Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 18 November 2004
Accepted 21 March 2005

The partial structure with errors: a probabilistic treatment

Rocco Caliandro, Benedetta Carrozzini, Giovanni Luca Cascarano, Liberato De Caro, Carmelo Giacovazzo,* Marat Moustiakimov and Dritan Siliqi

Institute of Crystallography – CNR, Via G. Amendola 122/O, 70126 Bari, Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

The method of the joint probability distribution functions has been applied to the case in which observed (with errors) and calculated structure factors are available, the latter referred to a part of the structure with finite errors in the coordinates, the thermal parameters and the scattering factors. Results obtained by other authors are confirmed and generalized. A new relationship is found to estimate the parameter σ_A, affecting the reliability of the estimates of $\cos(\varphi - \varphi_p)$. Some practical applications are described.

1. Notation

N: number of atoms in the unit cell of the model structure conventionally referred as *true structure*.

$p = l + t$: number of atoms in the structure conventionally referred as *model structure*. l atoms are located with small deviations with respect to the matched atoms in the *true structure*, t, with large deviations. Usually $p \leq N$ but also $p > N$ is possible.

$\mathbf{r}_j$, $j = 1, \ldots, N$: atomic positions in the true structure.

f_j, $j = 1, \ldots, N$: scattering factors of the atoms in true positions (f_j^0 is the scattering factor of the jth atom at rest).

$\mathbf{r}'_j = \mathbf{r}_j + \Delta\mathbf{r}_j$, $j = 1, \ldots, p$: positional vectors of the atoms belonging to the *model structure*.

B_{Tj}, $j = 1, \ldots, N$: isotropic temperature factors of the N atoms in the unit cell of the true structure.

g_j, $j = 1, \ldots., p$: scattering factors of the atoms of the model structure (g_j^0 is the scattering factor of the jth atom at rest).

$B'_{Tj} = B_{Tj} + \Delta B_{Tj}$, $j = 1, \ldots, p$: isotropic temperature factor for the atoms in the model structure.

$s = 2\sin\theta/\lambda$.

$F = \sum_{j=1}^{N} f_j^0 \exp(-s^2 B_{Tj}/4)\exp(2\pi i\mathbf{h}\mathbf{r}_j)$: 'true' structure factor for the reflection $\mathbf{h}$.

$F_p = \sum_{j=1}^{p} g_j^0 \exp(-s^2 B'_{Tj}/4)\exp(2\pi i\mathbf{h}\mathbf{r}'_j)$: structure factor of the model structure.

$E = A + iB = R\exp(i\phi)$, $E_p = A_p + iB_p = R_p\exp(i\phi_p)$: normalized structure factors of F and F_p, respectively.

ε: correction factor for expected intensities in reciprocal-lattice zones (from Wilson statistics).

$\Sigma_N = \varepsilon \sum_{j=1}^{N} f_j^2$.

$\Sigma_p = \varepsilon \sum_{j=1}^{l} g_j^2 + \varepsilon \sum_{j=l+1}^{p} g_j^2$.

$\Sigma_q = \varepsilon \sum_{j=p+1}^{N} f_j^2 \quad \text{if } p < N$.

$$\sigma_A = \left\langle \left\{ \sum_{j=1}^{l} f_j g_j \exp[-s^2\Delta B_j/4]\cos(2\pi\mathbf{h}\Delta\mathbf{r}_j) \right\} \right\rangle \Big/ (\Sigma_N\Sigma_p)^{1/2}: \tag{1}$$

the average involves the l atoms and is performed per resolution shell.

$$D = \langle \exp[-s^2\Delta B_T/4]\cos(2\pi\mathbf{h}\Delta\mathbf{r})\rangle: \tag{2}$$

the average involves the l atoms and is performed per resolution shell.

2. Introduction

The passage from an inaccurate and/or incomplete to a refined model of the scattering density is one of the most crucial points in structural crystallography. Luzzati (1952) applied the method of Wilson to determine the statistical distribution of the errors in the structure factors when the atomic positions have small but non-vanishing errors. To improve the efficiency of the weighting in Fourier series calculation of the electron density, Sim (1959) provided the probability distribution of the structure-factor phases when a partial model without errors is available. Srinivasan & Ramachandran (1965) derived the probability of the observed structure factor in a more general case, when the located atoms have errors in the coordinates. They showed that the probability distribution of the phase depends on two parameters, α and β, which are supposed to be constant in narrow spherical layers of s^2. Lunin & Urzhumtsev (1984) suggested that α and β should be chosen with a view to maximizing the likelihood. Lunin & Skovoroda (1995) underlined that maximum-likelihood-based estimates of α and β are good for models not subjected to refinement. Read (1986) showed that in a number of cases the two parameters may be reduced to a single one, the parameter σ_A. A general treatment of the previous contributions has been presented by Pannu & Read (1996) (see also Murshudov *et al.*, 1997), with particular emphasis on the refinement of macromolecular structures, where errors on the experimental values of structure-factor magnitudes were taken into account. Read (1986) originally calculated σ_A in resolution bins, each covering a narrow resolution range: a more general technique, relying on the fitting of smooth spline functions of resolution, has been proposed by Cowtan (2002).

