{pj}) = |F_p| \exp(i\varphi_p)$, where $\mathbf{r}_{pj}$ are the atomic positions in the model structure.
$E = A + iB = R\exp(i\varphi)$, $E_p = A_p + iB_p = R_p \exp(i\varphi_p)$: normalized structure factors of F and F_p, respectively.
$D_{\mathbf{h}} = \langle \cos(2\pi \mathbf{h}\Delta\mathbf{r}) \rangle$: the average is performed per resolution shell. If $D_{\mathbf{h}} = 1$, the model and target structure positions coincide.
$\sigma_A = D(\Sigma_p/\Sigma_N)^{1/2}$: σ_A is a statistical estimate of the correlation between the model and target structure. Ideally $\sigma_A = 0$ for uncorrelated models, $\sigma_A = 1$ for identical model and target structures.
$\sigma_R^2 = \langle |\mu|^2 \rangle / \sum_N : \langle |\mu|^2 \rangle$ is the measurement error.
$e = 1 + \sigma_R^2$.
$I_i(x)$: modified Bessel function of order i.
$m = \langle \cos(\varphi - \varphi_p) \rangle = I_1(X)/I_0(X)$, where $X = 2\sigma_A R R_p \times (e - \sigma_A^2)^{-1}$.
$\Phi = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}$.
$\Phi_p = \varphi_{p\mathbf{h}} + \varphi_{p\mathbf{k}} + \varphi_{p,-\mathbf{h}-\mathbf{k}}$.
DM: direct methods.

2. Introduction

Cochran's (1955) formula for the estimation of the triplet phase invariants, *i.e.*

$$P(\Phi | R_{\mathbf{h}}, R_{\mathbf{k}}, R_{-\mathbf{h}-\mathbf{k}}) \approx [2\pi I_0(G)]^{-1} \exp(G\cos\Phi), \qquad (1)$$

with

$$G = \frac{2R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h}+\mathbf{k}}}{N^{1/2}},$$

has been the probabilistic basis for the main DM tool, the tangent formula (Hauptman & Karle, 1956; Karle & Karle, 1966). Equation (1) has been derived in the absence of any structural model, *i.e.*, it is only able to exploit the normalized moduli of the observed structure factors. The coefficient $1/N^{1/2}$ is strictly valid only for equal-atom structures: in the most general case $1/N^{1/2}$ should be replaced by $t(\mathbf{h},\mathbf{k})$, which, in accordance with our notation, may be approximated by $1/(N_{\rm eq})^{1/2}$.

Over the last 40 years, thanks to the efforts of many authors, DM have been enriched by new tools: neighbourhoods

(Hauptman, 1976) and representation (Giacovazzo, 1977, 1980) approaches provided the theoretical basis for the use of seminvariants and invariants of higher order, and allowed a remarkable improvement of the Cochran triplet phase estimates (*e.g.*, *via* the so-called *P*10 formula: see Cascarano *et al.*, 1984).

Dual-space techniques (Miller *et al.*, 1993; Sheldrick, 1998; Burla *et al.*, 2005; Yao *et al.*, 2006) combined DM with real-space procedures, with a further gain in efficiency. As a result, DM solved in practice the phase problem at least for structures up to 1000 atoms in the asymmetric unit.

The literature on DM is immense and cannot exhaustively be quoted here: the reader is referred to Giacovazzo (1998) for a general review. We only state here that the success of DM for small- and medium-size structures was also the reason for a general loss of interest: they did the job so well that supplementary investigations were considered to be unnecessary. In more recent years, attention has mainly focused on:

(i) Alternative phasing techniques like charge flipping (Oszlányi & Sütő, 2004, 2005; Palatinus *et al.*, 2006; Palatinus & Chapuis, 2007; Dumas & van der Lee, 2008) and *VLD* (*Vive la difference*) (Burla, Caliandro *et al.*, 2010; Burla, Giacovazzo & Polidori, 2010; Burla, Giacovazzo & Polidori, 2011; Burla, Carrozzini *et al.*, 2011). For such approaches a deep theoretical knowledge of invariants and seminvariant theories is no longer necessary. The simplicity of the algorithms is very appealing and boosted their popularity. The program codes mostly require effective fast Fourier transform (FFT) algorithms and a limited number of instructions.

(ii) Patterson deconvolution techniques (Buerger, 1959), eventually based on the implication transformations and on superposition techniques (Richardson & Jacobson, 1987; Pavelčík *et al.*, 1992). In the evolved form they proved to be highly competitive both for *ab initio* protein phasing (Burla *et al.*, 2005, 2006; Caliandro *et al.*, 2008) and for finding the substructure in multiple-wavelength anomalous scattering (MAD) techniques (Burla *et al.*, 2007).

The question now is: the Cochran formula was derived in the absence of a structure model. Can this formula be replaced by a more useful expression if a model structure is available?

Formulas for estimating triplet invariants given a model are already known. The investigation started with Main (1976, 1979) [see also Heinerman (1977) and Heinerman *et al.* (1977)]. The most general treatment was described by Giacovazzo (1983), who studied the conditional distribution

$$P(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}} | E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}}) \qquad (2)$$

under the following hypotheses:

(i) only one set of diffraction data is measured (the target set); structure factors of the partial structure, in modulus and phase, are calculated from the model;

(ii) the reflection indices as well as the coordinates $\mathbf{r}_{pj}$, $j = 1, \ldots, p$, are fixed known parameters;

(iii) p positional vectors of the target structure perfectly coincide with the p vectors of the model;

(iv) the atomic positions $\mathbf{r}_j$, $j = p+1, \ldots, N$ are the primitive random variables of the probabilistic approach, uniformly distributed in the unit cell.

The aim of the method is to facilitate the recovery of the target structure when a highly correlated model is available. The approach was successfully tested (Camalli *et al.*, 1985) but the simultaneous success of dual-space methods made it obsolete. Under the conditions (i)–(iv) some structural parameters are perfectly defined: *e.g.* the scattering powers of the difference (target minus model) structure,

$$\textstyle\sum_{qi} = \sum_{Ni} - \sum_{pi}, \quad i = 1, 2, 3,$$

and the corresponding equivalent number of atoms $N_{q,\mathrm{eq}}$ defined by

$$\frac{1}{(N_{q,\mathrm{eq}})^{1/2}} = \frac{\sum_{j=Np+1}^{N} Z_j^3}{(\sum_{j=Np+1}^{N} Z_j^2)^{3/2}}. \qquad (3)$$

