Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 30 November 2009
Accepted 29 January 2010

The difference electron density: a probabilistic reformulation

Maria Cristina Burla,[a] Rocco Caliandro,[b] Carmelo Giacovazzo[b,c]* and Giampiero Polidori[a]

[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 Geomineralogico, Università di Bari, 70125 Bari, Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

The joint probability distribution function $P(E, E_p)$, where E and E_p are the normalized structure factors of the target and of a model structure, respectively, is a fundamental tool in crystallographic methods devoted to crystal structure solution. It plays a central role in any attempt for improving phase estimates from a given structure model. More recently the difference electron density $\rho_q = \rho - \rho_p$ has been revisited and methods based on its modifications have started to play an important role in combination with electron density modification approaches. In this paper new coefficients for the difference electron density have been obtained by using the joint probability distribution function $P(E, E_p, E_q)$ and by taking into account both errors in the model and in measurements. The first applications show the correctness of our theoretical approach and the superiority of the new difference Fourier synthesis, particularly when the model is a rough approximation of the target structure. The new and the classic difference syntheses coincide when the model represents the target structure well.

1. Notation

ρ, ρ_p: electron densities of the target and of the model structure, respectively.
$\rho_q = \rho - \rho_p$: *ideal difference Fourier synthesis*; summed to ρ_p it exactly provides ρ, no matter the quality of ρ_p.
N: number of atoms in the unit cell for the target structure.
p: number of atoms in the unit cell for the model structure; usually $p \leq N$, but it may also be $p > N$.
f_j, $j = 1, \ldots, N$: atomic scattering factors for the target structure (thermal factor included).
$F = \Sigma_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j) = |F|\exp(i\varphi)$: classical expression for the structure factor of the target structure; in our model, F, added by the experimental error, will represent the observed structure factor.
$F_p = \Sigma_{j=1}^{p} f_j \exp(2\pi i \mathbf{h}\mathbf{r}'_j) = |F_p|\exp(i\varphi_p)$, where $\mathbf{r}'_j = \mathbf{r}_j + \Delta\mathbf{r}_j$: structure factor of the model structure.
$F_q = F - F_p = |F_q|\exp(i\varphi_q)$: structure factor of the *ideal difference structure*.
$E = A + iB = R\exp(i\varphi)$, $E_p = A_p + iB_p = R_p\exp(i\varphi_p)$, $E_q = A_q + iB_q = R_q\exp(i\varphi_q)$: normalized structure factors of F, F_p and F_q, respectively.
$R'_p = |F_p|/\Sigma_N^{1/2}$, $R'_q = |F_q|/\Sigma_N^{1/2}$: structure factors of the model and of the difference structure pseudonormalized with respect to Σ_N.
$\Sigma_N = \Sigma_{j=1}^{N} f_j^2$.
$\Sigma_p = \Sigma_{j=1}^{p} f_j^2$.
$D = \langle\cos(2\pi\mathbf{h}\Delta\mathbf{r})\rangle$: the average is performed per resolution shell.
$\sigma_A = D(\Sigma_p/\Sigma_N)^{1/2}$.
$\sigma_R^2 = \langle|\mu|^2\rangle/\Sigma_N$, where $\langle|\mu|^2\rangle$ 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 $X = 2\sigma_A R R_p/(e - \sigma_A^2)$.

2. Introduction

The joint probability distribution of two normalized structure factors $P(E, E_p)$ relative to two isomorphous structures (the target and the model structure, respectively; the latter is usually part of the former and shows discrepancies in the atomic coordinates) is an important tool for the solution of the phase problem. For example, it is often employed to drive the model phases towards the phases of the target structure. The interest started with Luzzati (1952), who studied the statistical effects on the structure factors of the errors owing to lack of isomorphism. Sim (1959) provided the probability distribution of the target structure factor phases when a model without errors is available. His theory associates a suitable weight to the coefficients of the observed Fourier synthesis, so improving its efficiency. Srinivasan & Ramachandran (1965) derived the probability of the observed structure factors in a

doi:10.1107/S010876731000365X

more general case, when the model atoms show errors in the coordinates. Read (1986) approximated the likelihood function given by Lunin & Urzhumtsev (1984) to provide the probability of the structure factor magnitudes given errors in the parameters of the located atoms. General applications of the previous contributions were described by Murshudov *et al.* (1997), Lunin *et al.* (2002) and Cowtan (2002). Caliandro *et al.* (2005) derived a general expression for $P(E, E_p)$ when both measurement errors and errors in the model structure are present.

The central role of the distribution $P(E, E_p)$ arises also from a supplementary circumstance: the differences $(|E| - |E_p|)\exp(i\varphi_p)$ are the classical coefficients of the *difference electron density*. Read (1986) suggested replacing them by more suitable differences

$$\big(m|E| - \sigma_A|E_p|\big)\exp\big(i\varphi_p\big), \tag{1}$$

which may be considered, in absence of any prior supplementary information, as the most accurate approximation of E_q, the normalized structure factor of the *ideal difference electron density*.

The recently proposed *DEDM* (difference electron density modification) algorithm, based on the modification of the difference electron density (Caliandro *et al.*, 2008), opened new perspectives for the recovery of the target structure from a model. This approach aims at providing more accurate estimates of E_q by breaking down the collinearity between model and target phases. Indeed, once better estimates of E_q are obtained, $E_p + E_q$ will provide more accurate phase values for the target structure. Caliandro *et al.* (2009*a*,*b*,*c*) combined *DEDM* with electron density modification (EDM) techniques (Cowtan, 1994, 1999; Abrahams, 1997; Abrahams & Leslie, 1996; Zhang *et al.*, 2001; Refaat & Woolfson, 1993; Giacovazzo & Siliqi, 1997). The combination led very imperfect models to converge to the target structure.

The above considerations suggest that an important role in modern phasing techniques may be played by E_q if more accurate estimates of its value become available. There are cases in which R_q is experimentally known (like in isomorphous derivative techniques, where E_p coincides with the normalized structure factor of the heavy-atom substructure, and R_q is the normalized diffraction modulus of the native protein) and cases in which R_q is the unknown structure factor modulus of the difference between the target and the model electron density (*i.e.* $\rho_q = \rho - \rho_p$). We will show that the study of the six-variate distribution $P(R, R_p, R_q, \varphi, \varphi_p, \varphi_q)$ can lead to more accurate estimates of E_q than *via* the four-variate distribution $P(E, E_p)$. This study is the main aim of this paper. From the formulas relating E, E_p, E_q, new coefficients for the difference Fourier synthesis are obtained: they are the sum of the classic structure factor difference term and of a flipping term, which is dominant when the model is a poor approximation of the target structure. The first applications of our theoretical results are also described.

3. About the mathematical model

Since the choice of the modelling influences all the theoretical results, let us consider, as a first step, the modelling criteria at the basis of our approach. The four-variate distribution $P(A, B, A_p, B_p)$ has been studied (Caliandro *et al.*, 2005) by the following structure factor model,

$$\begin{aligned}
A &= \Big[\textstyle\sum_{j=1}^{N} f_j \cos(2\pi\mathbf{hr}_j) + |\mu|\cos\theta\Big]\big/\big(\varepsilon\Sigma_N\big)^{1/2},\\
B &= \Big[\textstyle\sum_{j=1}^{N} f_j \sin(2\pi\mathbf{hr}_j) + |\mu|\sin\theta\Big]\big/\big(\varepsilon\Sigma_N\big)^{1/2},\\
A_p &= \textstyle\sum_{j=1}^{p} f_j \cos\big[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)\big]\big/\big(\varepsilon\Sigma_p\big)^{1/2},\\
B_p &= \textstyle\sum_{j=1}^{p} f_j \sin\big[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)\big]\big/\big(\varepsilon\Sigma_p\big)^{1/2},
\end{aligned}$$

where $\mu\exp(i\theta)$ is the (complex) error and ε is the correction factor for expected intensities in reciprocal-lattice zones (from Wilson statistics).

Sim's (1959) results correspond to the assumption $|\mu| = 0$, $\Delta\mathbf{r}_j = 0, j = 1, \ldots, p$ (*i.e.* no error in measurements, no error in the model structure). Srinivasan & Ramachandran's (1965) and Read's (1986) results correspond to the case $|\mu| = 0$ and non-vanishing $\Delta\mathbf{r}_j$ vectors. The general expression for the four-variate distribution is (Caliandro *et al.*, 2005)

$$\begin{aligned}
P(R, R_p, \varphi, \varphi_p) = RR_p\pi^{-2}(e - \sigma_A^2)^{-1}\exp\bigg\{-\frac{1}{(e - \sigma_A^2)}\\
\times\big[R^2 + eR_p^2 - 2\sigma_A RR_p\cos(\varphi - \varphi_p)\big]\bigg\}.
\end{aligned} \tag{2}$$

Six-variate distributions of type $P(A, B, A_p, B_p, A_q, B_q)$ were studied by Giacovazzo & Siliqi (2002) to treat the *SIR* case and by Giacovazzo & Siliqi (2001) to treat the *SAD* case (in this latter case complex scattering factors were assumed). For *SIR* the following structure factor model was used: (E, E_p, E_q) represent normalized structure factors of the derivative, of the heavy-atom substructure and of the protein, respectively. Then,

$$\begin{aligned}
A &= \Big[\textstyle\sum_{j=1}^{N} f_j \cos(2\pi\mathbf{hr}_j) + |\mu|\cos\theta\Big]\big/\big(\varepsilon\Sigma_N\big)^{1/2},\\
B &= \Big[\textstyle\sum_{j=1}^{N} f_j \sin(2\pi\mathbf{hr}_j) + |\mu|\sin\theta\Big]\big/\big(\varepsilon\Sigma_N\big)^{1/2},\\
A_p &= \textstyle\sum_{j=1}^{p} f_j \cos(2\pi\mathbf{hr}_j)\big/\big(\varepsilon\Sigma_p\big)^{1/2},\\
B_p &= \textstyle\sum_{j=1}^{p} f_j \sin(2\pi\mathbf{hr}_j)\big/\big(\varepsilon\Sigma_p\big)^{1/2},\\
A_q &= \Big[\textstyle\sum_{j=p+1}^{N} f_j \cos(2\pi\mathbf{hr}_j)\Big]\big/\big(\varepsilon\Sigma_q\big)^{1/2},\\
B_q &= \Big[\textstyle\sum_{j=p+1}^{N} f_j \sin(2\pi\mathbf{hr}_j)\Big]\big/\big(\varepsilon\Sigma_q\big)^{1/2}.
\end{aligned} \tag{3}$$

In this case $|\mu|\exp(i\theta)$ was assumed to represent the cumulative error, the components of which are errors due to lack of isomorphism, error in measurements and errors in the heavy-atom substructure. This modelling has the following limit: the errors owing to lack of isomorphism are not represented as differences between the atomic coordinates of the derivative and of the protein (*i.e.* the derivative model contains, as a subset and without any modification, the atomic positions of the protein). In practice the lack of isomorphism was included in the mathematical model just to increase the size of measurement errors.

Such a mathematical model is not adequate for the case in which derivative data are not available: indeed the deviations of the model from the target electron density may be dominant in the case of a poor model structure. In the next section we will introduce a more realistic mathematical approach using again the six-variate distributions: it will lead to phase relationships which encompass previous results.

The reader more interested in the practical aspects of the theory and to its applications may more carefully read from §9 onwards, where the first practical result of the theory, the definition of the coefficients for the calculation of a new difference Fourier synthesis, is given.

4. The bases of our probabilistic approach

The distribution

$$P(R, R_p, R_q, \varphi, \varphi_p, \varphi_q) \tag{4}$$

has a practical value only if (i) measurement errors are included in the mathematical model; (ii) σ_A (to be calculated between the model and the target structure) is not unity.