The present paper is devoted to the generalization of the Srinivasan & Ramachandran (1965) distribution and of the Read (1986) contribution to estimate the value of σ_A (from which one can estimate the mean coordinate error for the model structure) by including measurement errors in the probabilistic treatment. We will use the method of the joint probability distribution functions, already used by Hauptman (1982) when he integrated direct methods and isomorphous replacement techniques. In his formulation, corresponding atoms in the two isomorphous structures may have different scattering factors but equal coordinates and no error in measurements. In this paper, these limitations will be overcome by using the approach described by Giacovazzo & Siliqi (2002), who considered errors of various natures. A new statistical relationship has been found, which is able to estimate the parameter σ_A. Practical applications of the new relationship are also shown.

3. Theory

Suppose that the atoms of a model structure are defined by the parameters $\mathbf{r}_j$, f_j^0, B_{Tj}, $j = 1, \ldots, N$. We will conventionally refer to this structure as the *true structure*. Suppose also that the observed values of the corresponding structure factor are measured with some error, *i.e.*

$$F = \sum_{j=1}^{N} f_j^0 \exp(-s^2 B_{Tj}/4)\exp(2\pi i \mathbf{h}\mathbf{r}_j) + \mu\exp(i\vartheta), \qquad (3)$$

where $\mu\exp(i\vartheta)$ is the (complex) error.

Let F_p be the structure factor of another model structure (from now on conventionally called the *model structure*) defined by the atomic parameters $\mathbf{r}'_j = \mathbf{r}_j + \Delta\mathbf{r}_j$, g_j^0 and $B'_{Tj} = B_{Tj} + \Delta B_{Tj}$, $j = 1, \ldots, p$:

$$F_p = \sum_{j=1}^{p} g_j^0 \exp(-s^2 B'_{Tj}/4)\exp(2\pi i \mathbf{h}\mathbf{r}'_j).$$

We will conventionally refer to the $\Delta\mathbf{r}_j$'s and the ΔB_{Tj}'s as errors (with respect to the true structure) in the atomic positions and in the vibrational parameters. They are supposed to be small for $j \le l$ and large for $j = l+1, \ldots, p$ (*i.e.* the model may contain atoms in completely the wrong positions).

We will calculate the joint probability distribution function $P(E, E_p)$ under the following conditions.

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

(*b*) The variables $\mathbf{r}'_j$, $j = 1, \ldots, p$, are riding variables: they are correlated with the corresponding $\mathbf{r}_j$'s through the local positional errors $\Delta\mathbf{r}_j$.

(*c*) The $\Delta\mathbf{r}_j$ are local variables, statistically independent of the $\mathbf{r}_j$: their moduli are restrained to assume sufficiently small values to secure, at least at low resolution, the isomorphism between the l-atom substructure of the model structure and the true structure.

(*d*) The coordinates of the vectors $\mathbf{r}_j$, $j = l+1, \ldots, p$, are primitive random variables, uniformly distributed in the unit cell. As a consequence, the t-atom substructure is completely uncorrelated with the true structure.

(*e*) The ΔB_{Tj}'s are local variables that are supposed to be close to zero for $j \le l$ and may assume any positive or negative value for $j = l+1, \ldots, p$.

(*f*) In accordance with Read (1990), the matched atoms into the two structures may have, in general, different scattering factors. While Read considers complex scattering factors and neglects errors in the measurements of the structure-factor amplitudes, we assume real scattering factors and we do consider the measurement errors. In this perspective, we introduce two supplementary primitive random variables, μ and θ, arising from the experimental uncertainty on the observed structure factor.

In accordance with §1,

$$A = \sum_{j=1}^{N} f_j^0 \exp(-s^2 B_j/4)\cos(2\pi\mathbf{h}\mathbf{r}_j)/(\Sigma_N)^{1/2}$$

$$B = \sum_{j=1}^{N} f_j^0 \exp(-s^2 B_j/4)\sin(2\pi\mathbf{h}\mathbf{r}_j)/(\Sigma_N)^{1/2}$$

$$A_p = \left\{ \sum_{j=1}^{l} g_j^0 \exp[-s^2(B_j + \Delta B_j)/4]\cos[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)] + \sum_{j=l+1}^{p} g_j^0 \exp[-s^2(B_j + \Delta B_j)/4]\cos[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)] \right\} / (\Sigma_p)^{1/2}$$

$$B_p = \left\{ \sum_{j=1}^{l} g_j^0 \exp[-s^2(B_j + \Delta B_j)/4]\sin[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)] + \sum_{j=l+1}^{p} g_j^0 \exp[-s^2(B_j + \Delta B_j)/4]\sin[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)] \right\} / (\Sigma_p)^{1/2}$$

are the real and imaginary parts of E and E_p, respectively. Then (see Appendix *A*), the joint probability distribution $P(A, A_p, B, B_p)$ is the four-dimensional Gaussian distribution

$$P(A, A_p, B, B_p) = \pi^{-2} e^{-1} (\det\mathbf{K})^{-1/2} \exp\left\{ -\frac{1}{(e - \sigma_A^2)} \times [(A^2 + B^2) + e(A_p^2 + B_p^2) - 2\sigma_A(AA_p + BB_p)] \right\}, \qquad (4)$$

where

$$e = (1 + \sigma_R^2), \quad \sigma_R^2 = \langle|\mu|^2\rangle/\Sigma_N.$$

In terms of normalized moduli and phases, equation (4) becomes

$$P(R, R_p, \phi, \phi_p) = RR_p \pi^{-2} \mathrm{e}^{-1} (\det\mathbf{K})^{-1/2} \exp\left\{ -\frac{1}{(e - \sigma_A^2)} \times [R^2 + eR_p^2 - 2\sigma_A RR_p \cos(\phi - \phi_p)] \right\}. \qquad (5)$$