From the conditional distribution (2) the following estimate for $\varphi_{\mathbf{h}}$ was derived:

$$P(\varphi_{\mathbf{h}} | \ldots) = [2\pi I_0(G_{\mathbf{h}})]^{-1} \exp[G_{\mathbf{h}} \cos(\varphi_{\mathbf{h}} - \xi_{\mathbf{h}})], \qquad (4)$$

where

$$G_{\mathbf{h}}^2 = a_{\mathbf{h}}^2 + b_{\mathbf{h}}^2,$$

$$\begin{aligned} a_{\mathbf{h}} = 2R_{\mathbf{h}} \Bigg\{ & \gamma_0 R_{p\mathbf{h}} \cos\varphi_{p\mathbf{h}} + \frac{2}{(N_{q,\mathrm{eq}})^{1/2}} \Big[\gamma_1 R_{\mathbf{k}} R_{\mathbf{h}+\mathbf{k}} \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}+\mathbf{k}}) \\ & - \gamma_2 R_{\mathbf{k}} R_{p\mathbf{h}+\mathbf{k}} \cos(\varphi_{\mathbf{k}} + \varphi_{p\mathbf{h}+\mathbf{k}}) - \gamma_3 R_{p\mathbf{k}} R_{\mathbf{h}+\mathbf{k}} \cos(\varphi_{p\mathbf{k}} + \varphi_{\mathbf{h}+\mathbf{k}}) \\ & + \gamma_4 R_{p\mathbf{k}} R_{p\mathbf{h}+\mathbf{k}} \cos(\varphi_{p\mathbf{k}} + \varphi_{p\mathbf{h}+\mathbf{k}}) \Big] \Bigg\} \end{aligned}$$

$$\begin{aligned} b_{\mathbf{h}} = 2R_{\mathbf{h}} \Bigg\{ & \gamma_0 R_{p\mathbf{h}} \sin\varphi_{p\mathbf{h}} + \frac{2}{(N_{q,\mathrm{eq}})^{1/2}} \Big[- \gamma_1 R_{\mathbf{k}} R_{\mathbf{h}+\mathbf{k}} \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}+\mathbf{k}}) \\ & + \gamma_2 R_{\mathbf{k}} R_{p\mathbf{h}+\mathbf{k}} \sin(\varphi_{\mathbf{k}} + \varphi_{p\mathbf{h}+\mathbf{k}}) + \gamma_3 R_{p\mathbf{k}} R_{\mathbf{h}+\mathbf{k}} \sin(\varphi_{p\mathbf{k}} + \varphi_{\mathbf{h}+\mathbf{k}}) \\ & - \gamma_4 R_{p\mathbf{k}} R_{p\mathbf{h}+\mathbf{k}} \sin(\varphi_{p\mathbf{k}} + \varphi_{p\mathbf{h}+\mathbf{k}}) \Big] \Bigg\}, \end{aligned}$$

$$\cos\xi_{\mathbf{h}} = a_{\mathbf{h}}/G_{\mathbf{h}}, \quad \sin\xi_{\mathbf{h}} = b_{\mathbf{h}}/G_{\mathbf{h}},$$

$$\gamma_0 = \frac{(\sum_{N1} \sum_{p1})^{1/2}}{\sum_{q1}}, \quad \gamma_1 = \left(\frac{\sum_{N1} \sum_{N2} \sum_{N3}}{\sum_{q1} \sum_{q2} \sum_{q3}} \right)^{1/2},$$

$$\gamma_2 = \left(\frac{\sum_{N1} \sum_{N2} \sum_{p3}}{\sum_{q1} \sum_{q2} \sum_{q3}} \right)^{1/2}, \quad \gamma_3 = \left(\frac{\sum_{N1} \sum_{p2} \sum_{N3}}{\sum_{q1} \sum_{q2} \sum_{q3}} \right)^{1/2},$$

$$\gamma_4 = \left(\frac{\sum_{N1} \sum_{p2} \sum_{p3}}{\sum_{q1} \sum_{q2} \sum_{q3}} \right)^{1/2}.$$

A related probabilistic approach was described by Hauptman (1982). He integrated DM with isomorphous-replacement techniques under the following assumptions:

(*a*) two sets of diffraction amplitudes are measured, corresponding to protein and derivative crystals: no model, and therefore no phase information is available;

(*b*) reflection indices are the primitive random variables, while the atomic positions in both the structures are unknown fixed parameters;

(*c*) the isomorphous derivative is obtained by addition of heavy atoms, with perfect isomorphism between the two crystals. That is, protein atoms occupy the same positions as the corresponding atoms in the derivative.

Under the above conditions the distribution

$$P\left(E_{\mathbf{dh}}, E_{\mathbf{dk}}, E_{\mathrm{d},-\mathbf{h}-\mathbf{k}}, E_{\mathbf{Ph}}, E_{\mathbf{Pk}}, E_{\mathrm{P},-\mathbf{h}-\mathbf{k}}\right) \quad (5)$$

was studied, where subscripts d and capital P stand for derivative and protein, respectively. The protein triplet phase Φ_{P} was estimated by Hauptman *via* the conditional distribution

$$P\left(\Phi_{\mathrm{P}}|R_{\mathbf{dh}}, R_{\mathbf{dk}}, R_{\mathrm{d},-\mathbf{h}-\mathbf{k}}, R_{\mathbf{Ph}}, R_{\mathbf{Pk}}, R_{\mathrm{P},-\mathbf{h}-\mathbf{k}}\right) = 2\pi I_0(A_{\mathrm{P}}) \exp(A_{\mathrm{P}} \cos \Phi_{\mathrm{P}}). \quad (6)$$

A_{P} is described by a rather cumbersome algebraic expression in Hauptman notation. We report here the simplified formula obtained by Giacovazzo *et al.* (1988), which well approximates the original expression:

$$A_{\mathrm{P}} = \frac{2}{(N_{\mathrm{P,eq}})^{1/2}} R_{\mathrm{P1}} R_{\mathrm{P2}} R_{\mathrm{P3}} + \frac{2}{(N_{q,\mathrm{eq}})^{1/2}} \frac{(|F_{\mathrm{d1}}| - |F_{\mathrm{P1}}|)(|F_{\mathrm{d2}}| - |F_{\mathrm{P2}}|)(|F_{\mathrm{d3}}| - |F_{\mathrm{P3}}|)}{(\sum_{q1} \sum_{q2} \sum_{q3})^{1/2}}. \quad (7)$$

It is easily seen that the first term on right-hand side of (7) is the standard Cochran parameter: the second term involves the differences $(|F_{\mathrm{d}i}| - |F_{\mathrm{P}i}|)$ normalized with respect to the scattering power of the difference structure (in this case coincident with the heavy-atom substructure: accordingly $\sum_q = \sum_{\mathrm{H}}$,where $\sum_{\mathrm{H}}$ represents the scattering power of the heavy atoms). Since $1/(N_{q,\mathrm{eq}})^{1/2} \ll 1/(N_{\mathrm{P,eq}})^{1/2}$, the second term on the right-hand side of equation (7) is usually large and dominant with respect to the first one. As a consequence, protein triplet phases are no longer expected to be distributed around 0: they may be accurately estimated as 0 or π according to the sign of A_{P}.