To clarify this important point let us return back to the definition of ρ_q (see §1). It is the *ideal difference Fourier synthesis*: it is not positive-definite and, by definition, when summed to ρ_p it exactly provides ρ, no matter the quality of ρ_p. Under this hypothesis, if condition (i) is violated, then $F_q = F - F_p$ is perfectly determined by the other two variables, and its distribution reduces to the Dirac delta function $\delta[F_q - (F - F_p)]$: indeed it vanishes for any value $F_q \neq (F - F_p)$ and the integral of the distribution is equal to unity.

If condition (ii) is violated (*i.e.* $\sigma_A = 1$) then $\rho_p \equiv \rho$ and $\rho_q \equiv 0$: again we do not need to calculate a six-variate distribution. Indeed F_q will be identically equal to zero and the F_q distribution will coincide with the Dirac delta function $\delta(F_q)$.

According to the above considerations, a correctly calculated six-variate distribution is expected to diverge when $\sigma_A = 1$ and/or when $e = 1$. In the first case the distribution (4) degrades to the distribution (2); in the second case the distribution (4) degrades to the classical Srinivasan & Ramachandran (1965) distribution which may be obtained from (2) by setting $e = 1$.

A different point of view may be obtained by considering the distribution (4) in the parameter plane (σ_A, e). When $\sigma_A = 1$ or $e = 1$ (*e.g.* in correspondence of two straight lines in the plane) the distribution (4) diverges: by no means does this imply a lack of accuracy, but only a minor usefulness of the distribution because the prior information perfectly defines the variable F_q. The situation is similar to that occurring for (2) when $\sigma_A = 1$ and $e = 1$, and for the classical Srinivasan & Ramachandran distribution when $\sigma_A = 1$: then $\rho_p \equiv \rho$ from which the identity $F = F_p$ arises, which corresponds to the maximum of prior information, making the four-variate distribution very accurate but not useful.

While it is clear what to do when $\sigma_A = 1$ [we just do not need the distribution (4), because it is overdetermined from the prior information] it may not be clear what to do in the ideal case in which $e = 1$. Then the chosen definition of F_q does not allow us to calculate the distribution (4) (again F_q is determined by the prior information) and the six-dimensional distribution should be degraded to a four-dimensional one. However, we will see below that the practical use of (4) is not critical when e comes near unity, exactly as occurs for (2) when σ_A becomes closer and closer to 1 (or when ρ_p approaches ρ).

These circumstances oblige us to adopt a general mathematical approach, in which the errors in the model structure are accompanied by errors in measurements. We will calculate the joint probability distribution (4) under the following conditions:

(*a*) The coordinates of the vectors $\mathbf{r}_j$, $j = 1, \ldots, N$, are the primitive random variables, assumed to be uniformly distributed in the unit cell.

(*b*) The variables $\Delta\mathbf{r}_j$, $j = 1, \ldots, p$, are local variables randomly distributed around zero. In the absence of any information on their distribution and on their mutual correlation we will assume that they are independent of each other and uniformly distributed around zero. In many practical problems this condition is violated (*e.g.* when molecular fragments of the model are rotated or translated with respect to the correct orientation or position), but in the absence of supplementary information it is the less demanding hypothesis we can assume. The same assumption coincides in practice with that usually employed for the calculation of the σ_A parameter.

(*c*) Two supplementary primitive random variables, μ and θ, are considered, arising from the experimental uncertainty on the observed structure factor moduli. We will write

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

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

Accordingly, we will adopt the following general mathematical model,

$$\begin{aligned}
A &= \Big[\sum_{j=1}^{N} f_j \cos(2\pi\mathbf{h}\mathbf{r}_j) + |\mu|\cos\theta\Big]\Big/\big(\varepsilon\Sigma_N\big)^{1/2},\\
B &= \Big[\textstyle\sum_{j=1}^{N} f_j \sin(2\pi\mathbf{h}\mathbf{r}_j) + |\mu|\sin\theta\Big]\Big/\big(\varepsilon\Sigma_N\big)^{1/2},\\
A_p &= \textstyle\sum_{j=1}^{p} f_j \cos[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)]\big/\big(\varepsilon\Sigma_p\big)^{1/2},\\
B_p &= \textstyle\sum_{j=1}^{p} f_j \sin[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)]\big/\big(\varepsilon\Sigma_p\big)^{1/2},\\
A_q &= \Big\{\textstyle\sum_{j=1}^{N} f_j \cos(2\pi\mathbf{h}\mathbf{r}_j) - \sum_{j=1}^{p} f_j \cos[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)]\Big\}\\
&\quad \big/\big(\varepsilon\Sigma_q\big)^{1/2},\\
B_q &= \Big\{\textstyle\sum_{j=1}^{N} f_j \sin(2\pi\mathbf{h}\mathbf{r}_j) - \sum_{j=1}^{p} f_j \sin[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)]\Big\}\\
&\quad \big/\big(\varepsilon\Sigma_q\big)^{1/2}.
\end{aligned} \tag{5}$$

We explicitly recall the reader's attention to the differences between the F_q definitions in equations (5) and (3): we will show below that they have important mathematical consequences. Equations (3) assume that only the $N - p$ atoms, not included in the model structure, contribute to F_q: in particular, no contribution to F_q arises from the errors in the model structure. On the contrary, equations (4) explicitly include in F_q both model distortions and the contributions of the $N - p$ atoms. The assumption (3) implies that ρ_q is always positive, while the assumption (5) implies the presence of positive and negative peaks. It is also noted that ρ_q, defined by equations (5), is exactly the information the crystallographer needs to attain the target from the model structure: indeed, no matter whether the model is a poor or a good approximation of the target structure, it is always $\rho_p + \rho_q = \rho$ by definition (measurement errors excluded). To have a simple graphical representation of our definitions, we show in Figs. 1–3 ρ, ρ_p and ρ_q in three typical cases:

(*a*) In Fig. 1, ρ_p is an imperfect partial model, for which both σ_A and D do not coincide either with 0 or with 1. Then ρ_q is constituted by $q = N - p$ positive peaks and by p pairs of positive and negative peaks which arise from the model structure errors.

(*b*) In Fig. 2, ρ_p is a perfect partial model ($\sigma_A \neq 1$ and $D = 1$). In this case the ρ_p positive peaks perfectly overlap with p of the N ρ peaks, and ρ_q is a positive definite function constituted by $q = N - p$ peaks.

(*c*) In Fig. 3, ρ_p has completely lost its isomorphism with ρ up to the limit case in which $D = 0$: then ρ_q is constituted by the N positive peaks of ρ and by the p negative peaks of ρ_p. The two sets do not overlap.

An exact Σ_q estimate is seldom available in most practical cases: indeed Σ_q is a D-dependent parameter, as the following relationship suggests:

$$\Sigma_q = \langle |F_q|^2 \rangle = \Sigma_N + \Sigma_p - 2D\Sigma_p = \Sigma_p(1 - 2D) + \Sigma_N.$$

Accordingly, Σ_q depends on the quality of the model: it tends to $\Sigma_N - \Sigma_p$ when $D = 1$, and to $\Sigma_N + \Sigma_p$ when ρ_p progressively loses (up to $D = 0$) its isomorphism with ρ.

From the above assumptions the following relations are obtained:

$$\langle EE_p \rangle = \sigma_A, \tag{6}$$

$$\langle E_p E_q \rangle = \frac{(D-1)\Sigma_p}{\Sigma_p^{1/2}\Sigma_q^{1/2}} = \frac{\sigma_A \Sigma_N^{1/2} - \Sigma_p^{1/2}}{\Sigma_q^{1/2}}, \tag{7}$$

$$\langle EE_q \rangle = \frac{\Sigma_N - D\Sigma_p}{\Sigma_N^{1/2}\Sigma_q^{1/2}} = \frac{\Sigma_N^{1/2} - \sigma_A \Sigma_p^{1/2}}{\Sigma_q^{1/2}}. \tag{8}$$

We note the following:

(*a*) Since $0 \leq \sigma_A \leq 1$, $\langle |EE_p| \cos(\varphi - \varphi_p) \rangle$ is expected to be non-negative: its value should increase (up to 1) when the model becomes closer to the target structure.

(*b*) Since $0 \leq D \leq 1$, $\langle |E_p E_q| \cos(\varphi_p - \varphi_q) \rangle$ is expected to be non-positive definite. It tends to zero in the case of good isomorphism, strongly negative in the case of a lack of isomorphism: its maximum negative value is $-(\Sigma_p/\Sigma_q)^{1/2} = -[\Sigma_p/(\Sigma_p + \Sigma_N)]^{1/2}$, which attains the value $-(1/2)^{1/2} \simeq -0.78$ for a complete (*e.g.* $\Sigma_p = \Sigma_N$) but incorrect (*e.g.* $\sigma_A = 0$) model. In general, E_p and E_q are always anticorrelated, and are uncorrelated only when $D = 1$. This result directly derives from the definition of ρ_q: its positive maxima are expected to lie in the region not frequented by the atoms in the model.

(*c*) E and E_q are positively correlated, particularly when $D = 0$: in this case $\langle |EE_q| \cos(\varphi - \varphi_q) \rangle = (\Sigma_N/\Sigma_q)^{1/2}$. Equation (8) suggests that having better estimates of E_q is particularly useful for the phasing process when this starts from a poor model (this paper aims at contributing to this subject). The correlation will diminish to zero as the model converges to the target structure: in this case $\sigma_A = 1$ and Σ_p converges to Σ_N.

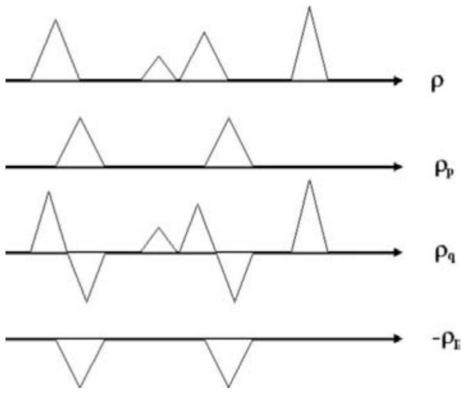

Figure 1
Schematic representation of ρ, ρ_p and ρ_q when D is different from 0 or 1.

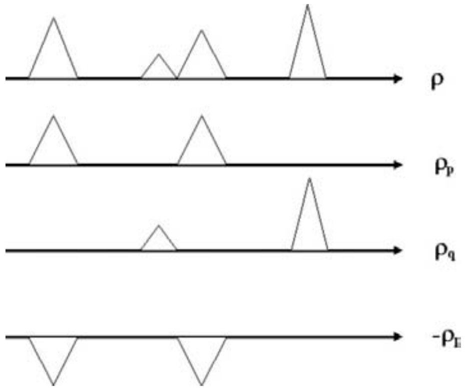

Figure 2
Schematic representation of ρ, ρ_p and ρ_q when $D = 1$.

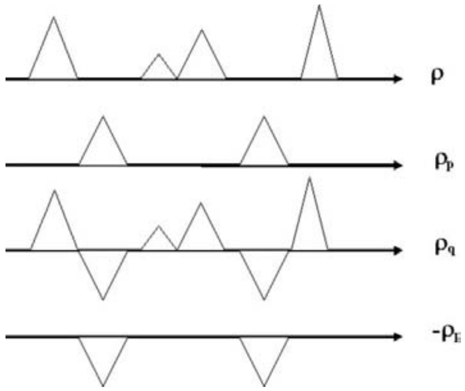

Figure 3
Schematic representation of ρ, ρ_p and ρ_q when $D = 0$.

To estimate the correlation between the pairs of a given triple E, E_p, E_q, the joint probability distribution function $P(E, E_p, E_q)$ should be calculated.