The distribution (5) is the most general result obtained in this paper: from it the following marginal and conditional distributions may be calculated:

$$P(R, R_p) = 4RR_p e^{-1}(\det \mathbf{K})^{-1/2} \times \exp\left\{-\frac{1}{(e-\sigma_A^2)}[R^2 + eR_p^2]\right\} I_0[X] \qquad (6)$$

$$P(R|R_p) = \frac{2R}{(e-\sigma_A^2)} \exp\left\{-\frac{1}{(e-\sigma_A^2)}[R^2 + \sigma_A^2 R_p^2]\right\} I_0[X] \qquad (7)$$

$$P(\phi|R, R_p, \phi_p) = [2\pi I_0(X)]^{-1} \exp\{X\cos(\phi - \phi_p)\}, \qquad (8)$$

where

$$X = \frac{2\sigma_A R R_p}{(e-\sigma_A^2)}. \qquad (9)$$

It may be worthwhile noticing the following points.

(i) Not all the p atoms of the model structure contribute to D (see §1), but only l of them; just those whose position is correlated with a corresponding atom in the true structure. Furthermore, the t (wrongly placed atoms) contribute to the denominator of σ_A: this reflects the fact that, if the wrong atoms are omitted from the model, the model would be improved as a result. Usually, l is an unknown parameter.

(ii) When $l = p$ (no atom of the model is in a completely wrong position), $\Delta\mathbf{r}_j = \Delta B_{Tj} = 0$ for $j = 1, \ldots, p$ and $e = 1$ (no measurement error on the diffraction intensities), then $\sigma_A = (\Sigma_p/\Sigma_N)^{1/2}$ and, in accordance with Sim's results, $X = 2R'R'_p$, where R' and R'_p are moduli normalized with respect to the unknown part of the structure.

(iii) If $l < p$, $f_j = g_j$ for $j = 1, \ldots, p$, $e = 1$ and errors in the model are allowed, then (Read, 1986) $\sigma_A = D(\Sigma_p/\Sigma_N)^{1/2}$.

(iv) When $f_j^0 = g_j^0$, $\Delta B_{Tj} = 0$ for $j = 1, \ldots, p$ and $e = 1$, then (7) coincides with the Srinivasan & Ramachandran (1965) distribution.

(v) The parameters e and σ_A enter individually into the expression (5): thus they can be estimated from some suitable moments of equation (5).

(vi) D, and therefore σ_A, depend on the values of the parameters $\Delta\mathbf{r}_j$ and ΔB_{Tj}. The separate accurate evaluation of their effects through an expression like

$$D = \langle \exp[-s^2 \Delta B_T/4]\rangle \langle \cos(2\pi \mathbf{h}\Delta\mathbf{r})\rangle \qquad (10)$$

is generally impossible because $\Delta\mathbf{r}_j$ and ΔB_{Tj} are often correlated (*e.g.* ΔB_{Tj} may assume a large positive value to compensate for a large location error). Usually both the averages at the right-hand side of equation (10) decrease with s^2 (consequently D and σ_A decrease with s^2), unless the ΔB_{Tj} are predominantly negative. In this case, D and σ_A are constant or increase with s^2 (see §6).

4. The estimation of σ_A

In accordance with all the quoted authors, σ_A is expected to be resolution dependent. Its estimate is crucial both for evaluating the reliability of the assigned phases [through the reliability parameter X] and for the efficiency of the maximum-likelihood refinement processes [through the application of equation (7)]. The average implicit in its definition should therefore be performed by varying the reflection index at a constant value of s. We can therefore rewrite σ_A as

$$\sigma_A = \frac{\sum_{j=1}^{l} f_j g_j}{(\Sigma_N \Sigma_p)^{1/2}} D.$$

When the p atoms are perfectly located ($l = p$, $g = f$), the practical use of equations (6)–(9) requires the estimation of Σ_q. Henderson & Moffat (1971) and Bricogne (1976) suggested the equivalences

$$\Sigma_q = \langle 2(|F| - |F_p|)^2/\varepsilon\rangle$$

and

$$\Sigma_q = \langle 2||F|^2 - |F_p|^2|/\varepsilon\rangle,$$

respectively. Estimates of σ_A (in the absence of errors in measurements) were provided by Lunin & Urzhumtsev (1984) by minimizing the product of the conditional distributions $\prod_{\rm refl} P(R_p|R)$ with respect to two parameters defining σ_A. Read (1986) showed that the Lunin & Urzhumtsev result is equivalent to estimating σ_A by finding the zero of the residual function

$$\mathrm{RES} = \sum_{\rm ref} 2(\sigma_A - mRR_p), \qquad (11)$$

where

$$m = \langle \cos(\phi - \phi_p)\rangle = I_1(X)/I_0(X)$$

and $I_i(x)$ is the modified Bessel function of order i.