The reader will certainly have noticed that prior knowledge of model moduli and phases is needed for estimating Φ from distribution (2), while the estimate of Φ_{P} *via* equation (6) does not imply any prior phase knowledge, but requires two sets of diffraction data, corresponding to protein and derivative. This observation would suggest that the Hauptman and Giacovazzo approaches are based on different assumptions and have different aims, but it may be shown that they are strictly related. Indeed:

(1) The protein and derivative structures in the Hauptman approach play the same roles as the partial and target structures in the Giacovazzo approach, respectively.

(2) Both approaches rely on the same joint probability distribution:

$$P\left(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}}, E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}}\right). \quad (8)$$

Choosing the conditional $P\left(E_{\mathbf{h}}|E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}}, E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}}\right)$ leads to Giacovazzo's formula, choosing $P(\Phi_p|R_{\mathbf{h}}, R_{\mathbf{k}}, R_{-\mathbf{h}-\mathbf{k}}, R_{p\mathbf{h}}, R_{p\mathbf{k}}, R_{p,-\mathbf{h}-\mathbf{k}})$ leads to Hautpman's distribution. Therefore, in accordance with statement (1), $P(\Phi_p|R_{\mathbf{h}}, R_{\mathbf{k}}, R_{-\mathbf{h}-\mathbf{k}}, R_{p\mathbf{h}}, R_{p\mathbf{k}}, R_{p,-\mathbf{h}-\mathbf{k}})$ coincides with the Hauptman distribution $P\left(\Phi_{\mathrm{P}}|R_{\mathbf{dh}}, R_{\mathbf{dk}}, R_{\mathrm{d},-\mathbf{h}-\mathbf{k}}, R_{\mathbf{Ph}}, R_{\mathbf{Pk}}, R_{\mathrm{P},-\mathbf{h}-\mathbf{k}}\right)$.

(3) Both approaches exploit eight types of triplet cosines, say

$$\begin{aligned} &\cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}), \quad \cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{p-\mathbf{h}-\mathbf{k}}),\\ &\cos(\varphi_{\mathbf{h}} + \varphi_{p\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}), \quad \cos(\varphi_{p\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}),\\ &\cos(\varphi_{\mathbf{h}} + \varphi_{p\mathbf{k}} + \varphi_{p-\mathbf{h}-\mathbf{k}}), \quad \cos(\varphi_{p\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{p-\mathbf{h}-\mathbf{k}}),\\ &\cos(\varphi_{p\mathbf{h}} + \varphi_{p\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}), \quad \cos(\varphi_{p\mathbf{h}} + \varphi_{p\mathbf{k}} + \varphi_{p-\mathbf{h}-\mathbf{k}}). \end{aligned} \quad (9)$$

What are the limits of the two approaches? For both of them the N_p atomic positions of the model perfectly coincide with the corresponding atomic positions in the target structure. This restriction is too limiting for real cases: a theory of triplet phase invariants in which the above hypothesis is relaxed may better fit a real scenario and contribute to DM renovation. In accordance with the above statements, a generalized expression of the distribution (8) would generalize both the Hauptman (1982) and Giacovazzo (1983) approaches. This is the first aim of this paper.

The mathematical work necessary for deriving such distributions is the occasion for establishing previously unnoticed properties connected to the assumptions under which the joint probability distributions are derived. For example, hypotheses on the primitive random variables so far considered mathematically equivalent are shown to provide different distributions and therefore different conclusive phasing formulas. This is the second aim of this paper. The theoretical results so obtained allow one to postulate a possible renovation of the present DM procedures.

3. The joint probability distribution $P(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}}, E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}})$ in the case of imperfect isomorphism: hypothesis I

We will calculate the conditional joint probability distribution (8) under the following assumptions:

(1) $\mathbf{r}_{pj}$, $j = 1, \ldots, N_p$, are the atomic positional parameters of the model structure. The $\mathbf{r}_{pj}$'s are primitive random variables of our approach, uniformly distributed in the unit cell.

(2) $\mathbf{r}_j$, $j = 1, \ldots, N$ are the atomic positional parameters of the target structure. N_p of them, say $\mathbf{r}_j = \mathbf{r}_{pj} + \Delta\mathbf{r}_j$, $j = 1, \ldots, N_p$, are riding variables: they are correlated with the corresponding $\mathbf{r}_{pj}$'s through local positional errors $\Delta\mathbf{r}_j$'s. The $\Delta\mathbf{r}_j$ moduli are restrained to assume sufficiently small values to secure, at least at low resolution, isomorphism between the model and target structures. The positional parameters $\mathbf{r}_j$, $j = N_p + 1, \ldots, N$, are primitive random variables uniformly distributed in the unit cell.

(3) Two supplementary primitive random variables, μ and ϑ, are introduced to take into account the experimental uncertainty of the observed structure-factor moduli. We will write for the target structure

$$F = \sum_{j=1}^{N} f_j \exp 2\pi i \mathbf{h}\mathbf{r}_j + \mu \exp(i\vartheta).$$

(4) All the primitive random variables are assumed to be statistically independent of each other.

The above assumptions (allowing different degrees of non-isomorphism between the model and target structures) have been already used by Srinivasan & Ramachandran (1965) and by Caliandro *et al.* (2005) for the derivation of the joint probability distribution $P(E_{\mathbf{h}}, E_{p\mathbf{h}})$, a very important tool for improving the model structure. It suggests optimal weights for calculated phases *via* the parameter $\sigma_A = D(\Sigma_p/\Sigma_N)^{1/2}$, the value of which depends on the correlation between the model and target structures.

In all the probabilistic approaches involving triplet invariants the assumptions (1)–(4) have never been used. The condition $D_i = 1$ for $i = 1, 2, 3$ (or equivalently, $\Delta\mathbf{r}_j = 0$ for $j = 1, \ldots, N_p$), was always introduced, according to which the positions of the N_p atoms of the model perfectly coincide with N_p atoms of the target structure. In this paper, according to conditions (1)–(4), we break down this condition and use the following definitions (from now on hypothesis I):

$$
\begin{aligned}
A &= \left\{ \sum_{j=1}^{Np} f_j \cos[2\pi\mathbf{h}(\mathbf{r}_{pj} + \Delta\mathbf{r}_j)] + \sum_{j=Np+1}^{N} f_j \cos 2\pi\mathbf{h}\mathbf{r}_j \right. \\
&\quad \left. + |\mu| \cos\vartheta \right\} \Big/ \left(\varepsilon\Sigma_N\right)^{1/2}, \\
B &= \left\{ \sum_{j=1}^{Np} f_j \sin[2\pi\mathbf{h}(\mathbf{r}_{pj} + \Delta\mathbf{r}_j)] + \sum_{j=Np+1}^{N} f_j \sin 2\pi\mathbf{h}\mathbf{r}_j \right. \\
&\quad \left. + |\mu| \sin\vartheta \right\} \Big/ \left(\varepsilon\Sigma_N\right)^{1/2}, \\
A_p &= \sum_{j=1}^{Np} f_j \cos 2\pi\mathbf{h}\mathbf{r}_{pj} / \left(\varepsilon\Sigma_p\right)^{1/2}, \\
B_p &= \sum_{j=1}^{Np} f_j \sin 2\pi\mathbf{h}\mathbf{r}_{pj} / \left(\varepsilon\Sigma_p\right)^{1/2},
\end{aligned}
$$

where ε is the Wilson statistical weight. To simplify the resulting formulas we will use the following notation:

$E_i = A_i + iB_i = E_{\mathbf{h}}, E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}}$, respectively, for $i = 1, 2, 3$.