5. The joint probability distribution $P(E, E_p, E_q)$

The characteristic function of the distribution (4) is

$$\begin{aligned} C\big(u, u_p, u_q, v, v_p, v_q\big) &= \Big\langle \exp i\Big(uA + u_pA_p + u_qA_q + vB \\ &\quad + v_pB_p + v_qB_q\Big)\Big\rangle \\ &= \exp\Big\{ -(1/4)\Big[e(u^2+v^2) + (u_p^2+v_p^2) \\ &\quad + (u_q^2+v_q^2) + 2\sigma_A(uu_p + vv_p) \\ &\quad + 2\sigma_{Aq}(uu_q + vv_q) \\ &\quad + 2\sigma_{Apq}(u_pu_q + v_pv_q)\Big]\Big\}, \qquad (9) \end{aligned}$$

where u, u_p, u_q, v, v_p, v_q are carrying variables associated with $A, A_p, A_q\ B, B_p, B_q$, respectively,

$$e = (1 + \sigma_R^2),$$

$$\sigma_R^2 = \langle|\mu|^2\rangle/\Sigma_N,$$

$$\sigma_{Aq} = \frac{\Sigma_N^{1/2} - \sigma_A\Sigma_p^{1/2}}{\Sigma_q^{1/2}},$$

$$\sigma_{Apq} = \frac{\sigma_A\Sigma_N^{1/2} - \Sigma_p^{1/2}}{\Sigma_q^{1/2}}.$$

The distribution $P(A, A_p, A_q, B, B_p, B_q)$ is the Fourier transform of (9). In polar coordinates we obtain

$$\begin{aligned} &P(R, R_p, R_q, \varphi, \varphi_p, \varphi_q) \\ &\quad = RR_pR_q\pi^{-3}e^{-1}(\det\mathbf{L})^{-1}\exp\Big\{ -\frac{1}{e(\det\mathbf{L})}\big[(1-\sigma_{Apq}^2)R^2 \\ &\qquad + (e-\sigma_{Aq}^2)R_p^2 + (e-\sigma_A^2)R_q^2 \\ &\qquad + 2(\sigma_{Aq}\sigma_{Apq} - \sigma_A)RR_p\cos(\varphi-\varphi_p) \\ &\qquad + 2(\sigma_A\sigma_{Apq} - \sigma_{Aq})RR_q\cos(\varphi-\varphi_q) \\ &\qquad + 2(\sigma_A\sigma_{Aq} - e\sigma_{Apq})R_pR_q\cos(\varphi_p-\varphi_q)\big]\Big\} \qquad (10) \end{aligned}$$

where

$$\mathbf{L} = \begin{vmatrix} 1 & \sigma_A/e^{1/2} & \sigma_{Aq}/e^{1/2} \\ \sigma_A/e^{1/2} & 1 & \sigma_{Apq} \\ \sigma_{Aq}/e^{1/2} & \sigma_{Apq} & 1 \end{vmatrix},$$

$$\det\mathbf{L} = \left(1 - \frac{\sigma_A^2}{e} - \frac{\sigma_{Aq}^2}{e} - \sigma_{Apq}^2 + 2\frac{\sigma_A\sigma_{Aq}\sigma_{Apq}}{e}\right).$$

Let us write the above equation in the form

$$\begin{aligned} P\big(R, R_p, R_q, \varphi, \varphi_p, \varphi_q\big) &\cong \pi^{-3}e^{-1}(\det\mathbf{L})^{-1}RR_pR_q \\ &\quad \times \exp\Big\{-\Big[\lambda_{11}R^2 + \lambda_{22}R_p^2 + \lambda_{33}R_q^2 \\ &\quad + 2\lambda_{12}RR_p\cos\big(\varphi-\varphi_p\big) \\ &\quad + 2\lambda_{13}RR_q\cos\big(\varphi-\varphi_q\big) \\ &\quad + 2\lambda_{23}R_pR_q\cos\big(\varphi_p-\varphi_q\big)\Big]\Big\}. \qquad (11) \end{aligned}$$

Lengthy calculations, not shown for brevity, lead to the following explicit expressions for the parameters λ_{ij},

$$(\det\mathbf{L}) = \frac{(e-1)(1-\sigma_A^2)\Sigma_N}{e\Sigma_q},$$

$$\lambda_{11} = \frac{(1-\sigma_{Apq}^2)}{e(\det\mathbf{L})} = \frac{1}{(e-1)},$$

$$\lambda_{22} = \frac{(e-\sigma_{Aq}^2)}{e(\det\mathbf{L})} = \frac{\Sigma_q}{\Sigma_N}\frac{1}{(1-\sigma_A^2)} + \frac{\Sigma_p}{\Sigma_N}\frac{1}{(e-1)},$$

$$\lambda_{33} = \frac{(e-\sigma_A^2)}{e(\det\mathbf{L})} = \frac{\Sigma_q}{\Sigma_N}\left[\frac{1}{(e-1)} + \frac{1}{(1-\sigma_A^2)}\right],$$

$$\lambda_{12} = \frac{(\sigma_{Aq}\sigma_{Apq} - \sigma_A)}{e(\det\mathbf{L})} = -\left(\frac{\Sigma_p}{\Sigma_N}\right)^{1/2}\frac{1}{(e-1)},$$

$$\lambda_{13} = \frac{(\sigma_A\sigma_{Apq} - \sigma_{Aq})}{e(\det\mathbf{L})} = -\left(\frac{\Sigma_q}{\Sigma_N}\right)^{1/2}\frac{1}{(e-1)},$$

$$\begin{aligned} \lambda_{23} = \frac{(\sigma_A\sigma_{Aq} - e\sigma_{Apq})}{e(\det\mathbf{L})} &= \frac{\big(\Sigma_p\Sigma_q\big)^{1/2}}{\Sigma_N}\left[\frac{e-\sigma_A^2}{(e-1)(1-\sigma_A^2)}\right] \\ &\quad - \left(\frac{\Sigma_q}{\Sigma_N}\right)^{1/2}\frac{\sigma_A}{(1-\sigma_A^2)}. \end{aligned}$$

It is worthwhile noting the following:

(a) When e and σ_A tend to unity, $(\det\mathbf{L})$ tends to zero and λ_{ij} tends to infinity. In this case the model tends to coincide with the target structure and $P(F, F_p, F_q)$ reduces to the Dirac delta function, in agreement with the expectations.

(b) The distribution (11) reduces to (2) when integrated over R_q and φ_q (we pass from three to two isomorphous structures). However, the coefficient of the term $2RR_p\cos(\varphi-\varphi_p)$ in (11) does not depend on σ_A, while the corresponding coefficient in (2) is σ_A dependent [*i.e.* it depends on $\sigma_A/(e-\sigma_A^2)$].

6. About conditional probabilities

The mathematical model described above suggests which conditional distributions are of interest for a phasing process. Since one may always suppose that a model (no matter whether rough or accurate) of the target structure is available, and that the diffraction intensities of the target structure were previously measured by a diffraction experiment, R, R_p and φ_p will always be known parameters in any conditional distribution of interest. In practice the following set of conditional distributions deserve to be studied:

(i) $P(\varphi_q, R_q|R, R_p, \varphi_p)$. Its derivation requires the previous integration of the distribution (11) over φ, which is supposed to be unknown. In practice, such conditional probability is the necessary intermediate step for calculating $P(\varphi_q|R, R_p, R_q, \varphi_p)$ and $P(R_q|R, R_p, \varphi_p, \varphi_q)$. In the absence of experimental errors, prior knowledge of R_q, R, R_p, φ_p geometrically defines φ_q; and *vice versa*, prior knowledge of $\varphi_q, R, R_p, \varphi_p$ defines R_q (see Fig. 4). In our mathematical modelling (which includes the experimental errors), R_q and φ_q will be two strongly correlated variables: therefore only one of $P(\varphi_q|R, R_p, R_q, \varphi_p)$ and $P(R_q|R, R_p, \varphi_p, \varphi_q)$ is of practical interest. Thanks to the results obtained by Caliandro *et al.* (2008), $P(R_q|R, R_p, \varphi_p, \varphi_q)$ may be neglected. Indeed the *DEDM* procedure provides simultaneous estimates of φ_q and R_q, and one of the two values may be used to estimate the other. To be more explicit, the first step of the *DEDM* algorithm requires the calculation, the modification and the inversion of the difference electron density. Such operations lead to new estimates of R_q and φ_q, which cannot be expected to satisfy the Carnot theorem for the triangle E, E_p, E_q: the algorithm accepts the φ_q estimate and derives R_q *via* the application of the Carnot theorem, in accordance with Fig. 4.

Finally we can limit our study to the distribution

$$P(\varphi_q|R, R_p, \varphi_p, \varphi_q).$$

(ii) $P(\varphi|R, R_p, R_q, \varphi_p, \varphi_q)$. This distribution is useful when estimates of φ_q and R_q become available during the phasing process. In practice, it constitutes our final tool for solving the phase problem when estimates of $R, R_p, R_q, \varphi_p, \varphi_q$ become available.

7. The conditional distributions $P(\varphi_q | R, R_p, R_q, \varphi_p)$

To derive conditional distributions of (11) and to reduce the complexity of the calculations, we first apply the approximation $\varphi_p \simeq \varphi$ (Giacovazzo & Siliqi, 2002): this relation is generally fulfilled when the model is close to the target structure and when R and R_p are sufficiently large. Using standard mathematical techniques we obtain, from (11), the following marginal and conditional distributions,

$$\begin{aligned} P(R, R_p, R_q, \varphi_p, \varphi_q) \cong{}& 2\pi^{-2}(\det \mathbf{L})^{-1} R R_p R_q \\ &\times \exp\Big\{-\Big[\lambda_{11}R^2 + \lambda_{22}R_p^2 + \lambda_{33}R_q^2 \\ &+2R_q(\lambda_{13}R + \lambda_{23}R_p)\cos(\varphi_q - \varphi_p)\Big]\Big\} \end{aligned}$$

from which

$$P(\varphi_q|R, R_p, R_q, \varphi_p) \cong [2\pi I_0(G_q)]^{-1} \exp\left[G_q \cos(\varphi_q - \varphi_p)\right], \qquad (12)$$

where

$$G_q = -2R_q(\lambda_{13}R + \lambda_{23}R_p).$$

In accordance with §5 we obtain

$$G_q = \frac{2R'_q}{e-1}\left[(R - R'_p) - (1-D)\left(\frac{e-1}{1-\sigma_A^2}\right)R'_p\right]. \qquad (13a)$$

In terms of F the above relation may be rewritten as

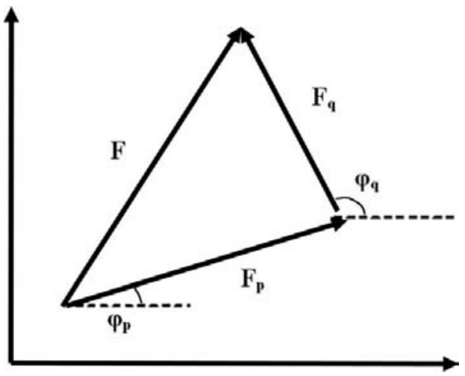

Figure 4
The triangle F, F_p, F_q with angles φ_p and φ_q emphasized.

$$G_q = \frac{2}{e-1}\frac{|F_q|}{\Sigma_N}\left[(|F| - |F_p|) - (1-D)\left(\frac{e-1}{1-\sigma_A^2}\right)|F_p|\right]. \qquad (13b)$$

Notice that the factor $(1 - D)/(1 - \sigma_A^2)$ is always positive, and is smaller than unity when $\Sigma_p < \Sigma_N$.