We follow a different approach. Equation (7) allows us to calculate any joint moment of the bivariate distribution $P(R, R_p)$, in particular

$$\langle R^2 R_p^2\rangle = \int_0^\infty\int_0^\infty R^2 R_p^2 P(R, R_p)\,\mathrm{d}R\,\mathrm{d}R_p,$$

which may be obtained by application of the relation

$$\int_0^\infty x^\mu \exp(-\alpha x^2) I_0(\beta x)\,\mathrm{d}x = \frac{\Gamma[(\mu+1)/2]}{2\alpha^{(\mu+1)/2}}\, {}_1F_1\left(\frac{\mu+1}{2}; 1; \frac{\beta^2}{4\alpha}\right), \qquad (12)$$

where Γ is the gamma function and ${}_1F_1$ is the confluent hypergeometric function.

We obtain the main result of this paper:

$$\langle R^2 R_p^2\rangle = (e + \sigma_A^2). \qquad (13)$$

If e is known from the counting statistics, equation (13) directly provides an estimate of σ_A, as defined by the general expression (1). In other words, there is no need to distinguish between the contribution of the parameters $\Delta\mathbf{r}_j$ to D and σ_A from the contribution of the parameters ΔB_{Tj}. Furthermore, in our treatment σ_A may be evaluated without the prior information on the l value. When this information is not available, as usual, we assume $l = p$, $g_j = f_j$ and rewrite equation (1) as

$$\sigma_A = \left(\frac{\Sigma_p}{\Sigma_n}\right)^{1/2} D, \qquad (14)$$

where

$$D = \langle \cos(2\pi \mathbf{h}\Delta\mathbf{r})\rangle \tag{15}$$

and

$$e - \sigma_A^2 = \sigma_R^2 + \frac{\Sigma_q}{\Sigma_N} + (1 - D^2)\frac{\Sigma_p}{\Sigma_N}. \tag{16}$$

From (15),

$$D \le 1 \tag{17}$$

and, from (14),

$$\sigma_A \le \left(\frac{\Sigma_p}{\Sigma_N}\right)^{1/2}. \tag{18}$$

Equation (13) generalizes a previous result by Hauptman (1982) obtained when $l = p$ and $e = 1$.

Equation (16) generalizes a previous result by Srinivasan & Ramachandran, originally obtained in the absence of errors on measurements. However, (16) is not generally valid when $l < p$ (that is, when the model contains atoms completely uncorrelated with the true structure). Furthermore, equations (14) and (15) do not take into account the effects of the ΔB_{Tj}: consequently also equations (17) and (18) are not generally valid.

Let us now calculate the expected value $\langle RR_p\rangle$. Owing to (12) and to the integral

$$\int_0^\infty \exp(-st)t^{b-1}{}_1F_1(a; c; kt)\,\mathrm{d}t = \Gamma(b)(s-k)^{-b}F\left(c-a, b; c; \frac{k}{k-s}\right)$$

(valid when $[|s-k| > |k|]$ and $[\mathrm{Re}(b) > 0, \mathrm{Re}(s) > \max(0, \mathrm{Re}(k))]$), we obtain (see Srinivasan & Parthasarathy, 1976, for the case $e = 1$)

$$\langle RR_p\rangle = \frac{\pi}{4}\frac{(e-\sigma_A^2)^2}{e^{3/2}}F\left(\frac{3}{2}, \frac{3}{2}; 1; \frac{\sigma_A^2}{e}\right),$$

where $F(\alpha, \beta; \gamma; z)$ is the Gaussian hypergeometric function. Since

$$F(\alpha, \beta; \gamma; z) = (1-z)^{\gamma-\alpha-\beta}F(\gamma-\alpha; \gamma-\beta; \gamma; z),$$

we have

$$\langle RR_p\rangle = \frac{\pi}{4}e^{1/2}F\left(\frac{-1}{2}, \frac{-1}{2}; 1; \frac{\sigma_A^2}{e}\right). \tag{19}$$

The series F always converges except when $\sigma_A^2 = e$. Its trend when $e = 1$ is shown in Fig. 1 by the full line: it is almost linear in σ_A^2. Accordingly, $\langle RR_p\rangle$ may be approximated by the function $(\pi/4)e^{1/2}[1 + (\pi/12)(\sigma_A^2/e)]$ (dashed line in Fig. 1). Using (19) is more complicated than using (13): we will mostly refer to (13) in the application section.

The relation (11) has been obtained by Lunin & Urzhumtsev (1984) *via* a maximum-likelihood criterion. By using the distribution (5) an equivalent result is obtained:

$$\langle mRR_p\rangle = \sigma_A. \tag{20}$$

We note: (*a*) the average at the left-hand side of (20) directly provides σ_A (the error e does not appear in the formula but it is used in the calculation of the X parameter); (*b*) the application of (20) as well as the use of (11) requires the previous knowledge of σ_A (to estimate m). Since this information is not available *a priori*, a refinement procedure is necessary (Read, 1986); (*c*) equations (13) and (19) are computationally more convenient than (20). We will employ (13) in our applications rather than (19), since its use requires no approximation.

5. The estimate of $\cos(\phi - \phi_p)$

Sim (1959) suggested that X could be rewritten in a simple form when the partial structure is perfect: $X = 2R'R'_p$, where $R' = F/\Sigma_q$, $R'_p = F_p/\Sigma_q$ (structure factors normalized with respect to the rest of the structure). Let us introduce (non-general) equation (14) into equation (10): the latter may be rewritten in the form (see Read, 2003)

$$X = \frac{2F(DF_p)}{(e\Sigma_N - D^2\Sigma_p)}. \tag{21}$$

Equation (21) may be interpreted as follows.