$E_{pi} = A_{pi} + iB_{pi} = E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}}$, respectively, for $i = 1, 2, 3$.

$R_i = |E_i|$, $R_{pi} = |E_{pi}|$, $i = 1, 2, 3$.

f_i, $i = 1, 2, 3$ are equal to $f(\mathbf{h})$, $f(\mathbf{k})$, $f(\mathbf{h}+\mathbf{k})$, respectively.

$\sigma_{A1} = \sigma_A(\mathbf{h})$, $\sigma_{A2} = \sigma_A(\mathbf{k})$, $\sigma_{A3} = \sigma_A(\mathbf{h}+\mathbf{k})$.

We will calculate the joint probability distribution $P(A_i, A_{pi}, B_i, B_{pi}, i = 1, 2, 3)$, the characteristic function of which is

$$
\begin{aligned}
&C(u_i, u_{pi}, v_i, v_{pi},\ \ i = 1, 2, 3) \\
&\quad = \left\langle \exp\left\{ i \sum_{i=1}^{3} (u_i A_i + u_{pi} A_{pi} + v_i B_i + v_{pi} B_{pi}) \right\} \right\rangle \\
&\quad = \exp\left\{ -(1/4) \sum_{i=1}^{3} [e_i(u_i^2 + v_i^2) + u_{pi}^2 + v_{pi}^2 \right. \\
&\qquad + 2\sigma_{Ai}(u_i u_{pi} + v_i v_{pi})] \\
&\qquad - [i/4(N_{\rm eq})^{1/2}](u_1u_2u_3 - v_1v_2u_3 - v_1u_2v_3 - u_1v_2v_3) \\
&\qquad - [i/4(N_{p,\rm eq})^{1/2}][\sigma_{A2}\sigma_{A3}(u_{p1}u_2u_3 - v_{p1}v_2u_3 \\
&\qquad - v_{p1}u_2v_3 - u_{p1}v_2v_3) \\
&\qquad + \sigma_{A1}\sigma_{A3}(u_1u_{p2}u_3 - v_1v_{p2}u_3 - v_1u_{p2}v_3 - u_1v_{p2}v_3) \\
&\qquad + \sigma_{A1}\sigma_{A2}(u_1u_2u_{p3} - v_1v_2u_{p3} - v_1u_2v_{p3} - u_1v_2v_{p3}) \\
&\qquad + \sigma_{A3}(u_{p1}u_{p2}u_3 - v_{p1}v_{p2}u_3 - v_{p1}u_{p2}v_3 - u_{p1}v_{p2}v_3) \\
&\qquad + \sigma_{A2}(u_{p1}u_2u_{p3} - v_{p1}v_2u_{p3} - v_{p1}u_2v_{p3} - u_{p1}v_2v_{p3}) \\
&\qquad + \sigma_{A1}(u_1u_{p2}u_{p3} - v_1v_{p2}u_{p3} - v_1u_{p2}v_{p3} - u_1v_{p2}v_{p3}) \\
&\qquad \left. + (u_{p1}u_{p2}u_{p3} - v_{p1}v_{p2}u_{p3} - v_{p1}u_{p2}v_{p3} - u_{p1}v_{p2}v_{p3})] \right\}, \quad (10)
\end{aligned}
$$

where u_i, u_{pi}, $i = 1, 2, 3$, are carrying variables associated with A_i, A_{pi} $i = 1, 2, 3$, respectively; v_i, v_{pi}, $i = 1, 2, 3$, are carrying variables associated with B_i, B_{pi}, $i = 1, 2, 3$, respectively.

The Fourier transform of equation (10) in terms of polar coordinates provides the required joint probability distribution:

$$
\begin{aligned}
&P(R_i, R_{pi}, \varphi_i, \varphi_{pi}, i = 1, 2, 3) \\
&\quad = \prod_{i=1}^{3} \frac{R_i R_{pi}}{e_i - \sigma_{Ai}^2} \exp\left\{ -\sum_{i=1}^{3} \frac{1}{e_i - \sigma_{Ai}^2} \right. \\
&\qquad \left. \times [R_i^2 + e_i R_{pi}^2 - 2\sigma_{Ai} R_i R_{pi} \cos(\varphi_i - \varphi_{pi})] \right\} \\
&\qquad \times \exp\left\{ \frac{2}{(N_{p,\rm eq})^{1/2}} R_{p1}R_{p2}R_{p3} \cos(\varphi_{p1} + \varphi_{p2} + \varphi_{p3}) \right. \\
&\qquad + 2\beta[R_1R_2R_3 \cos(\varphi_1 + \varphi_2 + \varphi_3) \\
&\qquad - \sigma_{A3} R_1R_2R_{p3} \cos(\varphi_1 + \varphi_2 + \varphi_{p3}) \\
&\qquad - \sigma_{A2} R_1R_{p2}R_3 \cos(\varphi_1 + \varphi_{p2} + \varphi_3) \\
&\qquad - \sigma_{A1} R_{p1}R_2R_3 \cos(\varphi_{p1} + \varphi_2 + \varphi_3) \\
&\qquad + \sigma_{A2}\sigma_{A3} R_1R_{p2}R_{p3} \cos(\varphi_1 + \varphi_{p2} + \varphi_{p3}) \\
&\qquad + \sigma_{A1}\sigma_{A3} R_{p1}R_2R_{p3} \cos(\varphi_{p1} + \varphi_2 + \varphi_{p3}) \\
&\qquad + \sigma_{A1}\sigma_{A2} R_{p1}R_{p2}R_3 \cos(\varphi_{p1} + \varphi_{p2} + \varphi_3) \\
&\qquad \left. - \sigma_{A1}\sigma_{A2}\sigma_{A3} R_{p1}R_{p2}R_{p3} \cos(\varphi_{p1} + \varphi_{p2} + \varphi_{p3})] \right\}, \qquad (11)
\end{aligned}
$$

where

$$
\begin{aligned}
\beta &= \frac{1}{(e_1 - \sigma_{A1}^2)(e_2 - \sigma_{A2}^2)(e_3 - \sigma_{A3}^2)} \\
&\quad \times \left(\frac{1}{(N_{\rm eq})^{1/2}} - \frac{1}{(N_{p,\rm eq})^{1/2}} \sigma_{A1}\sigma_{A2}\sigma_{A3} \right).
\end{aligned}
$$

Equation (11) is the distribution we were looking for.