The condition $\varphi_p \simeq \varphi$ used to derive (13) from (11) may be replaced by a better approximation (Read, 1986): $|F|\exp(i\varphi) = m|F|\exp(i\varphi_p)$. Then (13) may be rewritten as follows:

$$G_q = \frac{2R'_q}{e-1}\left[(mR - \sigma_A R_p) - R'_p(1-D)\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right)\right] \qquad (14a)$$

$$G_q = \frac{2F_q}{(e-1)\sum_N}\left[(m|F| - D|F_p|) - |F_p|(1-D)\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right)\right] \qquad (14b)$$

Equations (13*b*) and (14*b*) suggest the following considerations [analogous considerations hold for equations (13*a*) and (14*a*)]:

(1) The well known phase relationship (classically used in difference Fourier syntheses)

$$\begin{aligned} &\varphi_q \simeq \varphi_p \qquad \text{if } m|F| > D|F_p|, \\ &\varphi_q \simeq \varphi_p + \pi \qquad \text{if } m|F| < D|F_p|, \end{aligned}$$

is not generally supported. Indeed φ_q is expected to be close to φ_p or close to $\varphi_p + \pi$ according to whether G_q is positive or negative.

(2) The sign of G_q does not always coincide with the sign of $(m|F| - D|F_p|)$. Indeed the right-hand side of (13*b*) is the sum of two contributions, the first depending on the value of $(m|F| - D|F_p|)$, called here *difference term*; the second {say $|F_p|(1 - D)[(e - \sigma_A^2)/(1 - \sigma_A^2)]$}, called the *flipping term*, is always negative and proportional (*via* a positive factor) to $-|F_p|$. Its contribution depends on the quality of the model structure, and increases with the poorness of the model. It is dominant when the model is very poor (then $m|F| - D|F_p| = 0$).

(3) The *flipping term* concurs to establish the anticorrelation between E_p and E_q foreseen by (7). It is not negligible when D is small: then the $\cos(\varphi_q - \varphi_p)$ values are expected to be more negative than the estimates provided by the *difference term* only. Accordingly, a statistical asymmetry in the phase esti-

mates is expected when the reflections are ordered according to $(m|F| - D|F_p|)$. In particular, according to equation (14b), the most probable value of φ_q and its reliability, for two reflections having the same value of $\left|m|F| - D|F_p|\right|$, is different according to whether $(m|F| - D|F_p|)$ is positive or negative. We will check such expectation in the applications.

Let us now omit the condition $\varphi \simeq \varphi_p$ in our calculations (too restrictive when the model is a poor approximation of the target structure) and look for a different mathematical approach. Since φ is now a free variable, equation (11) may be integrated over φ, according to

$$\begin{aligned} P(R, R_p, R_q, \varphi_p, \varphi_q) &\cong \pi^{-3}(\det \mathbf{L})^{-1}RR_pR_q \\ &\quad\times \exp\Big\{-\Big[\lambda_{11}R^2 + \lambda_{22}R_p^2 + \lambda_{33}R_q^2 \\ &\quad + 2\lambda_{23}R_pR_q\cos(\varphi_p - \varphi_q)\Big]\Big\} \\ &\quad\times \int_0^{2\pi} \exp\Big[2\lambda_{12}RR_p\cos(\varphi - \varphi_p) \\ &\quad + 2\lambda_{13}RR_q\cos(\varphi - \varphi_q)\Big]\,\mathrm{d}\varphi \\ &= 2\pi^{-2}(\det \mathbf{L})^{-1}RR_pR_q \\ &\quad\times \exp\Big\{-\Big[\lambda_{11}R^2 + \lambda_{22}R_p^2 + \lambda_{33}R_q^2 \\ &\quad + 2\lambda_{23}R_pR_q\cos(\varphi_p - \varphi_q)\Big]\Big\}I_0(G_{qp}), \end{aligned}$$

where

$$G_{qp} = 2R\big[\lambda_{12}^2R_p^2 + \lambda_{13}^2R_q^2 + 2\lambda_{12}\lambda_{13}R_pR_q\cos(\varphi_p - \varphi_q)\big]^{1/2}.$$

To perform the integral the following approximation may be used (Giacovazzo, 1979),

$$\begin{aligned} I_o\big[G_1^2 + G_2^2 + 2G_1G_2\cos(\varphi - \theta)\big]^{1/2} &= \frac{I_0(G_1)I_0(G_2)}{I_0(G)} \\ &\quad\times \exp[G\cos(\varphi - \theta)] \end{aligned}$$

where G satisfies the equation

$$D_1(G) = D_1(G_1)D_1(G_2)$$

and $D_1(x) = I_1(x)/I_0(x)$ is the ratio of two modified Bessel functions of order 1 and 0, respectively.

Accordingly,

$$\begin{aligned} P(R, R_p, R_q, \varphi_p, \varphi_q) &= 2\pi^{-2}\,\frac{I_0(2\lambda_{12}RR_p)I_0(2\lambda_{13}RR_q)}{I_0(G_{qp})} \\ &\quad\times (\det \mathbf{L})^{-1}RR_pR_q \\ &\quad\times \exp\Big\{-\Big[\lambda_{11}R^2 + \lambda_{22}R_p^2 + \lambda_{33}R_q^2 \\ &\quad + (2\lambda_{23}R_pR_q - G_{qp})\cos(\varphi_p - \varphi_q)\Big]\Big\} \end{aligned}$$

and

$$P(\varphi_q|R, R_p, R_q, \varphi_p) = [2\pi I_0(S_q)]^{-1}\exp\big[S_q\cos(\varphi_p - \varphi_q)\big] \quad (15)$$

where

$$S_q = \big(G_{qp} - 2\lambda_{23}R_pR_q\big), \qquad (16a)$$

$$D_1(G_{qp}) = D_1\big(2\lambda_{12}RR_p\big)D_1\big(2\lambda_{13}RR_q\big). \qquad (16b)$$

In accordance with the λ_{ij} expressions derived in §5 we can rewrite equations (16) as

$$S_q = \frac{2R'_q}{e-1}\left\{\left[\frac{G_{qp}}{2R'_q}(e-1) - R'_p\right] - \frac{(1-D)(e-1)}{1-\sigma_A^2}R'_p\right\}, \qquad (17a)$$

$$D_1(G_{qp}) = D_1\left(-\frac{2}{e-1}RR'_p\right)D_1\left(-\frac{2}{e-1}RR'_q\right). \qquad (17b)$$

We note the following:

(*a*) G_{qp} is always positive. One can recognize a *difference term* {say $[(G_{qp}/2R'_q)(e-1) - R'_p]$} and a *flipping term* {say $[(1-D)(e-1)/(1-\sigma_A^2)]R'_p$}.

(*b*) The reliability parameter S_q may be positive or negative. If $S_q > 0$, then $\varphi_q \simeq \varphi_p$, if $S_q < 0$ then $\varphi_q \simeq \varphi_p + \pi$.

(*c*) If $2RR'_p$ is sufficiently large (then the relation $\varphi_q \simeq \varphi_p$ should hold), we have $D_1\{-[2/(e-1)]RR'_p\} \simeq -1$, $G_{qp} \simeq [2/(e-1)]RR'_q$, and S_q coincides with G_q, as expected.

(*d*) If both R and R'_q are large and R'_p is small (this is an important subset of reflections, owing to the large experimental amplitudes), then

$$D_1\left(-\frac{2}{e-1}RR'_q\right) \simeq -1, \qquad G_{qp} \simeq \frac{2}{e-1}RR'_p,$$

and

$$\begin{aligned} S_q &\simeq \frac{2R'_q}{e-1}\left[\big(R - R'_q\big)\frac{R'_p}{R'_q} - \frac{(1-D)(e-1)}{1-\sigma_A^2}R'_p\right] \\ &\simeq \frac{2}{e-1}\big(R - R'_q\big)R'_p - 2R'_qR'_p\frac{(1-D)}{1-\sigma_A^2}, \end{aligned}$$

which is in strong disagreement with (13*a*).

(*e*) If the R, R'_p and R'_q values are sufficiently small to allow the approximation $D_1(x) \simeq x/2$, then

$$\begin{aligned} D_1\left(-\frac{2}{e-1}RR'_p\right) &\simeq -\frac{1}{e-1}RR'_p, \\ D_1\left(-\frac{2}{e-1}RR'_q\right) &\simeq -\frac{1}{e-1}RR'_q, \end{aligned}$$

and

$$S_q \simeq \frac{2R'_q}{e-1}\left[\left(\frac{R^2}{e-1} - 1\right)R'_p - (e-1)\left(\frac{1-D}{1-\sigma_A^2}\right)R'_p\right],$$

which again diverges from (13*a*). Indeed, the difference term is now proportional to R'_p and its sign depends on whether $R^2/\langle|\mu|^2\rangle$ is larger or smaller than 1. A special case occurs when R'_p and R'_q are large while R is very small: then both the *difference* and the *flipping term* are negative. This is the case in which S_q attains the strongest negative values, and the relation $\varphi_q \simeq \varphi_p + \pi$ is suggested, in accordance with (13*a*).

(*f*) At difference with (13) and (14), the value of $|R_q|$ contributes to fix both the value of φ_q and its reliability. As an example, let us consider the point (*d*) above, where both R and R'_q are large and R'_p is small. Then the sign of S_q is strongly influenced by the value of $(R - R'_q)R'_p$: this term is positive or negative according to whether $R > R'_q$ or $R < R'_q$.

The above considerations show that breaking down the approximation $|F|\exp(i\varphi) = m|F|\exp(i\varphi_p)$ provides phase estimates (*i.e.* through the S_q parameter) which may substantially differ from the G_q estimates. Both of them, however, provide the most accurate estimates when $|F_p|$ is much larger than $|F|$. This feature allows us to introduce the latter observation: the equations derived in this section confirm a result obtained by Caliandro *et al.* (2008) based on the algebraic calculation of the variance σ_q^2 of F_q, *i.e.*

$$\sigma_q^2 = \left(1 - m^2\right)|F|^2 + \left\langle|\mu|^2\right\rangle. \tag{18}$$

According to the relationship (18), the accuracy of the φ_q estimates are expected to be inversely correlated with σ_q: in particular, reflections with the same value of $(m|F| - D|F_p|)$ are expected to have different values of σ_q according to whether $|F|$ is large and $\left|F_p\right|$ is small, or *vice versa*. The phase estimates corresponding to strong negative values of $(m|F| - D|F_p|)$ are indicated by (18) as the most accurate ones, in agreement with the present theory.

Equation (18) was the basis of the *DEDM* procedure: it was able to assign variances to the phase estimates, but not to modify the phase estimates indicated by the sign of $(m|F| - D|F_p|)$. The present theory provides, through equations (12)–(17), a more solid basis to the results by Caliandro *et al.* (2008): in addition, it modifies the φ_q estimates, which no longer exclusively depend on the sign of $(m|F| - D|F_p|)$.

8. Considerations on the conditional distribution $P(\varphi_q\,|\,R, R_p, R_q, \varphi_p)$

The theory so far described provides estimates of φ_q no matter the quality of the model structure. For example, even in the limit case in which σ_A and D are zero (model and target structure completely uncorrelated, as in Fig. 3), the parameter G_q may be large and therefore the φ_q estimate may be reliable. That is equivalent to the following assumption: it is possible to obtain a meaningful estimate of ρ_q even when ρ_p and ρ are completely uncorrelated (*e.g.* when ρ_p is randomly fixed). At first sight this property seems to be without fundamentals: indeed under these conditions the classical estimate $\langle|F_q|\rangle = |m|F| - D|F_p||$ is vanishing, and consequently the intensity of any pixel of the difference Fourier synthesis calculated *via* those coefficients is expected to vanish. In practice, no information in, no information out.