(*a*) $F'_p = DF_p$ is the structure factor statistically representative of the partial structure. Accordingly, $D^2\Sigma_p$ is the expected value of $\langle |F'_p|^2\rangle$.

(*b*) $(e\Sigma_N - D^2\Sigma_p)$ is statistically representative of Σ_q.

Accordingly, X may be again expressed in the Sim form $X = 2R'R'_p$.

Let us now introduce the general relation (13) into equation (10). An estimate of X in terms of statistical averages is obtained:

$$X = \frac{2(\langle R^2R_p^2\rangle - e)^{1/2}}{(2e - \langle R^2R_p^2\rangle)}RR_p. \tag{22}$$

Equation (22) suggests that the expected value of $\langle R^2R_p^2\rangle$ should lie in the interval $(e, 2e) \equiv (1 + \sigma_R^2, 2 + 2\sigma_R^2)$. Indeed, when two isomorphous structures perfectly coincide (*i.e.* when $nl = 0$, $p = N$, $\Delta\mathbf{r}_j = \Delta B_j = 0$, for $j = 1, \ldots, n$) and there is no measurement error in the data (*i.e.* $e = 1$) then $R_p = R$ and

$$\langle R^2R_p^2\rangle = \langle R^4\rangle = 2,$$

according to Wilson statistics. The larger the deviation from 2, the more imperfect is the model. In the limit case in which the

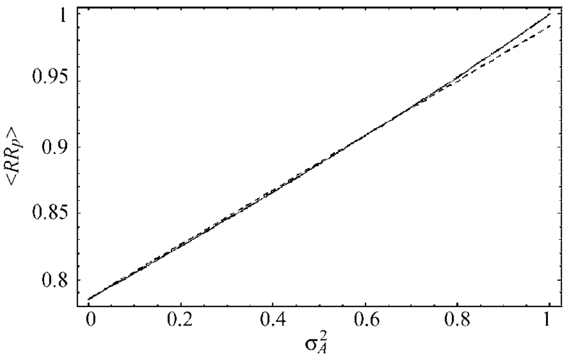

Figure 1
The function $(\pi e^{1/2}/4)F(\frac{-1}{2}, \frac{-1}{2}; 1; \sigma_A^2/e)$ (full line) and its approximation $y = (\pi e^{1/2}/4)[1 + (\pi/12)(\sigma_A^2/e)]$ (dashed line) in the interval (0, 1), for $e = 1$.

two structures are completely uncorrelated and there is no measurement error in the data then

$$\langle R^2 R_p^2 \rangle = \langle R^2 \rangle \langle R_p^2 \rangle = 1.$$

The curves $\langle R^2 \rangle$, $\langle R_p^2 \rangle$ and $\langle R^2 R_p^2 \rangle$ show strong Debye effects, which are responsible for the deviations of $\langle R^2 \rangle$ and $\langle R_p^2 \rangle$ from unity, and of $\langle R^2 R_p^2 \rangle$ from the interval $(e, 2e)$. In accordance with Read (1986), it is better to reduce these effects by renormalizing R^2 and R_p^2 within each resolution shell.

Equation (22) is equivalent to (21): we will refer to equation (22) in our applications, where we apply the following restriction: the experimental value of $\langle R^2 R_p^2 \rangle$ is not allowed to lie outside the interval $(e + 0.2, 2e - 0.2)$.

6. Practical applications

The theoretical results obtained in §§4 and 5 were implemented into a modified version of the program *SIR2004* (Burla *et al.*, 2005) and applied to several practical cases. Owing to their different behaviours, we will report the results obtained for the experimental data of two of them.

● The protein crambin (Weeks *et al.*, 1995), space group $P2_1$, 362 atoms in the asymmetric unit. The original data include 28725 reflections at 0.83 Å resolution.

● The protein RNase59 (Berisio *et al.*, 2002), space group $P2_1$, 950 atoms in the symmetric unit. The original data include 53354 reflections at 1.05 Å resolution.

To elucidate the effect of measurement errors, we first applied equation (13) to calculated data of crambin. The 362 atoms in the asymmetric unit were used to compute the F values ($B_{Tj} = 6$ for all the j's): a subset of atoms, corresponding to the fraction $\Sigma_p/\Sigma_N = 0.48$ was used to calculate the F_p's (again we assigned $B_{Tj} = 6$ to all the atoms of the subset). Wilson plot normalization returned an average B_T factor equal to 5.99 in both the cases. In Fig. 2, we plot σ_A [as obtained from equation (13)] *versus* resolution: obviously the parameter e was 1 for all the reflections because we are dealing with calculated error-free data. In the same figure, the real $\langle \cos(\phi - \phi_p) \rangle$ values are compared with their expected m values. We note: (*a*) all three least-squares straight lines, fitting σ_A, $\langle \cos(\phi - \phi_p) \rangle$ and m, respectively, are practically horizontal, as expected for a perfect partial structure. In particular, the average value of σ_A is 0.69, in quite good agreement with equation (18); (*b*) m and $\langle \cos(\phi - \phi_p) \rangle$ plots practically overlap. In Fig. 3, we plot (independently of the resolution) $\langle \cos(\phi - \phi_p) \rangle$ *versus* m: the plot is very close to the diagonal of the figure (the ideal case).