The parameter β plays a central role in equation (11) and deserves to be illustrated. It is easy shown (the reader should use the definitions given in §1) that

$$\beta = \prod_{i=1}^{3}\left(\frac{\sum_{Ni}}{e_i\sum_{Ni} - D_i^2\sum_{pi}}\right)^{1/2} \times \left(\frac{\sum_{j=1}^{N} f_{j1}f_{j2}f_{j3} - D_1D_2D_3\sum_{j=1}^{Np} f_{j1}f_{j2}f_{j3}}{\prod_{i=1}^{3}(e_i\sum_{Ni} - D_i^2\sum_{pi})^{1/2}}\right). \quad (12)$$

We notice:

(*a*) The term $(e_i\sum_{Ni} - D_i^2\sum_{pi})$ is the expected scattering power of the difference structure for the ith reflection (say $\sum_{qi}$), given hypothesis I. Indeed $e_i\sum_{Ni}$ is the scattering power of the target structure given the measurement error, and $\sum_{pi}$ is the equivalent scattering factor of the model structure. The difference $(e_i\sum_{Ni} - D_i^2\sum_{pi})$ takes into account the correlation between target and model structure: if $D_i = 1$ and $e_i \approx 1$ then $\sum_{qi} = \sum_{Ni} - \sum_{pi}$, the classical value in the literature. If no correlation exists, the scattering power of the difference structure increases and becomes equal to that of the target structure (say $\sum_{qi} = e_i\sum_{Ni}$). It is immediately seen that $\sum_{qi}$ does not coincide with the value $\sum_{pi}(1 - 2D_i) + \sum_{Ni}$ obtained by Burla, Caliandro *et al.*, 2010) *via* the study of the joint probability distribution $P(E_{\mathbf{h}}, E_{p\mathbf{h}}, E_{q\mathbf{h}})$: the discrepancy between the two estimates is the effect of the different amount of prior information available in the two distributions $P(E_{\mathbf{h}}, E_{p\mathbf{h}})$ and $P(E_{\mathbf{h}}, E_{p\mathbf{h}}, E_{q\mathbf{h}})$.

(*b*) The term

$$\frac{\sum_{j=1}^{N} f_{j1}f_{j2}f_{j3} - D_1D_2D_3\sum_{j=1}^{p} f_{j1}f_{j2}f_{j3}}{\prod_{i=1}^{3}(e_i\sum_{Ni} - D_i^2\sum_{pi})^{1/2}} \quad (13)$$

may be identified with $1/(N_{q,\mathrm{eq}})^{1/2}$. Indeed, if $D_i, e_i = 1$ for $i = 1, 2, 3$, then equation (13) reduces to equation (3); if the correlation between the target and model structures vanishes then $D_i = 0$ for $i = 1, 2, 3$, and equation (13) reduces to $1/N^{1/2}$, a value larger than $1/(N_q)^{1/2}$. In this condition the triplet phase reliability attains its minimum.

4. New phasing tools

The general distribution (11) may be used for obtaining two basic phasing formulas, the first for estimating triplet phase invariants given six diffraction moduli and the second for estimating $\varphi_{\mathbf{h}}$ given six moduli and five phases.

4.1. Triplet invariants estimates

From equation (11) standard mathematical techniques lead to the conditional distribution

$$P(\Phi|R_1, R_2, R_3, R_{p1}, R_{p2}, R_{p3}) \approx [2\pi I_0(G)]^{-1}\exp[G\cos\Phi], \quad (14)$$

where

$$G = \frac{2}{(N_{p,\mathrm{eq}})^{1/2}} m_1m_2m_3R_{p1}R_{p2}R_{p3} + 2\beta[(R_1 - \sigma_{A1}m_1R_{p1})(R_2 - \sigma_{A2}m_2R_{p2})(R_3 - \sigma_{A3}m_3R_{p3})]. \quad (15)$$

Furthermore,

$$P(\Phi_p|R_1, R_2, R_3, R_{p1}, R_{p2}, R_{p3}) \approx [2\pi I_0(G_p)]^{-1}\exp(G_p\cos\Phi_p), \quad (16)$$

where

$$G_p = \frac{2}{(N_{p,\mathrm{eq}})^{1/2}} R_{p1}R_{p2}R_{p3} + 2\beta[(m_1R_1 - \sigma_{A1}R_{p1})(m_2R_2 - \sigma_{A2}R_{p2})(m_3R_3 - \sigma_{A3}R_{p3})]. \quad (17)$$

Equations (14)–(17) suggest the following conclusions:

(i) If a model is available which is (weakly or strongly) correlated with the target structure, the triplet phase Φ is no longer expected to be always distributed around zero, as suggested by Cochran formula. Indeed the distribution (14) is centered around π (rather than around zero) when the second term on the right-hand side of equation (15) is sufficiently negative. The percentage of triplet phases close to π depends on the correlation between the target and model structures.

(ii) If diffraction data for two isomorphous structures have been measured and no model is available, then the triplet phase Φ_p is no longer expected to be always distributed around zero. Indeed, the distribution (16) is centered around π (rather than around zero) when the second term on the right-hand side of equation (17) is sufficiently negative. Again, the percentage of triplet phases close to π depends on the correlation between the target and model structures.

Let us simplify our expressions (15) and (17) by assuming $e_i = 1$ for $i = 1, 2, 3$: this is a reasonable assumption when the diffraction amplitudes are sufficiently large. We will consider two extreme cases:

(1) The target and model structure completely uncorrelated. In this case $D_i = 0$ for $i = 1, 2, 3$, $\sigma_{Ai} \approx m_i \approx 0$ for $i = 1, 2, 3$, $\beta = 1/(N_{\mathrm{eq}})^{1/2}$, $G = [2/(N_{\mathrm{eq}})^{1/2}]R_1R_2R_3$ and $G_p = [2/(N_{p,\mathrm{eq}})^{1/2}]R_{p1}R_{p2}R_{p3}$. In simple words, and in accordance with expectations, a uncorrelated model does not provide any information supplementary to that contained in the Cochran formula.