However, the correctness of our approach becomes clear if one considers Fig. 3, where simplified models for ρ, ρ_p and ρ_q are schematized when $\sigma_A = 0$. On assuming that ρ is unknown and that ρ_p is uncorrelated with ρ (they have no peak in common), then ρ_q is constituted by N positive and by p negative peaks. It is easily seen that ρ_q and $-\rho_p$ are positively correlated because they have p negative peaks in common (corresponding to the wrongly located atoms in the ρ_p structure). Equivalently, ρ_p and ρ_q are anticorrelated, and it is just the *flipping term* in equations (13), (14) and (17) which guarantees the anticorrelation. Accordingly an estimate of ρ_q is always possible, even when the atoms of the model structure are randomly located.

Similar observations hold also for Fig. 1, where σ_A and D do not coincide either with 0 or with 1. In this case ρ_q is constituted by $N - p$ positive peaks and by p electron density residuals, constituted by pairs of positive and negative peaks. Now ρ_q and ρ_p are weakly anticorrelated: such a relation may be fully taken into account by using both the *difference* and the *flipping term*.

If we consider Fig. 2, where the case $D = 1$ is schematized, it is seen that ρ_q and ρ_p are uncorrelated. Accordingly the *flipping term* vanishes and the structure completion may occur only through the *difference term*.

If so, it is important to recognize the primary source from which the *flipping term* arises. Let us return to equations (5): the definition $F_q = F - F_p$ (in vectorial sense) is equivalent to the assumption that ρ_q is the *ideal difference Fourier synthesis* $\rho - \rho_p$ (non-positive definite). It is just this mathematical model which allows the estimate of ρ_q no matter the quality of ρ_p. On the contrary, the mathematical model described by (3) is unable to generate the *flipping term* in equations (13), (14) and (17).

A last consideration deserves to be made. *SIR–MIR* techniques have a special advantage with respect to the case in which experimental diffraction data from isomorphous structures are not available: *e.g.* at least two diffraction moduli (*i.e.* of the protein and of the derivative) are experimentally measured. When only the diffraction data of the target structure are experimentally available, as in the case treated so far, the simultaneous knowledge of $|F|$, $|F_p|$, φ_p is unable to provide exact estimates of $|F_q|$: indeed, the coefficients of the difference Fourier synthesis are usually rough approximations of $|F_q|$ when the model is a poor approximation of the target structure. Luckily, the *DEDM* procedure (Caliandro *et al.*, 2008), based on the cyclic modification of the difference Fourier synthesis, is progressively able to improve such an approximation and makes the theory presented here fully applicable.

It may be worthwhile noting that we were unable to derive a flipping term from the distribution (2). The reason is trivial: if one tries to derive the probability density $P(\varphi_q, R_q|R, R_p, \varphi_p)$ as a derivative of the distribution (2) by imposing the condition $E_q = E - E_p$, then the Dirac delta function $\delta[F_q - (F - F_p)]$ should obviously be obtained, which is not useful for the phasing problem. On the contrary, defining a new variate E_q through conditions which allow a correlation with E and E_p but not its identity with $E - E_p$ allows a larger flexibility and unexpected results.

9. About the $(F_o - F_c)$ Fourier synthesis

The properties of the traditional Fourier synthesis $(F_o - F_c)$ were studied, among others, by Cochran (1951) and by Henderson & Moffat (1971). Main (1979) proposed to use the coefficient $(m|F| - |F_p|)$ to take into account the uncertainty in the phases of the target structure. Ursby & Bourgeois

(1997) studied, *via* the Bayesian statistics, the influence of measurement errors on the efficiency of the synthesis. The most popular coefficient for calculating a difference Fourier synthesis, say $(m|F| - D|F_p|)$, has been suggested by Read (1986); the D term was introduced to compensate for errors in the atomic positions, scattering and B factors.

Let us derive, from the theoretical results obtained in §7, suitable coefficients for the difference Fourier synthesis. When the assumption $|F|\exp(i\varphi) = m|F|\exp(i\varphi_p)$ is made, φ, φ_p and φ_q are necessarily collinear: this is the classical situation met when difference maps are calculated. Then φ_q is assumed to be φ_p or $\varphi_p + \pi$ according to whether $(mR - \sigma_A R_p)$ is positive or negative. Equation (13) suggests, under the same collinearity conditions, a different criterion: φ_q is expected to be close to φ_p or to $\varphi_p + \pi$ according to whether $(mR - \sigma_A R_p) - R'_p(1-D)[(e-\sigma_A^2)/(1-\sigma_A^2)]$ is positive or negative. Accordingly, the following difference Fourier coefficient may be conjectured,

$$\left[(mR - \sigma_A R_p) - R'_p(1-D)\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right)\right]\exp(i\varphi_p). \qquad (19)$$

Let us now examine the properties of a difference Fourier synthesis when the coefficients (19) are used. Three different types of peaks are expected, the properties of which are determined by the poorness of the model. Let us first consider a very poor model. Then the difference electron density will show the following peaks:

(*a*) Very strong negative peaks where model atoms do not overlap with target atoms: in this case both the *difference* and the *flipping* term will generate negative electron density.

(*b*) Medium-intensity negative peaks, where model and target atoms overlap: in this case the *difference term* does not provide any contribution to the electron density while the *flipping term* will generate negative electron density.

(*c*) Medium intensity positive peaks, where target atoms do not overlap with model atoms: in this case the *difference term* provides a positive electron density while the *flipping term* does not provide any contribution to the electron density.

The intensity ratio between the peaks (*a*) and the peaks (*b*)–(*c*) will decrease when the model becomes a better approximation of the target structure. In particular (i) the intensities of the peaks (*a*) will become weaker because the *flipping term* contribution diminishes; (ii) the peaks (*b*) tend to vanish or to become very weak; (iii) the peaks (*c*) will continue to show their intensity. In other words, the difference Fourier synthesis will then show the classical peaks generated by the *difference term*.

Let us now consider equations (17) in order to derive useful coefficients for a difference electron density. The assumption $\varphi \simeq \varphi_p$ is no longer made and therefore the difference Fourier synthesis may be calculated under more general conditions: in particular φ is not compelled to be collinear with φ_p, so that φ_q may assume any value between 0 and 2π.

Let us rewrite equation (17*a*) in the form

$$S_q = \frac{2R'_q}{e-1}\left\{\frac{G_{qp}}{2R'_q}(e-1) - R'_p\left[\frac{e-\sigma_A^2 - D(e-1)}{1-\sigma_A^2}\right]\right\}. \qquad (20)$$

This form is mathematically consistent (*i.e.* S_q does not present any discontinuity) even when R'_q goes to zero. Indeed, when R'_q is sufficiently small, $D_1\{-[2/(e-1)]RR'_q\}$ tends to $\{-[1/(e-1)]RR'_q\}$ and $D_1(G_{qp}) \simeq -[1/(e-1)]RR'_q \times D_1\{-[2/(e-1)]RR'_p\}$.

Then, in accordance with the criteria used to define the coefficient (19), the following difference Fourier coefficient arises from (20),

$$\left\{\frac{G_{qp}}{2R'_q}(e-1) - R'_p\left[\frac{e-\sigma_A^2 - D(e-1)}{1-\sigma_A^2}\right]\right\}\exp(i\varphi_p). \qquad (21)$$

There is a remarkable difference between the coefficients (19) and (21). In (19) R'_q does not influence the φ_q value but only its reliability (that is confirmed by trivial geometrical considerations: if φ, φ_p and φ_q are collinear, the difference $R - R_p$ is sufficient to define φ_q). *Vice versa*, in (21) R'_q concurs to define both φ_q and the reliability of the estimate: this is an advantage when R'_q is experimentally available, but it may degrade the φ_q estimate if R'_q is roughly evaluated.

Let us now consider the influence of the completeness of the model on the quality of the estimated difference Fourier map. Rewrite the algebraic expression of the *flipping term* in equation (13*a*) in the more explicit form

$$-R_p\left[\left(\frac{\Sigma_p}{\Sigma_N}\right)^{1/2} - \sigma_A\right]\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right) = -R_p\left[\left(\frac{\Sigma_p}{\Sigma_N}\right)^{1/2} - \sigma_A\right] \times\left(1+\frac{\sigma_R^2}{1-\sigma_A^2}\right),$$

and consider (see Fig. 5) the trend of the factor

$$FF = \left[\left(\frac{\Sigma_p}{\Sigma_N}\right)^{1/2} - \sigma_A\right]\left(1+\frac{\sigma_R^2}{1-\sigma_A^2}\right)$$

as a function of σ_A, for different values of Σ_p/Σ_N and when σ_R^2 = 1.0. The curves are essentially straight lines, which are almost symmetrical with respect to the diagonal of the figure:

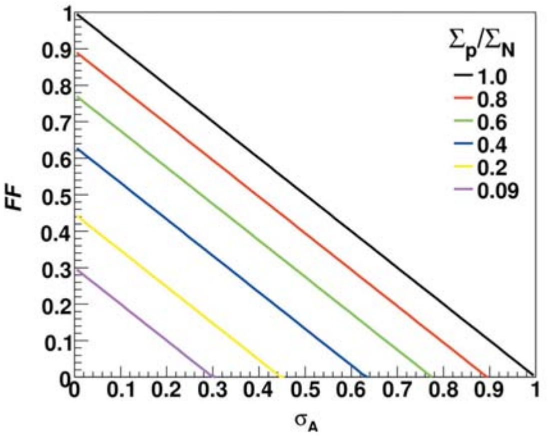

Figure 5
The factor *FF versus* σ_A for different values of Σ_p/Σ_N.

they intersect the FF and σ_A axes at exactly/about $(\Sigma_p/\Sigma_N)^{1/2}$, and translate towards higher values of FF when σ_R^2 increases (they are close to straight lines as long as e remains close to unity). Accordingly, for small values of σ_A, FF is close to $(\Sigma_p/\Sigma_N)^{1/2}$, and decreases up to zero when σ_A coincides with $(\Sigma_p/\Sigma_N)^{1/2}$. The flipping term, as part of a Fourier coefficient, is maximally dominant with respect to the difference term when $\Sigma_p \simeq \Sigma_N$: then it attains $-R_p[(1-\sigma_A)+(\sigma_R^2)/(1+\sigma_A)]$, and almost coincides with $-R_p$ when $\sigma_A = 0$ (just when the difference term vanishes).

The above result suggests that the *flipping term* may improve the ρ_q estimate more when the model is complete rather than when it is incomplete, no matter its poorness. In practice the correlation between the estimated and the true ρ_q is expected to be higher if an incomplete but good model is completed by addition of a complementary poor model. This feature will be verified in §11.

Let us now derive a simplified expression of equation (19) when the model is complete: then $(\Sigma_p/\Sigma_N)^{1/2} = 1$, $D = \sigma_A$ and (19) reduces to

$$\left[\left(mR - R_p\right)\frac{e+\sigma_A}{1+\sigma_A}\right]\exp(i\varphi_p)$$
$$= \left[\left(mR - R_p\right) - \frac{\sigma_R^2}{1+\sigma_A}R_p\right]\exp(i\varphi_p).$$

Under the above assumption, the term (19) is the sum of the Fourier coefficient proposed by Main (1979) plus a negative term depending on the measurement errors, on the quality of the model and on R_p (the reader should also consider that, according to this theory, σ_R^2 cannot vanish).

The results obtained in this section may be used to define the coefficients for a $(2F_{\rm o} - F_{\rm c})$ Fourier synthesis. The suggested coefficients are

$$\left(2mR - \sigma_A R_p\right) - R'_p(1-D)\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right)$$

or

$$mR + \frac{G_{qp}}{2R'_q}(e-1) - R'_p\left[\frac{e-\sigma_A^2 - D(e-1)}{1-\sigma_A^2}\right]$$

according to whether equation (19) or (21) is used as the coefficient of the difference synthesis.