The tests with the experimental data were performed as follows. *SIR2004* is able to solve *ab initio* both the structures: the structural refinement (automatically performed by the program *via* modified electron-density techniques) was interrupted at the stages at which the average phase error was 58° for crambin and 73° for RNase59. For the first, we selected a partial structure of 164 atoms (S included), characterized by the ratio $\Sigma_p/\Sigma_N \approx 0.48$, for the second a partial structure of 304 atoms (S included) corresponding to the value $\Sigma_p/\Sigma_N \approx 0.23$

In Fig. 4, σ_A, as calculated by equation (13) for the experimental diffraction data of crambin, is plotted as a function of the resolution. In the same figure, the real $\langle \cos(\phi - \phi_p) \rangle$ values are compared with the expected m values. We note the following. (*a*) The least-squares straight lines (not shown to allow a simple reading of the figure) fitting the σ_A and m values decrease with $\sin\theta/\lambda$. This feature well agrees with the trend of $\langle \cos(\phi - \phi_p) \rangle$. (*b*) The m values fit well the experimental $\langle \cos(\phi - \phi_p) \rangle$ values when calculated according to equation (10). This behaviour is confirmed by Fig. 3, where the $\langle \cos(\phi - \phi_p) \rangle$'s are compared (independently of the resolution) with the expected m's.

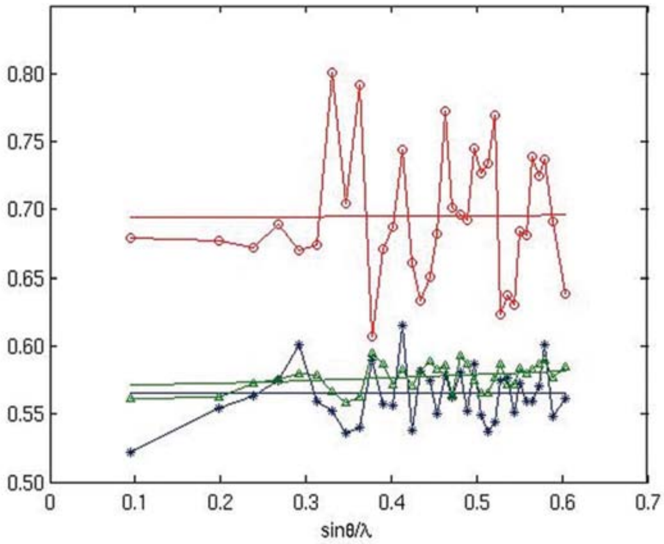

Figure 2
Crambin: calculated data. Open circles: σ_A plot according to equation (13) *versus* resolution. The least-squares straight line is calculated by omitting the point with the lowest value of $\sin\theta/\lambda$. Triangles: m plot according to $m = I_1(X)/I_0(X)$, where X is given by equation (9). Least-squares straight line calculated as before. Asterisks: experimental $\langle \cos(\phi - \phi_p) \rangle$ plot. Least-squares straight line calculated as before.

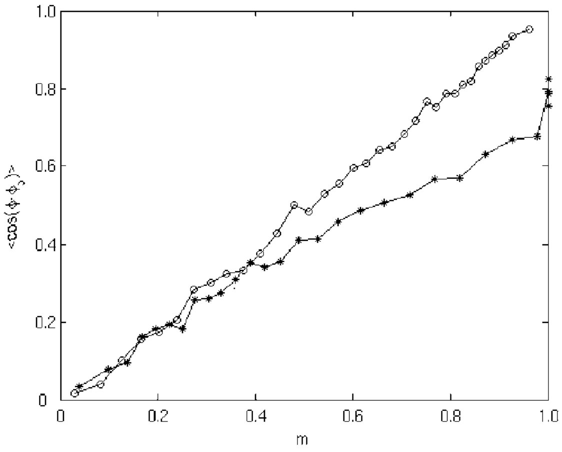

Figure 3
Crambin: m estimates [where $m = I_1(X)/I_0(X)$ and X is given by equation (9)] against $\langle \cos(\phi - \phi_p) \rangle$ values. Circles: calculated data; asterisks: experimental data.

In order to compare our results with those obtainable by application of previous theories, we plot in Fig. 4 the σ_A and m values obtained by the program (maximum-likelihood-based) *SIGMAA* (Collaborative Computational Project, Number 4, 1994)). We observe that the m's provided by *SIGMAA* overestimate the $\langle\cos(\phi - \phi_p)\rangle$'s.

The above results suggest the following conclusion. The overestimation of σ_A and, correspondingly, of $\langle\cos(\phi - \phi_p)\rangle$ may depend (besides other sources) on: (*a*) the experimental errors in the data; (*b*) on how much $\langle|\mu|^2\rangle$ is representative of the experimental error. In particular, it may be recalled that measurement errors are usually evaluated from diffraction intensities *via* Poisson statistics (integrated by standard deviations criteria when more than the asymmetric unit in the reciprocal space has been measured) and that systematic errors are not taken into account by such statistics (*e.g.* the solvent effects at low s values). Therefore, errors in the $\langle|\mu|^2\rangle$ moduli and in the $\langle|\mu|^2\rangle$ trend (with $\sin\theta/\lambda$) will cause, according to equation (13), errors in the σ_A estimates. To check this last observation, we plot in Fig. 5 the σ_A values, as calculated by equation (13) for RNase59, against the resolution, together with the corresponding m values. In the same figure, we also plot σ_A and m as calculated by *SIGMAA*. The m plots are compared with the experimental $\langle\cos(\phi - \phi_p)\rangle$ values. We observe the following.