(2) The model and target structure are well correlated: *i.e.*, the N_p atomic positions of the model perfectly coincide with corresponding atomic positions in the target structure and the scattering power of the model structure is non-negligible with respect to the target. Then, for most of the reflections (*i.e.*, those with largest intensity), $D_i = m_i = 1$ for $i = 1, 2, 3$. In this case, the number of atoms in the difference structure is properly defined by $N_q = N - N_p$, or, in a more general way, by $N_{q,\mathrm{eq}}$ (see §2): furthermore, the scattering power $\sum_q$ of the difference structure reduces to $\sum_{qi} = \sum_{Ni} - \sum_{pi}$ for $i = 1, 2, 3$. We obtain

$$\sigma_{Ai} = (\textstyle\sum_{pi} / \sum_{Ni})^{1/2}, \quad \beta = \dfrac{1}{(N_{q,\mathrm{eq}})^{1/2}} \left(\dfrac{\sum_{N1} \sum_{N2} \sum_{N3}}{\sum_{q1} \sum_{q2} \sum_{q3}} \right)^{1/2},$$

$$G = \frac{2}{(N_{p,\mathrm{eq}})^{1/2}} R_{p1} R_{p2} R_{p3} + \frac{2}{(N_{q,\mathrm{eq}})^{1/2}} \frac{(|F_1| - |F_{p1}|)(|F_2| - |F_{p2}|)(|F_3| - |F_{p3}|)}{(\sum_{q1} \sum_{q2} \sum_{q3})^{1/2}} \tag{18}$$

and

$$G_p = \frac{2}{(N_{p,\mathrm{eq}})^{1/2}} R_{p1} R_{p2} R_{p3} + \frac{2}{(N_{q,\mathrm{eq}})^{1/2}} \frac{(|F_1| - |F_{p1}|)(|F_2| - |F_{p2}|)(|F_3| - |F_{p3}|)}{(\sum_{q1} \sum_{q2} \sum_{q3})^{1/2}}. \tag{19}$$

Equation (19) exactly coincides with equation (7) if the subscript p is replaced by P and $|F_i|$ by $|F_{\mathrm{di}}|$: accordingly, the Hauptman formula may be considered a special case of equation (19), valid only when N_p atoms of the target structure perfectly coincide with the N_p atoms of the model.

For large structures, equations (15)–(17) suggest the following conclusions: the second terms on the right-hand sides of equations (15) and (17) may be dominant with respect to the first terms. Thus, in protein cases, the triplet phase estimations cannot be based on Cochran formula, but should be allowed to benefit by the supplementary information provided by a model, even if it is weakly correlated with the target. The expressions (15)–(17) are therefore the general relations to use: they require the prior knowledge of the σ_{Ai} parameters, which in turn may be estimated by standard statistical methods, as suggested by Read (1986) or Burla, Giacovazzo, Mazzone *et al.* (2011).

4.2. Conditional phase estimates

From equation (11) the conditional distribution $P(\varphi_1 | R_1, R_2, R_3, R_{p1}, R_{p2}, R_{p3}, \varphi_2, \varphi_3, \varphi_{p1}, \varphi_{p2}, \varphi_{p3})$ may be obtained by standard techniques. We obtain

$$P(\varphi_1 | \ldots) = [2\pi I_0(G_1)]^{-1} \exp[G_1 \cos(\varphi_1 - \xi_1)], \tag{20}$$

where

$$G_1^2 = a_1^2 + a_2^2,$$

$$a_1 = 2R_1 \left\{ \frac{\sigma_{A1}}{e_1 - \sigma_{A1}^2} R_{p1} \cos\varphi_{p1} + \beta[R_2 R_3 \cos(\varphi_2 + \varphi_3) - \sigma_{A3} R_2 R_{p3} \cos(\varphi_2 + \varphi_{p3}) - \sigma_{A2} R_{p2} R_3 \cos(\varphi_{p2} + \varphi_3) + \sigma_{A2}\sigma_{A3} R_{p2} R_{p3} \cos(\varphi_{p2} + \varphi_{p3})] \right\},$$

$$a_2 = 2R_1 \left\{ \frac{\sigma_{A1}}{e_1 - \sigma_{A1}^2} R_{p1} \sin\varphi_{p1} + \beta[-R_2 R_3 \sin(\varphi_2 + \varphi_3) + \sigma_{A3} R_2 R_{p3} \sin(\varphi_2 + \varphi_{p3}) + \sigma_{A2} R_{p2} R_3 \sin(\varphi_{p2} + \varphi_3) - \sigma_{A2}\sigma_{A3} R_{p2} R_{p3} \sin(\varphi_{p2} + \varphi_{p3})] \right\},$$

$$\tan\xi_1 = a_2/a_1.$$

If the model and target structures are uncorrelated, (*i.e.*, $\sigma_{Ai} \approx m_i \approx 0$, for $i = 1, 2, 3$) then $\beta = 1/(N_{\mathrm{eq}})^{1/2}$, and

$$a_1 = \frac{2}{(N_{\mathrm{eq}})^{1/2}} R_1 R_2 R_3 \cos(\varphi_2 + \varphi_3),$$

$$a_2 = -\frac{2}{(N_{\mathrm{eq}})^{1/2}} R_1 R_2 R_3 \sin(\varphi_2 + \varphi_3),$$

in agreement with the Cochran formula. The expressions for a_1 and a_2 become much more useful if the correlation between the target and model structures is not vanishing. In the extreme case in which p atoms of the model perfectly coincide with p atoms of the target, then for the most intense reflections $D_i = m_i = 1$ for $i = 1, 2, 3$, and the distribution (4) is obtained. This result shows that equation (4) is a particular case of equation (20), only valid under conditions rarely met in practice.

5. The new reliability parameters

The distribution (8) deals with normalized structure factors of two isomorphous structures: as stated before, it is just for convenience that they were called 'target' and 'model', just for establishing a difference between the structure we want to phase (target) and the structure suggested by some source of information (model). From the mathematical point of view the target and model structures play a symmetrical role in equation (8) (the reader can be more easily convinced if they assume that $p = N$). In spite of this statement, equation (10) shows a lack of symmetry: in the triplet cosine terms σ_{Ai} is always associated to R_{pi}, while $\sigma_{Ai} R_i$, $i = 1, 2, 3$, are never present. The asymmetry is due to the following basic assumption (see §3): $\mathbf{r}_j$, $j = 1, \ldots, N$, are random variables, the first N_p of which (say $\mathbf{r}_j = \mathbf{r}_{pj} + \Delta\mathbf{r}_j$, $j = 1, \ldots, N_p$) are riding variables, correlated with the corresponding $\mathbf{r}_{pj}$'s through local positional errors $\Delta\mathbf{r}_j$'s.

It may be worthwhile comparing the reliability parameters (15) and (17) with (7). The first term on the right-hand side of equation (7) coincides with the Cochran term, while the second term is determined by structure-factor differences. As an effect of the lack of isomorphism (*i.e.*, the $\mathbf{r}_j$'s are variables freely riding over the primitive random variables $\mathbf{r}_{pj}$'s), the scenario is completely changed if we consider equation (15): we can no longer speak of the Cochran term and of difference terms. Indeed, the first term on the right-hand side of equation (15) is the Cochran parameter multiplied by the weight $w = m_1 m_2 m_3$, which may be very close to zero if the isomorphism between the model and target structures is poor.