10. The conditional probability $P(\varphi\,|\,R, R_p, R_q, \varphi_p, \varphi_q)$

Let us suppose that we know the moduli R, R_p, R_q and the phases φ_p, φ_q. This assumption includes the case in which, like in the *EDM–DEDM* procedures, E_p and E_q are not collinear. Then, from the joint probability distribution (11), the following conditional distribution may be obtained,

$$P(\varphi|R, R_p, R_q, \varphi_p, \varphi_q) = [2\pi I_0(Q)]^{-1}\exp[Q\cos(\varphi-\theta)],$$

where θ is the most probable value of φ, given by

$$\tan\theta = \frac{-\lambda_{12}R_p\sin\varphi_p - \lambda_{13}R_q\sin\varphi_q}{-\lambda_{12}R_p\cos\varphi_p - \lambda_{13}R_q\cos\varphi_q}$$
$$= \frac{R'_p\sin\varphi_p + R'_q\sin\varphi_q}{R'_p\cos\varphi_p + R'_q\cos\varphi_q} = \frac{Q_{\rm T}}{Q_{\rm B}}, \qquad (22)$$

and

$$Q = 2R(e-1)^{-1}(Q_{\rm T}^2 + Q_{\rm B}^2)^{1/2}$$
$$= 2R(e-1)^{-1}\left[R_p'^2 + R_q'^2 + 2R'_pR'_q\cos(\varphi_p - \varphi_q)\right]^{1/2} \qquad (23)$$

is its reliability factor.

Equations (22) and (23) provide the best estimate of φ given R, R_p, R_q, φ_p, φ_q *via* the sum of two contributions, the first arising from ρ_p and the second from the difference electron density map. We note that (22) is unweighted. The reason has to be searched for in the definitions (5): given E_p and E_q, the value of E may be fixed without any weighting scheme owing to the fact that ρ is just the sum of ρ_p and ρ_q. However, in practice, while E_p is fixed by the structure model without any uncertainty, only estimates of E_q are available. Accordingly, weights may be involved in the tangent formula (22) so that

$$\tan\theta = \frac{w_pR'_p\sin\varphi_p + w_qR'_q\sin\varphi_q}{w_pR'_p\cos\varphi_p + w_qR'_q\cos\varphi_q} = \frac{Q_{\rm T}}{Q_{\rm B}} \qquad (24)$$

and

$$Q = 2R(e-1)^{-1}(Q_{\rm T}^2 + Q_{\rm B}^2)^{1/2}$$
$$= 2R(e-1)^{-1}\Big[w_p^2R_p'^2 + w_q^2R_q'^2$$
$$+ 2w_pw_qR'_pR'_q\cos(\varphi_p - \varphi_q)\Big]^{1/2}. \qquad (25)$$

11. The applications

The theory described above can potentially influence a wide crystallographic area: *ab initio* crystal structure solution of small as well as of macromolecules, phase assignment and refinement in cooperation with *SIR–MIR*, molecular replacement, and *EDM–DEDM* techniques. The following preliminary tests aim at checking the correctness of the theory and its usefulness in the most basic case, *e.g.* the calculation and the study of the properties of the difference Fourier synthesis. We will limit ourselves to checking the properties of the coefficients (19): comparison will be made with the Fourier syntheses calculated *via* the coefficients (1), to assess whether the theory described here leads to more informative maps. The study of the properties of the coefficients (21) and the use of equations (22)–(25) are deferred to a future paper: the latter require the availability of specific computing codes for running practical phasing procedures, particularly in the protein field. For the tests presented in this paper we selected 18 cases, listed in Table 1. They are proteins to which the program *REMO09* (Caliandro *et al.*, 2009*c*), included into the package *IL MILIONE* (Burla *et al.*, 2007), was applied to find structural models *via* molecular replacement. The columns 'Target' and 'Model' indicate the Protein Data Bank (PDB) code of the

Table 1
Test structures used in the analysis.

For each test structure, PDB is the PDB code of the protein structure, Res is the data resolution limit in Å, NresT is the number of residues of the target structure, Model is the PDB code of the model structure used in the molecular replacement procedure, NresM is the number of residues of the model structure and CORR is the correlation factor between the electron density map calculated by using observed normalized moduli and phases φ_p and the map calculated *via* observed normalized moduli and phases φ calculated from deposited coordinates.

PDB	Res	NresT	Model	NresM	CORR
1kf3	1.0	124	7rsa	124	0.93
6rhn	2.2	115	4rhn	104	0.89
1zs0	1.6	163	1i76	163	0.82
1na7	2.4	329	1m2r	327	0.74
1s31	2.7	273	1c8z	265	0.73
1a6m	1.0	151	1mbc	153	0.72
2p0g	2.3	318	2oka	336	0.72
2sar	1.8	192	1ucl, chain A	96	0.70
1kqw	1.8	134	1opa	133	0.62
1lys	1.7	258	2ihl	129	0.57
6ebx	1.7	124	3ebx	62	0.43
1cgn	2.2	127	2ccy	122	0.39
2iff	2.6	556	2hem	129	0.36
1yxa	2.1	740	1qlp	744	0.34
2bpy	2.1	1155	1mki	1248	0.32
9pti	1.2	58	3ebx	62	0.02
6ebx′	1.7	124	3ebx, 2 copies	124	0.01
9pti′	1.2	58	1lri	98	0.01

Table 2
Correlation coefficients and average phase errors of difference Fourier syntheses, with respect to the ideal difference Fourier synthesis.

CORRq_1 and $\langle\Delta\varphi_q\rangle_1$, CORR$q_{19}$ and $\langle\Delta\varphi_q\rangle_{19}$, and CORR$q_{26}$ and $\langle\Delta\varphi_q\rangle_{26}$ refer to difference Fourier syntheses calculated *via* coefficients (1), (19) and (26), respectively.

PDB	CORRq_1	$\langle\Delta\varphi_q\rangle_1$	CORRq_{19}	$\langle\Delta\varphi_q\rangle_{19}$	CORRq_{26}	$\langle\Delta\varphi_q\rangle_{26}$
1kf3	0.51	57	0.62	53	0.09	76
6rhn	0.40	63	0.62	54	0.34	75
1zs0	0.38	64	0.61	54	0.41	70
1na7	0.31	63	0.58	59	0.46	66
1s31	0.34	63	0.60	55	0.50	64
1a6m	0.30	64	0.60	56	0.42	66
2p0g	0.30	66	0.60	59	0.48	68
2sar	0.36	62	0.59	55	0.46	66
1kqw	0.28	64	0.61	54	0.55	60
1lys	0.31	65	0.51	62	0.39	72
6ebx	0.26	64	0.53	61	0.51	54
1cgn	0.19	65	0.62	53	0.61	55
2iff	0.18	71	0.51	67	0.49	67
1yxa	0.18	64	0.64	51	0.64	51
2bpy	0.12	68	0.64	52	0.63	52
9pti	0.01	73	0.62	54	0.64	54
6ebx′	0.00	70	0.66	46	0.69	46
9pti′	−0.02	74	0.66	50	0.68	50

target and the model structure, respectively; the columns ‘NresT’ and ‘NresM’ show the corresponding number of residues, ‘RES’ is the data resolution and ‘CORR’ is the correlation factors between the electron density map calculated by using observed moduli and phases φ_p (the phase values available at the end of the molecular replacement process) and the map calculated *via* observed moduli and phases φ calculated from deposited coordinates. We used normalized moduli to calculate CORR, in order to allow a direct comparison among the E-type syntheses used for the difference Fourier map. In Tables 1 and 2 the test structures are listed according to CORR: its values span the interval (−0.03 to 0.95). The highest values of CORR correspond to the best models.

To compare the efficiency of various difference Fourier coefficients with the ideal ones we will use in the tables and in the figures the following notation,

$$\langle E'_q\rangle_1 = (mR - \sigma_A R_p), \qquad \langle\varphi_q\rangle_1 = \exp[i(\varphi_p + s_1\pi)],$$

$$\langle E'_q\rangle_{19} = \left[(mR - \sigma_A R_p) - R'_p(1-D)\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right)\right],$$

$$\langle\varphi_q\rangle_{19} = \exp[i(\varphi_p + s_{19}\pi)],$$

where $E'_q = F_q/\Sigma_N^{1/2}$ and s_1 and s_{19} are 0 or 1 according to whether $\langle E'_q\rangle_1$ and $\langle E'_q\rangle_{19}$ are positive or negative, respectively. This notation allows us to discuss the role of the signs s_1 and s_{19} in the phase error estimation. The parameter e has been calculated in the tests by using the definition given in the notations and the measurement errors reported in the reflection files taken from the PDB. We define now in detail the ideal difference Fourier synthesis which will be used as a term of comparison. $F_q = F - F_p$ are the natural coefficients of the ideal F-type difference Fourier synthesis: in terms of normalized or pseudo-normalized structure factors these coefficients may be written as $E_q\Sigma_q^{1/2} = E\Sigma_N^{1/2} - E_p\Sigma_p^{1/2}$. Therefore we will consider as an ideal E-type difference Fourier synthesis that calculated *via* coefficients $E_q(\Sigma_q^{1/2}/\Sigma_N^{1/2}) = E'_q = E - E'_p$. Then,

$\langle\Delta\varphi_q\rangle_1 = \langle|\langle\varphi_q\rangle_1 - \varphi_q|\rangle$ is the average phase error of the difference Fourier synthesis, calculated *via* coefficients (1), with respect to the *ideal difference Fourier synthesis*. CORRq_1 is the correlation of the corresponding maps.

$\langle\Delta\varphi_q\rangle_{19} = \langle|\langle\varphi_q\rangle_{19} - \varphi_q|\rangle$ is the average phase error of the difference Fourier synthesis, calculated *via* coefficients (19), with respect to *the ideal difference Fourier synthesis*. CORRq_{19} is the correlation of the corresponding maps.

$\langle\Delta\varphi_q\rangle_{26} = \langle|\langle\varphi_q\rangle_{26} - \varphi_q|\rangle$ is the average phase error, with respect to the *ideal difference Fourier synthesis*, of the difference synthesis calculated *via* coefficients

$$-R'_p \exp(i\varphi_p). \tag{26}$$

CORRq_{26} is the correlation of the corresponding maps. Accordingly $\langle E'_q\rangle_{26} = R_p$, $\langle\varphi_q\rangle_{26} = \exp[i(\varphi_p + \pi)]$. This is a limit case for the coefficients (19), occurring when $\sigma_A = 0$. The corresponding results may be useful for better understanding the main features of the coefficients (19).

In Table 2 the average values of CORRq_1, $\langle\Delta\varphi_q\rangle_1$, CORR$q_{19}$, $\langle\Delta\varphi_q\rangle_{19}$, CORR$q_{26}$ and $\langle\Delta\varphi_q\rangle_{26}$ are reported for all the test cases considered. To better appreciate the general trends, the correlation values are plotted in Fig. 6, together with the values of CORR. We observe the following:

(1) CORRq_1 tends to decrease with decreasing values of CORR. It is remarkably smaller than CORR for all the test structures except for the worst quality models (9pti, 6ebx′, 9pti′), for which CORRq_1 and CORR have about the same

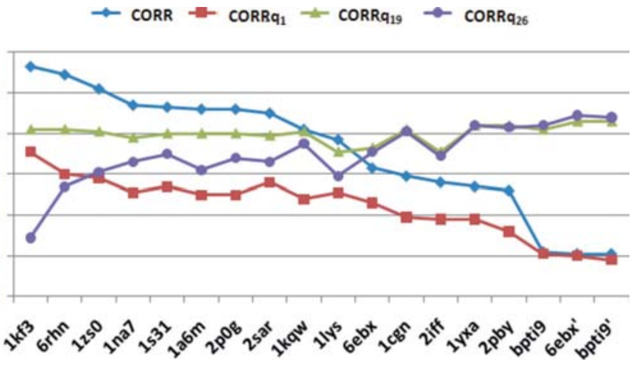

Figure 6
Correlation factors for all the test cases.

values. CORRq_1 varies between −0.02 (for the worst structure model) and 0.51 (for the best model): correspondently $\langle\Delta\varphi_q\rangle_1$ varies between 74° and 57°. The result confirms the common belief that the classic difference Fourier synthesis provides useful information only when the model structure is sufficiently accurate.