(*a*) $\langle\cos(\phi - \phi_p)\rangle$ increases with $\sin\theta/\lambda$. This trend is shared by both m plots and is rather unexpected if one relies on equation (15) (this equation suggests that σ_A should always decrease with $\sin\theta/\lambda$). Actually, in all the literature quoted above, the model structures are divided into two categories: partial structures without coordinate errors and partial structures with coordinate errors. For the first category, σ_A is expected to be constant *versus* the resolution, while σ_A is expected to decrease for increasing values of $\sin\theta/\lambda$. It is worthwhile noting that the general expression for D is given by equation (1): the value of D depends also on the errors in the vibrational parameters, and these can be responsible for the inverse σ_A behaviour. Such a behaviour is not infrequent when the phasing process is at an intermediate state (*e.g.* when electron-density modification techniques are used to improve the phase estimates).

(*b*) Both (our and *SIGMAA*) m estimates overestimate the experimental $\langle\cos(\phi - \phi_p)\rangle$ values. To explain this result, we plot in Fig. 6 the value of e as a function of resolution for crambin and RNase59. It easy seen that crambin e values are remarkably larger than the corresponding values for RNase. If, in all the calculations for RNase, we multiply by a factor of 2, the experimental e values, the m plot calculated according to our approach will be remarkably closer to the experimental $\langle\cos(\phi - \phi_p)\rangle$ values. In conclusion, if $\langle|\mu|^2\rangle$ is too rough a representative of the experimental error, then the accuracy of the σ_A estimates will decrease. This is another reason for involving measurement errors in any mathematical model describing the isomorphism between a partial and the complete structure [see Murshudov *et al.* (1997) and Bricogne & Irwin (1996) for maximum-likelihood applications in the refinement context].

7. Conclusions

The probabilistic theory on the structure-factor distribution for two isomorphous structures, one of which is part of the second and shows errors in the atomic coordinates, has been generalized to include errors in the vibrational parameters and

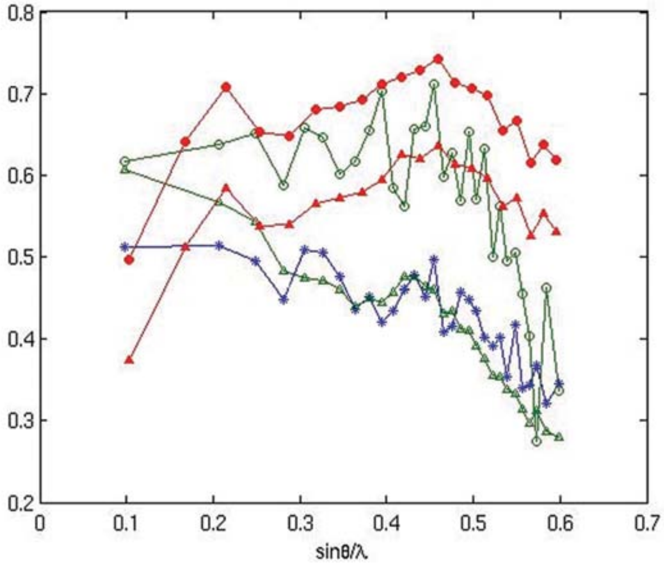

Figure 4
Crambin: experimental diffraction data. Open circles: σ_A plot according to equation (13) *versus* resolution. Asterisks: experimental $\langle\cos(\phi - \phi_p)\rangle$ plot. Open triangles: m plot according to $m = I_1(X)/I_0(X)$, where X is given by equation (9). Filled circles: σ_A plot, calculated by maximum-likelihood criteria, by the program *SIGMAA*. Filled triangles: m plot calculated by the program *SIGMAA*.

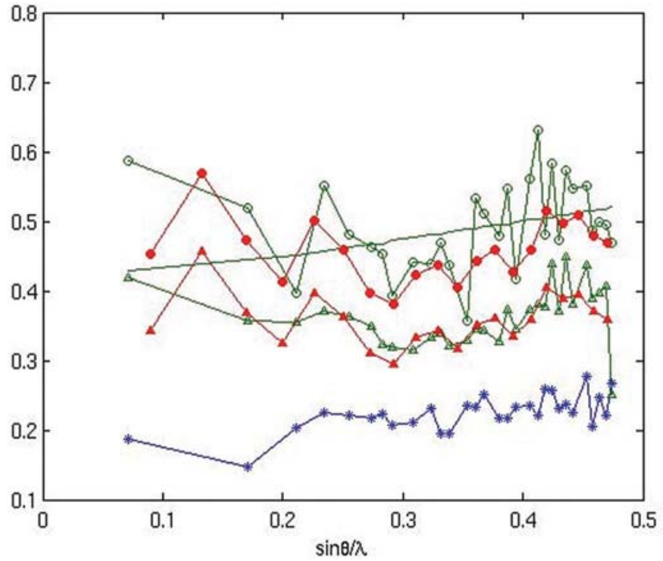

Figure 5
RNase59: experimental diffraction data. Open circles: σ_A plot according to equation (13) *versus* resolution. Asterisks: experimental $\langle\cos(\phi - \phi_p)\rangle$ plot. Open triangles: m plot according to $m = I_1(X)/I_0(X)$, where X is given by equation (9). Filled circles: σ_A plot, calculated by maximum-likelihood criteria, by the program *SIGMAA*. Filled triangles: m plot calculated by the program *SIGMAA*.

in measurements. A simple probabilistic expression has been found, which, applied to practical cases, allows one to improve the phase estimates provided by the classical Sim relationship.