Let us consider two asymptotic situations: if the isomorphism is very poor, the Cochran term coincides with the second term on the right-hand side of equation (15); if the isomorphism is strong, the Cochran term coincides with the first term on the right-hand side of equation (15). If the current situation is not asymptotic, the Cochran contribution arises from both terms on the right-hand side of equation (15).

Let us now compare the parameters (7) and (17). In agreement with hypothesis 1, Φ_p as calculated from distribution (16) does not suffer by riding effects, and the first term on the right-hand side of equation (17) coincides with the traditional Cochran parameter for the model triplet.

6. The asymmetry of the distribution $P(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{-\mathbf{h}-\mathbf{k}}, E_{p\mathbf{h}}, E_{p\mathbf{k}}, E_{p,-\mathbf{h}-\mathbf{k}})$

The distribution (8) can be studied under a different assumption (hypothesis 2): the target atomic positions $\mathbf{r}_j$, $j = 1, \ldots, N$, are primitive random variables, and $\mathbf{r}_{pj} = \mathbf{r}_j + \Delta\mathbf{r}_j$, $j = 1, \ldots, N_p$, are riding variables, correlated with the corresponding $\mathbf{r}_j$'s through local positional errors $\Delta\mathbf{r}_j$'s. In symbols

$$A = \left\{ \sum_{j=1}^{N} f_j \cos 2\pi\mathbf{h}\mathbf{r}_j + |\mu| \cos\vartheta \right\} \Big/ \left(\varepsilon\Sigma_N\right)^{1/2},$$

$$B = \left\{ \sum_{j=1}^{N} f_j \sin 2\pi\mathbf{h}\mathbf{r}_j + |\mu| \sin\vartheta \right\} \Big/ \left(\varepsilon\Sigma_N\right)^{1/2},$$

$$A_p = \sum_{j=1}^{Np} f_j \cos[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)] \Big/ \left(\varepsilon\Sigma_p\right)^{1/2},$$

$$B_p = \sum_{j=1}^{Np} f_j \sin[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)] \Big/ \left(\varepsilon\Sigma_p\right)^{1/2}.$$

Of the two basic hypotheses, the first emphasizes that some positional vectors of the target structure are distributed around the model positional vectors, the second that the model positional vectors (the only ones available during the phasing process) are expected to be distributed around the target positions. To see how the characteristic function (10) changes under hypothesis 2, we calculate the value of $\langle A_{\mathbf{h}} A_{\mathbf{k}} A_{p-\mathbf{h}-\mathbf{k}} \rangle$ (the coefficient of $u_1u_2u_{p3}$) according to both hypotheses. We obtain, according to the first,

$$\begin{aligned}
&\langle A_{\mathbf{h}} A_{\mathbf{k}} A_{p-\mathbf{h}-\mathbf{k}} \rangle \\
&\quad = \left\langle \frac{1}{\left(\sum_{N1}\sum_{N2}\sum_{p3}\right)^{1/2}} \sum_{j=1}^{Np} f_j(\mathbf{h}) f_j(\mathbf{k}) f_j(\mathbf{h}+\mathbf{k}) \right. \\
&\qquad \left. \times \cos 2\pi\mathbf{h}(\mathbf{r}_{pj} + \Delta\mathbf{r}_j) \cos 2\pi\mathbf{k}(\mathbf{r}_{pj} + \Delta\mathbf{r}_j) \cos[2\pi(\mathbf{h}+\mathbf{k})\mathbf{r}_{pj}] \right\rangle \\
&\quad = \frac{1}{4} \frac{1}{(N_{p,\mathrm{eq}})^{1/2}} \sigma_{A1}\sigma_{A2},
\end{aligned}$$

and according to the second hypothesis,

$$\begin{aligned}
&\langle A_{\mathbf{h}} A_{\mathbf{k}} A_{p-\mathbf{h}-\mathbf{k}} \rangle \\
&\quad = \left\langle \frac{1}{\left(\sum_{N1}\sum_{N2}\sum_{p3}\right)^{1/2}} \sum_{j=1}^{N_p} f_j(\mathbf{h}) f_j(\mathbf{k}) f_j(\mathbf{h}+\mathbf{k}) \right. \\
&\qquad \left. \times \cos 2\pi\mathbf{h}\mathbf{r}_j \cos 2\pi\mathbf{k}\mathbf{r}_j \cos[2\pi(\mathbf{h}+\mathbf{k})(\mathbf{r}_j + \Delta\mathbf{r}_j)] \right\rangle \\
&\quad = \frac{1}{4} \frac{1}{(N_{p,\mathrm{eq}})^{1/2}} \sigma_{A3}.
\end{aligned}$$

As a consequence, under hypothesis 2, the following characteristic function arises:

$$\begin{aligned}
&C(u_i, u_{pi}, v_i, v_{pi}, i = 1, 2, 3) \\
&\quad = \exp\left\{ -\frac{1}{4}\sum_{i=1}^{3} [e_i(u_i^2 + v_i^2) + u_{pi}^2 + v_{pi}^2 + 2\sigma_{Ai}(u_i u_{pi} + v_i v_{pi})] \right. \\
&\qquad - \frac{i}{4(N_{\mathrm{eq}})^{1/2}} (u_1u_2u_3 - v_1v_2u_3 - v_1u_2v_3 - u_1v_2v_3) \\
&\qquad - \frac{i}{4(N_{p,\mathrm{eq}})^{1/2}} [\sigma_{A1}(u_{p1}u_2u_3 - v_{p1}v_2u_3 - v_{p1}u_2v_3 - u_{p1}v_2v_3) \\
&\qquad + \sigma_{A2}(u_1u_{p2}u_3 - v_1v_{p2}u_3 - v_1u_{p2}v_3 - u_1v_{p2}v_3) \\
&\qquad + \sigma_{A3}(u_1u_2u_{p3} - v_1v_2u_{p3} - v_1u_2v_{p3} - u_1v_2v_{p3}) \\
&\qquad + \sigma_{A1}\sigma_{A2}(u_{p1}u_{p2}u_3 - v_{p1}v_{p2}u_3 - v_{p1}u_{p2}v_3 - u_{p1}v_{p2}v_3) \\
&\qquad + \sigma_{A1}\sigma_{A3}(u_{p1}u_2u_{p3} - v_{p1}v_2u_{p3} - v_{p1}u_2v_{p3} - u_{p1}v_2v_{p3}) \\
&\qquad + \sigma_{A2}\sigma_{A3}(u_1u_{p2}u_{p3} - v_1v_{p2}u_{p3} - v_1u_{p2}v_{p3} - u_1v_{p2}v_{p3}) \\
&\qquad \left. + (u_{p1}u_{p2}u_{p3} - v_{p1}v_{p2}u_{p3} - v_{p1}u_{p2}v_{p3} - u_{p1}v_{p2}v_{p3})] \right\}. \quad (21)
\end{aligned}$$

The joint probability distribution corresponding to the characteristic function (21) is certainly different from the distribution (11), but it will be not calculated in this paper. Indeed, we are mainly interested in estimating the target given the model phases and that implies a selection of the distribution to use. In real cases the parameters $\mathbf{r}_{pj}$ are perfectly available during the phasing process, as well as the R_p and φ_p values. In this situation the conditional probability $P(\Phi_p|R_1, R_2, R_3, R_{p1}, R_{p2}, R_{p3})$ adheres perfectly to hypothesis 1.