(2) CORRq_{19} is almost constant with CORR and is rather insensitive to the quality of the model: it lies in the interval 0.51–0.66, and $\langle\Delta\varphi_q\rangle_{19}$ lies in the narrow interval 46°–67° in spite of the quite large range covered by CORR (*e.g.* from 0.01 to 0.93). CORRq_{19} is smaller than CORR for high-quality models, and is larger when the model is poor: in this latter case the quality of the difference structure model is higher than the quality of the structure model. Therefore CORRq_1 always provides useful structural information, even in the case in which the model is very poor.

(3) In all cases CORRq_{19} > CORRq_1 and $\langle\Delta\varphi_q\rangle_{19} < \langle\Delta\varphi_q\rangle_1$: the differences are quite remarkable for very poor models. For example, in the test cases 9pti, 6ebx′ and 9pti′, very poor models were used, whose electron density maps are uncorrelated with the maps of the target (*e.g.* CORR $\simeq$ 0 in all three cases). While CORRq_1 is constantly close to 0, CORRq_{19} lies in the range 0.62–0.66 and $\langle\Delta\varphi_q\rangle_{19}$ lies in the range 46°–54°.

(4) CORRq_{26} is comparable with CORRq_{19} for poor models [in these cases $\sigma_A \simeq 0$ and coefficients (19) reduce to (26)], but is quite small when the quality of the model is sufficiently high.

The results of points (1)–(4) are in perfect agreement with the theory expectations and with Figs. 1–3.

The test cases denoted by 9pti, 6ebx′ and 9pti′ deserve further comment. They deliberately correspond to false MR solutions for (*a*) both copies of a homologous model for 6ebx′, (*b*) a homologous model for 9pti, (*c*) a model completely different from the target in size and folding for 9pti′. As a result, all of them show vanishing values of CORR and CORRq_1, but at the same time they exhibit high values of CORRq_{19} and CORRq_{26}. This may be explained by considering Fig. 3: for very poor models $D \simeq 0$ and ρ_q is well correlated with $-\rho_p$ because they have negative density in common. It is worthwhile noting that the values of CORRq_{19} and $\langle\Delta\varphi_q\rangle_{19}$ for 6ebx′ are better than for 6ebx (0.66 against 0.53), so confirming our prediction (see §9) according to which the correlation between the estimated and the true ρ_q is expected to be higher if the model is more complete even if rough.

It may be interesting to calculate the correlation coefficient between the ideal difference Fourier synthesis and the difference syntheses calculated *via* the coefficients (1), (19) and (26), by considering separately the positive and negative parts of the maps. Indeed the ideal difference map will contain in the positive region target atoms not present in the model (or underestimated in terms of number of electrons), and in the negative region model atoms in incorrect positions (or overestimated in terms of electrons). A good correlation of the maps with coefficients (1), (19) and (26) with the ideal difference Fourier synthesis should indicate the capacity for discovering new atoms or for eliminating wrongly positioned model atoms. In Fig. 7 we show the values of such correlations for all the test structures (the corresponding diagrams are indicated by CORRq_{1+}, CORRq_{1-}, CORRq_{19+}, CORRq_{19-}, CORRq_{26+}, CORRq_{26-}). The following may be noted:

(i) CORRq_{1-} shows the lowest correlation values. This indicates a very low capacity for eliminating model atoms in the wrong position (*i.e.* large dependency on the model).

(ii) In all cases CORR$q_{1+} \geq$ CORRq_{1-}. This suggests a better capacity for locating target atoms not in the model, but only when the model is sufficiently good. For bad models both CORRq_{1+} and CORRq_{1-} tend to zero.

(iii) CORRq_{19+} and CORRq_{19-} are both well over CORRq_{1+} and CORRq_{1-}. In particular the coefficients (19) show a larger capacity for eliminating model atoms in the wrong position than locating new atoms not in the model: this behaviour is fully expected as an effect of the flipping term. The larger [with respect to the coefficients (1)] usefulness of the coefficients (19) for correcting the model is, however, remarkable. This capacity still persists when the model is very bad.

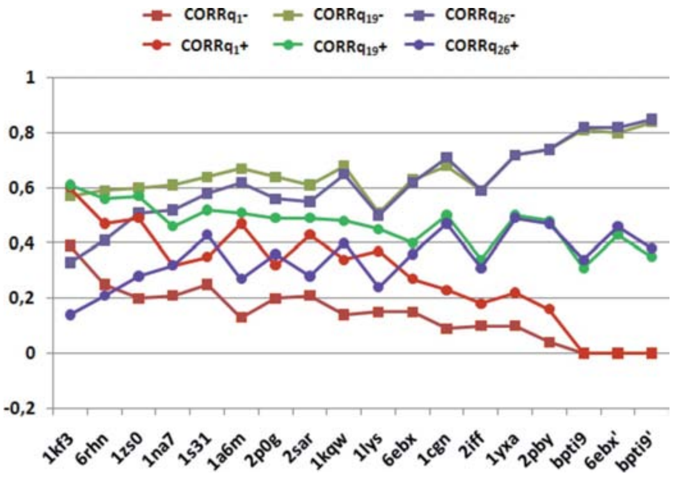

Figure 7
For each test structure the correlations between the ideal difference Fourier synthesis and the difference syntheses calculated *via* the coefficients (1), (19) and (26), by considering separately the positive and negative parts of the maps, are shown. They are denoted by CORRq_{1+}, CORRq_{1-}, CORRq_{19+}, CORRq_{19-}, CORRq_{26+} and CORRq_{26-}, respectively.

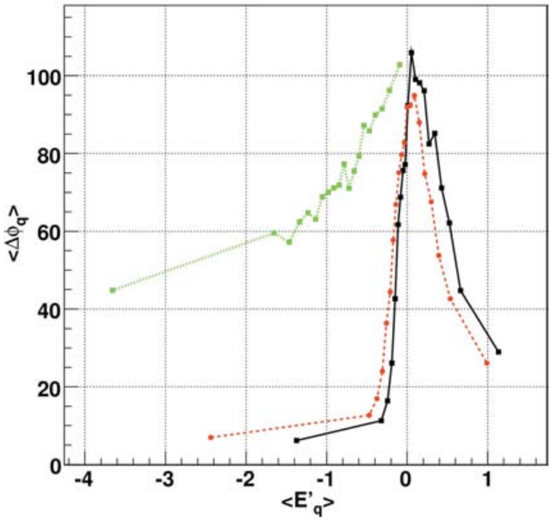

Figure 8
6rhn: (i) in black $\langle \Delta\varphi_q \rangle_1$ *versus* coefficients $\langle E'_q \rangle_1$; (ii) in red $\langle \Delta\varphi_q \rangle_{19}$ *versus* the coefficient $\langle E'_q \rangle_{19}$; (iii) in green, $\langle \Delta\varphi_q \rangle_{26}$ *versus* coefficients $\langle E'_q \rangle_{26}$.

(iv) CORRq_{26+} and CORRq_{26-} tend to coincide with CORRq_{19+} and CORRq_{19-} for bad models, but they are inferior when the quality of the model improves.

Let us now analyse the distribution of the phase errors *versus* the estimated difference Fourier coefficient $\langle E'_q \rangle$: we selected three test cases, 6rhn, 2p0g and 1yxa, having comparable RES values but different model qualities: very poor, medium and high, respectively. For each structure we divided the reflections into 20 batches, each batch containing an equal number of reflections and corresponding to a given value of $\langle E'_q \rangle$. In Figs. 8–10 we show (i) in black $\langle \Delta\varphi_q \rangle_1$ *versus* coefficients $\langle E'_q \rangle_1$; (ii) in red $\langle \Delta\varphi_q \rangle_{19}$ *versus* the coefficient $\langle E'_q \rangle_{19}$; (iii) in green, $\langle \Delta\varphi_q \rangle_{26}$ *versus* coefficients $\langle E'_q \rangle_{26}$. The following features may be noted:

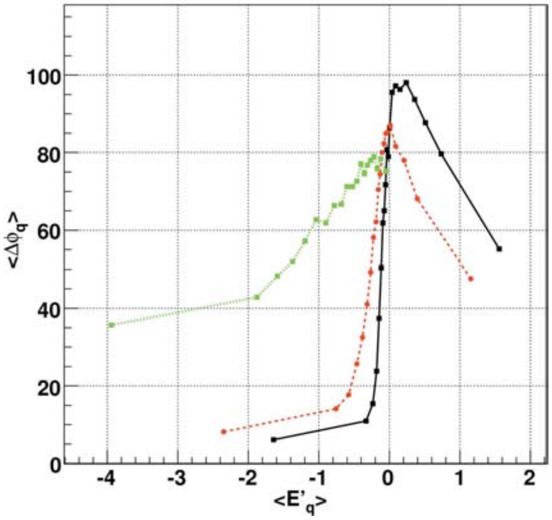

Figure 9
2p0g: (i) in black $\langle \Delta\varphi_q \rangle_1$ *versus* coefficients $\langle E'_q \rangle_1$; (ii) in red $\langle \Delta\varphi_q \rangle_{19}$ *versus* the coefficient $\langle E'_q \rangle_{19}$; (iii) in green, $\langle \Delta\varphi_q \rangle_{26}$ *versus* coefficients $\langle E'_q \rangle_{26}$.

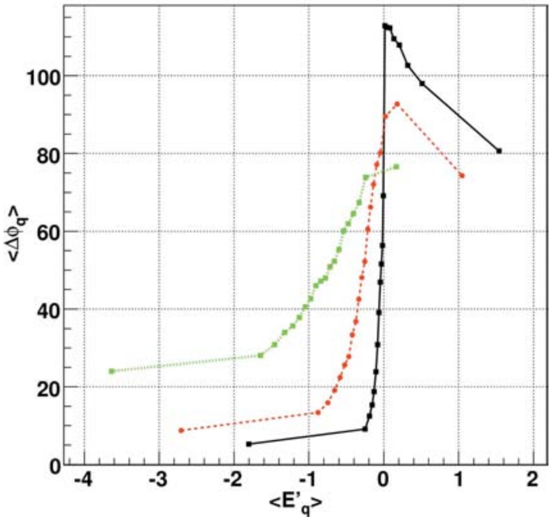

Figure 10
1yxa: (i) in black $\langle \Delta\varphi_q \rangle_1$ *versus* coefficients $\langle E'_q \rangle_1$; (ii) in red $\langle \Delta\varphi_q \rangle_{19}$ *versus* the coefficient $\langle E'_q \rangle_{19}$; (iii) in green, $\langle \Delta\varphi_q \rangle_{26}$ *versus* coefficients $\langle E'_q \rangle_{26}$.

The black curves are all asymmetric with respect to $\langle E'_q \rangle_1 = 0$, with higher mean phase errors for the positive part. In accordance with our theory, reflections with the same value of $(m|F| - D|F_p|)$ are expected to have different values of σ_q according to whether $|F|$ is large and $|F_p|$ is small, or *vice versa*: the phase estimates corresponding to strong negative values of $(m|F| - D|F_p|)$ are the most accurate.

(*a*) As an effect of the flipping term, the red curves are shifted to the left with respect to the black curves: the shift increases with the poorness of the model. They exhibit a smaller global mean phase error. The shift, however, is not sufficient to establish a perfect symmetry with respect to the zero point. The reason for this is unknown at the moment.

The green curves are monotonic, since they cover only the negative part of the $\langle E'_q \rangle$ axis. They show low phase errors for very poor models, but are completely inadequate when the model is sufficiently good.