APPENDIX *A*

To calculate the distribution $P(A, A_p, B, B_p)$, we first calculate the characteristic function

$$\begin{aligned}C(u, u_p, v, v_p) &= \langle \exp i(uA + u_pA_p + vB + v_pB_p)\rangle\\ &= \exp\{-\tfrac{1}{4}[e(u^2 + v^2) + u_p^2 + v_p^2\\ &\quad + 2\sigma_A(uu_p + vv_p)]\},\end{aligned}$$

where u, u_p, v, v_p are carrying variables associated with A, A_p, B, B_p, respectively. Then $P(A, A_p, B, B_p)$ is the Fourier transform of C. We have

$$\begin{aligned}P(A, A_p, B, B_p) &= \int\limits_{-\infty}^{+\infty} \ldots \int\limits_{-\infty}^{+\infty} \exp(i\overline{\mathbf{T}}\mathbf{U})\exp(-\tfrac{1}{2}\overline{\mathbf{U}}\mathbf{K}\mathbf{U})\,\mathrm{d}\mathbf{U}\\ &= (2\pi)^{n/2}(\det\mathbf{K})^{-1/2}\exp(-\tfrac{1}{2}\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}),\end{aligned}$$

where

$$\begin{aligned}\overline{\mathbf{U}} &= [(2/e)^{1/2}u', (2)^{1/2}u'_p, (2/e)^{1/2}v', (2)^{1/2}v'_p]\\ \overline{\mathbf{T}} &= [(2/e)^{1/2}A, (2)^{1/2}A_p, (2/e)^{1/2}B, (2)^{1/2}B_p].\end{aligned}$$

Since

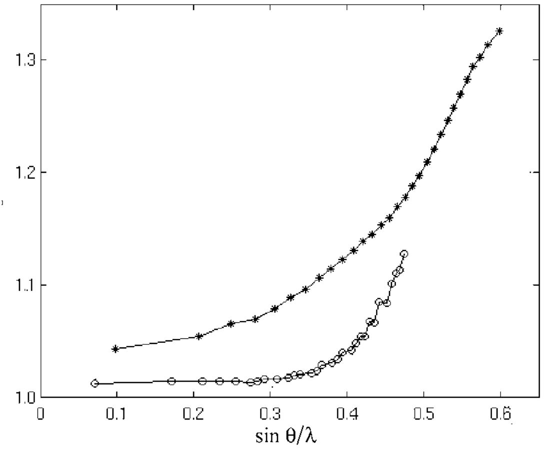

Figure 6
The *e* values for crambin (asterisks) and for RNase59 (circles) are plotted against resolution.

$$\mathbf{K} = \left|\begin{matrix}\mathbf{L} & 0\\ 0 & \mathbf{L}\end{matrix}\right|, \quad \mathbf{K}^{-1} = \left|\begin{matrix}\mathbf{L}^{-1} & 0\\ 0 & \mathbf{L}^{-1}\end{matrix}\right|, \quad \det\mathbf{K} = (e - \sigma_A^2)^2/e^2,$$

where

$$\mathbf{L} = \left|\begin{matrix}1 & \sigma_A/e^{1/2}\\ \sigma_A/e^{1/2} & 1\end{matrix}\right|,$$

we obtain

$$\begin{aligned}P(A, A_p, B, B_p) = \pi^{-2}e^{-1}(\det\mathbf{K})^{-1/2}\exp\bigg\{&-\frac{1}{(e - \sigma_A^2)}\\ &\times[(A^2 + B^2) + e(A_p^2 + B_p^2)\\ &- 2\sigma_A(AA_p + BB_p)]\bigg\},\end{aligned}$$

which coincides with equation (4).

We thank R. J. Read for enlightening discussions.

References

Berisio, R., Sica, F., Lamzin, V. S., Wilson, K. S., Zagari, A. & Mazzarella, L. (2002). *Acta Cryst.* D**58**, 441–450.
Bricogne, G. (1976) *Acta Cryst.* A**32**, 832–847.
Bricogne, G. & Irwin, J. J. (1996). Macromolecular Refinement. Proceedings of the CCP4 Study Weekend, SERC Daresbury Laboratory, Warrington, England, pp. 85–92.
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.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (2002). *J. Appl. Cryst.* **35**, 655–663.
Giacovazzo, C. & Siliqi, D. (2002). *Acta Cryst.* A**58**, 590–597.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Henderson, R. & Moffat, J. K. (1971). *Acta Cryst.* B**27**, 1414–1420.
Lunin, V. Yu. & Skovoroda, T. P. (1995). *Acta Cryst.* A**51**, 880–887.
Lunin, V. Y. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Read, R. J. (2003). *Acta Cryst.* D**59**, 1891–1902.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Srinivasan, R. & Parthasarathy, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.
Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.
Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.