An interesting detail is the following: the characteristic functions (10) and (21) coincide up to terms of order N^0: thus the distribution $P(E_{\mathbf{h}}, E_{p\mathbf{h}})$, as obtained by Srinivasan & Ramachandran (1965) and generalized by Caliandro *et al.* (2005), does not change, regardless of whether we use hypothesis 1 or 2.

7. Conclusions

The Cochran reliability parameter for triplet phase invariants is based on the conditional probability distribution $P(\Phi|\ R_1, R_2, R_3)$: accordingly it cannot exploit the information contained in the models eventually available during the phasing process. In particular, during the tangent-formula cycles phases progressively change: correspondently, in direct space new structural models become available, but, according

to the Cochran scheme, every triplet phase is constantly estimated *via* the same probability parameter.

This paper establishes the theoretical bases for a radical change of the traditional direct-methods approach: the Cochran formula should no longer be necessary. Indeed the new formulas (14)–(20) are based on the distributions

$$P(\Phi|R_1, R_2, R_3, R_{p1}, R_{p2}, R_{p3})$$

and

$$P(\varphi_1|R_1, R_2, R_3, R_{p1}, R_{p2}, R_{p3}, \varphi_2, \varphi_3, \varphi_{p1}, \varphi_{p2}, \varphi_{p3}),$$

respectively, which can take into account both the observations and the models progressively available during the phasing process.

The application of equations (14)–(20) require the previous estimate of the quality of the model. That is made *via* the statistical approaches described by Read (1986) or Burla, Giacovazzo, Mazzone *et al.* (2011). If the quality is poor (*i.e.*, when the σ_{Ai}'s are close to zero), the model contribution in equations (15), (17) and (20) will be weak, but it will increase as soon as the σ_{Ai} values increase.

The practical use of equation (20) deserves some additional remarks. Equation (20) provides a new value for φ_1 given five phases and six moduli. While φ_{p1}, φ_{p2} and φ_{p3} are perfectly known from the model, φ_2 and φ_3 are unknown because they refer to the target structure. On the other hand, the assumptions $\varphi_2 \approx \varphi_{p2}$ and $\varphi_3 \approx \varphi_{p3}$:

(i) presuppose that a high correlation exists between target and model: that is the less interesting case for this paper, which aims at phasing the target also starting from weakly correlated or uncorrelated model structures;

(ii) may trap the phase refinement away from the correct values, probably not far from the initial φ_p values. Again that does not comply with the hypothesis that the model and target structures may be weakly correlated or uncorrelated.

The problem may be solved exactly as proposed and effectively used by the *VLD* algorithm, in which, given the best available map (regardless of its quality), the corresponding phases play the φ role, while the φ_p's are obtained from a suitably modified electron-density map (*e.g.* by Fourier inversion of a small percentage of the map, that with the largest density values). The application of equation (20) is expected to progressively improve the phases φ and therefore the model phases φ_p, as in *VLD* procedures.

If we consider phasing from a random model, the phasing scenario may be the following: the procedure starts from a random phase set, as in traditional direct-methods applications. Each cycle of tangent formula, based on equation (20), would produce a new set of phases from which a model electron-density map may be derived. This model may be used to the improve the estimation of the triplet invariants, and so on in a cyclic way. This practice is expected to make triplet invariants more suitable for larger structures.

References

Buerger, M. J. (1959). *Vector Space.* New York: Wiley.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *J. Appl. Cryst.* **38**, 381–388.

Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2006). *J. Appl. Cryst.* **39**, 527–535.

Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2007). *J. Appl. Cryst.* **40**, 211–217.

Burla, M. C., Caliandro, R., Giacovazzo, C. & Polidori, G. (2010). *Acta Cryst.* A**66**, 347–361.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2011). *J. Appl. Cryst.* **44**, 1143–1151.

Burla, M. C., Giacovazzo, C., Mazzone, A., Polidori, G. & Siliqi, D. (2011). *Acta Cryst.* A**67**, 276–283.

Burla, M. C., Giacovazzo, C. & Polidori, G. (2010). *J. Appl. Cryst.* **43**, 825–836.

Burla, M. C., Giacovazzo, C. & Polidori, G. (2011). *J. Appl. Cryst.* **44**, 193–199.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2008). *J. Appl. Cryst.* **41**, 548–553.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Moustiakimov, M. & Siliqi, D. (2005). *Acta Cryst.* A**61**, 343–349.

Camalli, M., Giacovazzo, C. & Spagna, R. (1985). *Acta Cryst.* A**41**, 605–613.

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**, 473–478.

Dumas, C. & van der Lee, A. (2008). *Acta Cryst.* D**64**, 864–873.

Giacovazzo, C. (1977). *Acta Cryst.* A**33**, 933–944.

Giacovazzo, C. (1980). *Acta Cryst.* A**36**, 362–372.

Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 685–692.

Giacovazzo, C. (1998). *Direct Phasing in Crystallography.* Oxford University Press.

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

Hauptman, H. (1976). *Acta Cryst.* A**32**, 934–940.

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

Hauptman, H. & Karle, J. (1956). *Acta Cryst.* **9**, 45–55.

Heinerman, J. J. L. (1977). *Acta Cryst.* A**33**, 100–106.

Heinerman, J. J. L., Krabbendam, H. & Kroon, J. (1977). *Acta Cryst.* A**33**, 873–878.

Karle, J. & Karle, I. L. (1966). *Acta Cryst.* **21**, 849–859.

Main, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, pp. 97–105. Copenhagen: Munksgaard.

Main, P. (1979). *Acta Cryst.* A**35**, 779–785.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

Oszlányi, G. & Sütő, A. (2004). *Acta Cryst.* A**60**, 134–141.

Oszlányi, G. & Sütő, A. (2005). *Acta Cryst.* A**61**, 147–152.

Palatinus, L. & Chapuis, G. (2007). *J. Appl. Cryst.* **40**, 786–790.

Palatinus, L., Dušek, M., Glaum, R. & El Bali, B. (2006). *Acta Cryst.* B**62**, 556–566.

Pavelčík, F., Kuchta, L. & Sivý, J. (1992). *Acta Cryst.* A**48**, 791–796.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Richardson, J. W. & Jacobson, R. A. (1987). *Patterson and Pattersons*, edited by J. P. Glusker, B. K. Patterson & M. Rossi, pp. 310–317. Oxford University Press.

Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publishers.

Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.

Yao, J. X., Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *Acta Cryst.* D**62**, 901–908.