To investigate the local properties of the new difference Fourier maps we used 6ebx (CORR = 0.50) as a test case: this is particularly interesting, since only one of the two symmetry-independent monomers of 6ebx (say monomer I) is covered by the molecular replacement model. By using the CCP4 suite (Collaborative Computational Project, Number 4, 1994), we calculated the main-chain residue-by-residue correlation coefficients (CORRES) between the electron density map calculated from the published coordinates and the difference electron density maps calculated by using (see Fig. 11): ideal coefficients (black, squares), coefficients (1) (red, circles), coefficients (19) (blue, triangles pointing up) and coefficients (26) (green, triangles pointing down). Both the electron density and the difference electron density maps have been calculated by using E-type syntheses. To understand the main features of Fig. 11 we divide the unit cell into two domains: domain I, occupied by the monomer I and its symmetry equivalents, and domain II, occupied by the second monomer and its equivalents. We will denote by CORRES$_{\rm id}$, CORRES$_1$,

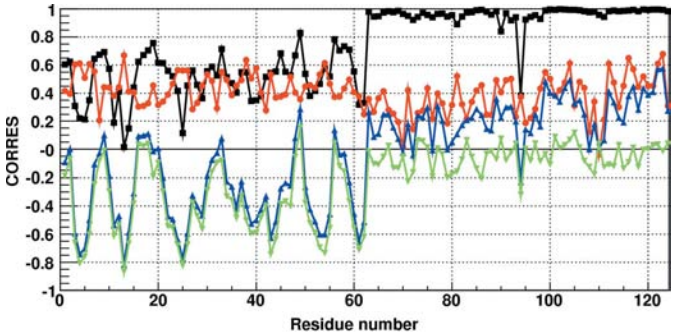

Figure 11
6ebx: main-chain residue-by-residue correlation coefficients (CORRES) between the electron density map calculated from the published coordinates and (*a*) the ideal difference electron density map (black, squares), (*b*) the map calculated *via* coefficients (1) (red, circles), (*c*) coefficients (19) (blue, upward-pointing triangles), (*d*) coefficients (26) (green, downward-pointing triangles).

CORRES_{19} and CORRES_{26} the CORRES values obtained *via* the ideal difference synthesis and by those obtained *via* coefficients (1), (19), (26), respectively.

Black line: in domain II the ρ_q peaks are expected to coincide with the ρ peaks: therefore $\mathrm{CORRES}_{\mathrm{id}}$ is expected to be close to unity in this domain. Differences from unity may be noted in Fig. 11: they are mostly due to finite resolution effects in the difference Fourier synthesis. In domain I $\mathrm{CORRES}_{\mathrm{id}}$ will attain its maximum in the correct residue positions, provided ρ_p did not locate sufficient electron density there. The minima of CORRES are expected where ρ_p peaks do not overlap with the correct residue positions: there ρ_q, and therefore $\mathrm{CORRES}_{\mathrm{id}}$, will show strong minima.

Red line: CORRES_1 is slightly higher in domain I than in domain II. Its values are probably too low to allow the recovery of the full structure.

Green line: ρ_{qflip}, the electron density calculated *via* the flipping term, is essentially coincident with $-\rho_p$, which is expected to be flat and almost vanishing in domain II. Accordingly, in this domain CORRES_{26} is expected to be zero (in Fig. 11 CORRES_{26} fluctuates around zero because of rounding errors in the Fourier synthesis). In domain I ρ_{qflip} should present deep negative minima where model atoms are located: accordingly, CORRES_{26} is expected to show strong negative minima where ρ_p peaks overlap residues, and maxima (slightly positive or negative) if ρ_p peaks lie in wrong positions.

Blue line: $\langle\rho_q\rangle_{19}$ is the sum of two electron densities: that calculated *via* the coefficients (1) (say $\langle\rho_q\rangle_1$) and ρ_{qflip}. This second component vanishes in domain II, where the peak distribution is determined by $\langle\rho_q\rangle_1$ only. Accordingly, CORRES_{19} and CORRES_1 are expected to coincide in domain II, as Fig. 11 shows.

Since the model structure for 6ebx is rather poor, ρ_{qflip} will be the dominant component in domain I. Therefore CORRES_{19} and CORRES_{26} should practically coincide. Accordingly, CORRES_{19} is expected to show minima where the residues are correctly located, and maxima (slightly positive or negative) where they are in wrong positions.

Fig. 11 suggests the following final considerations:

(*a*) In domain II the curves corresponding to coefficients (19) and (1) are almost coincident because the flipping term (arising from ρ_p) does not significantly influence this domain. Therefore the source of the inequality $\mathrm{CORR}q_{19} \gg \mathrm{CORR}q_1$ (see Table 2) should mostly lie in domain I. There the curves corresponding to coefficients (19) and to the ideal difference synthesis have a very similar trend (maxima and minima of the first almost coincide with maxima and minima of the second) but they are shifted by a constant quantity, so that the curve (19) is much more negative on average. This is an effect of the flipping term, an effect which may be overcome as described in point (*b*) below.

(*b*) If a function with the same trend of $-\rho_{\mathrm{qflip}}$ is added to $\langle\rho_q\rangle_{19}$ in such a way that the maxima $-\rho_{\mathrm{qflip}} + \langle\rho_q\rangle_{19}$ are strongly positive (its minima may remain negative or weakly positive), then the corresponding value of CORRES will be high and positive. That is what we expect to accomplish by the tangent formula (24): necessarily w_q is expected to be a function of σ_A and R_p. These last aspects concern the practical applications of the present theory and are deferred to a future paper.

A further test may be useful to have a direct estimate of the quality of different kinds of calculated syntheses: to compare the main-chain residue-by-residue correlation coefficients (CORRESD) between the ideal difference map and the difference maps calculated *via* coefficients (1), (19) and (26). The results for 6ebx are shown in Fig. 12 by red circles, blue upward-pointing triangles and green downward-pointing triangles, respectively. We note the following:

(i) The red curve is poorly correlated with the ideal difference Fourier map in domain I, while in domain II the correlation increases.

(ii) The green curve, on the contrary, is poorly correlated with the ideal difference Fourier map in domain II, while it is well correlated in domain I.

(iii) The blue curve follows the green curve in domain I and the red one in domain II, so taking the best part from the two curves.

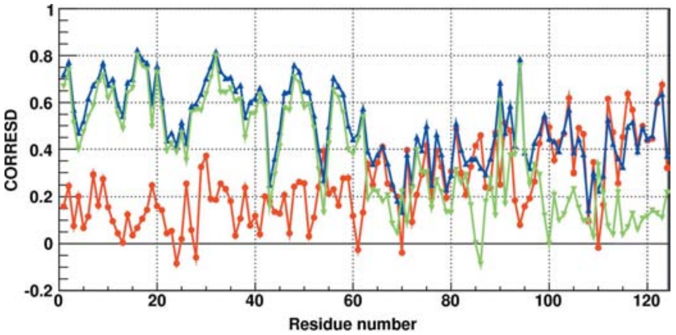

Figure 12
6ebx: main-chain residue-by-residue correlation coefficients (CORRESD) between the ideal difference electron density map and (*a*) the map calculated *via* coefficients (1) (red, circles), (*b*) coefficients (19) (blue, upward-pointing triangles), (*c*) coefficients (26) (green, downward-pointing triangles).

Finally, a window on the next developments may be useful to better understand both the meaning of the formulas and the potential of the present theory. We have previously stated that an estimation of ρ_q (or, in reciprocal space, of φ_q) is always possible even for random models if equation (19) is used. This does not imply that equation (19) directly carries the information on the target structure, but only that it is well correlated with the ideal electron density map ρ_q: our tests clearly show the correctness of this expectation. However, the principal component of the ρ_q estimate arises from the random map $-\rho_p$ and therefore does not directly contain information on the target map.

Let us now suppose the following:

(i) The ρ_q map is calculated *via* the coefficients (19) and suitably modified. Usually the modifications improve the quality of the map only if it is well correlated with the true map: if it is far away, the quality of the original map is degraded. This trend is widely confirmed by our previous experience, gained through the applications of the *DEDM–EDM* algorithm to protein models obtained *ab initio*, by molecular replacement or by *SAD–MAD* techniques.

(ii) The modification improves the quality of the map, since it is well correlated (as testified by our applications) with the ideal one.

(iii) The structure factor F_q obtained by inverse Fourier transform of the original map is summed to the structure factors F_p by the tangent formulas (22) or (24).

Then it may be expected that a fragment of the target structure may be found in the observed Fourier synthesis calculated by using the computed φ phases.

The above scheme outlines a new algorithm for obtaining correct structures from random models: it tends to coincide with the *DEDM* algorithm when the model becomes sufficiently good. Presently such a new algorithm is under development, but the first applications show without any doubt the potential of the approach and the full correctness of this theoretical contribution.

12. Conclusions

We have studied the joint probability distribution functions of three isomorphous crystal structures, the target structure defined by the electron density ρ, a model defined by the electron density ρ_p and the difference structure $\rho_q = \rho - \rho_p$. The distribution takes into account both model and measurement errors. Useful marginal and conditional distributions were obtained which may be applied to a wide crystallographic area. Indeed the theory suggests:

(*a*) New coefficients for the difference Fourier synthesis. They are the sum of the *difference term* [*i.e.* $(mR - \sigma_A R_p)$] and of the *flipping term*. If the model is very poor the flipping term is dominant; it is negligible only when the model is an accurate approximation of the target structure.

(*b*) The usefulness of the new Fourier coefficients for passing from a random model [*e.g.* from atoms located in random positions] to a realistic model of the target structure. Such potential arises from the flipping term, which may provide useful information even when the model is completely random [*e.g.* when the difference terms $(mR - \sigma_A R_p)$ are vanishing].

(*c*) The usefulness of the new Fourier coefficients for refining a model obtained, *e.g.* by direct or Patterson methods, molecular replacement and *SIR–MIR* techniques.

(*d*) A more solid theoretical background to the *EDM–DEDM* procedures.

The first applications described above show the correctness of our mathematical approach. We can anticipate that the theory described above will really be able to obtain the correct target structure from a random model [point (*b*)] and that it may also be successfully applied to the areas described in points (*c*) and (*d*). The description of the corresponding phasing procedures will be the object of future papers.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G., Siliqi, D. & Spagna, R. (2007). *J. Appl. Cryst.* **40**, 609–613.

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

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2008). *Acta Cryst.* A**64**, 519–528.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2009*a*). *Acta Cryst.* D**65**, 249–256.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2009*b*). *Acta Cryst.* D**65**, 477–484.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2009*c*). *Acta Cryst.* A**65**, 512–527.

Cochran, W. (1951). *Acta Cryst.* **4**, 408–411.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. (1999). *Acta Cryst.* D**55**, 1555–1567.

Cowtan, K. (2002). *J. Appl. Cryst.* **35**, 655–663.

Cowtan, K. D. (1994). *Jnt CCP4/ESF-EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Giacovazzo, C. (1979). *Acta Cryst.* A**35**, 757–764.

Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.

Giacovazzo, C. & Siliqi, D. (2001). *Acta Cryst.* A**57**, 40–46.

Giacovazzo, C. & Siliqi, D. (2002). *Acta Cryst.* A**58**, 590–597.

Henderson, R. & Moffat, J. K. (1971). *Acta Cryst.* B**27**, 1414–1420.

Lunin, V. Y., Afonine, P. V. & Urzhumtsev, A. G. (2002). *Acta Cryst.* A**58**, 270–282.

Lunin, V. Yu. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

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

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

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

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

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

Ursby, T. & Bourgeois, D. (1997). *Acta Cryst.* A**53**, 564–575.

Zhang, K. Y. J., Cowtan, K. D. & Main, P. (2001). *International Tables for Crystallography*, Vol. F, pp. 311–331. Dordrecht: Kluwer Academic Publishers.