Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 16 September 2010
Accepted 26 February 2011

About the σ_A estimate

Maria Cristina Burla,[a,b] Carmelo Giacovazzo,[b,c]* Annamaria Mazzone,[b] Giampiero Polidori[a,b] and Dritan Siliqi[b]

[a]Department of Earth Sciences, University of Perugia, 06100 Perugia, Italy, [b]Institute of Crystallography, CNR, Via G. Amendola 122/O, 70126 Bari, Italy, and [c]Dipartimento Geomineralogico, Università di Bari, 70125 Bari, Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

The resolution parameter σ_A is currently used for evaluating the degree of similarity between a model and the target structure. Here, quasi-Wilson distributions are used to represent the local statistics of the normalized amplitudes both for the target and for the model structure. The study uses the joint probability distribution approach to provide (i) a description of the statistical properties of the σ_A parameter; (ii) a deeper insight into the role, for the σ_A estimate, of the high-order moments of the target and of the model structure-factor distributions; and (iii) new statistical formulas for estimating σ_A. The theoretical results have been checked using test proteins.

1. Notation

N, p: number of atoms in the unit cell for the target and for the model structure, respectively. 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).
$F = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{hr}_j) = |F| \exp(i\varphi)$: structure factor of the target structure.
$F_p = \sum_{j=1}^{p} f_j \exp(2\pi i \mathbf{hr}'_j) = |F_p| \exp(i\varphi_p)$, where $\mathbf{r}'_j = \mathbf{r}_j + \Delta\mathbf{r}_j$: structure factor of the model structure.
$E = A + iB = R\exp(i\varphi)$, $E_p = A_p + iB_p = R_p \exp(i\varphi_p)$: normalized structure factors of F and F_p, respectively.
$\Sigma_N = \sum_{j=1}^{N} f_j^2$, $\Sigma_p = \sum_{j=1}^{p} f_j^2$.
$D = \langle \cos(2\pi \mathbf{h}\Delta\mathbf{r}) \rangle$. $\langle \Delta\mathbf{r} \rangle$ is the average vectorial difference between the p positional vectors of the model atoms and the corresponding vectors in the target structure. The calculation of D has to be made per resolution shell (D is expected to diminish for higher-resolution reflections).
$\sigma_A = D(\Sigma_p/\Sigma_N)^{1/2}$.
$\sigma_R^2 = \langle |\mu|^2 \rangle / \Sigma_N$, where $\langle |\mu|^2 \rangle$ is the measurement error.
$e = 1 + \sigma_R^2$.
$I_i(x)$: modified Bessel function of order i.
$s = \sin^2\theta/\lambda^2$.

2. Introduction

A wide literature exists on the joint probability distributions of structure factors of isomorphous structures: it has been used for studying, in the reciprocal space, the relationships between model and target structure, between native proteins and their derivatives, for finding substructures *via* anomalous-scattering effects *etc.* The simplest joint distribution is also the most important one: it relies on the structure factors of the target and of a model for the same reflection $\mathbf{h}$, and is denoted here as $P(E, E_p)$. It is often employed to drive the model phases towards the phases of the target structure. We quote the most important results in the study of this distribution, all obtained by considering the atomic positional vectors $\mathbf{r}_j$ as the primitive random variables and, when the case, $\Delta\mathbf{r}_j$ as local variables:

(*a*) Sim (1959) assumed a model structure whose atoms are located at the same sites of the target atoms (*i.e.* $\Delta\mathbf{r}_j = 0$ for $j = 1, \ldots, p$):

$$E = \left\{ \sum_{j=1}^{N} f_j \exp 2\pi i \mathbf{hr}_j \right\} \Big/ \left(\varepsilon \Sigma_N\right)^{1/2},$$

$$E_p = \sum_{j=1}^{p} f_j \exp 2\pi i \mathbf{hr}_j \Big/ \left(\varepsilon \Sigma_p\right)^{1/2}.$$

The theory associates the weight

$$m_S = D_1\left[2R'R'_p\right], \tag{1}$$

with the phase of the target structure, where R' and R'_p are structure-factor moduli normalized with respect to the rest of the structure, and $D_i(x) = I_i(x)/I_0(x)$.

(*b*) Srinivasan & Ramachandran (1965) used a more realistic approach, by allowing errors in the atomic coordinates of the model structure: *e.g.*

$$E = \sum_{j=1}^{N} f_j \exp 2\pi i \mathbf{hr}_j \Big/ \left(\varepsilon \Sigma_N\right)^{1/2},$$

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

(*c*) The same model was used by Read (1986): he used the likelihood function given by Lunin & Urzhumtsev (1984) to

provide the probability of the structure-factor magnitudes. The weight (1) was generalized into

$$m_{\rm SR} = D_1\big[2\sigma_A RR_p/(1-\sigma_A^2)\big]. \quad (2)$$

(*d*) Caliandro *et al.* (2005) derived a more general expression for $P(E, E_p)$, by considering both measurement errors [represented by the complex number $|\mu|\exp(i\vartheta)$] and errors in the model structure,

$$E = \left\{\sum_{j=1}^{N} f_j \exp(2\pi i\mathbf{hr}_j) + |\mu|\exp(i\vartheta)\right\}\Big/\big(\varepsilon\Sigma_N\big)^{1/2},$$
$$E_p = \sum_{j=1}^{p} f_j \exp[2\pi i\mathbf{h}(\mathbf{r}_j+\Delta\mathbf{r}_j)]\Big/\big(\varepsilon\Sigma_p\big)^{1/2}.$$

For each shell the value of σ_A may be obtained by the following relation,

$$\sigma_A^2 = \big(\langle R^2R_p^2\rangle - e\big). \quad (3)$$

The average is calculated per resolution shell. Accordingly, the weights (1) and (2) were generalized into

$$m = D_1\big[2\sigma_A RR_p/(e-\sigma_A^2)\big]. \quad (4)$$

From the above considerations the crucial role of the parameter σ_A for the efficiency of the phasing process is evident, particularly in protein crystallography where phasing is not straightforward. It may also be considered a useful figure of merit, monitoring the phasing progress: as an example we consider the expression

$$\langle|E-E_p|^2\rangle = \langle R^2\rangle + \langle R_p^2\rangle - 2\langle RR_p\cos(\varphi-\varphi_p)\rangle$$
$$\simeq \langle R^2\rangle + \langle R_p^2\rangle - 2\langle mRR_p\rangle,$$

which is expected to be minimum when the model coincides with the target. Since (Read, 1986; Caliandro *et al.*, 2005) $\langle mRR_p\rangle = \sigma_A$, the larger σ_A the smaller the vectorial difference between E and E_p.

In the *CCP4* package (Collaborative Computational Project, Number 4, 1994) a specific program (*SIGMAA*; Read, 1986) is dedicated to the σ_A estimation: measured reflections are partitioned in resolution shells (σ_A is a resolution-dependent parameter) and for each shell a maximum-likelihood estimate is derived. The procedure implies the normalization of the structure factor shell-by-shell, *i.e.* the locally normalized quantities

$$\frac{\langle|FF_p|^2\rangle}{\langle|F|^2\rangle\langle|F_p|^2\rangle}. \quad (5)$$

In terms of normalized structure factors (calculated according to the Wilson plot) the quantity (5) may be replaced by

$$\frac{\langle R^2R_p^2\rangle}{\langle R^2\rangle\langle R_p^2\rangle}. \quad (6)$$

Local renormalization has never been theoretically justified, but it is necessary in practice to relate the local values of $\langle F^2F_p^2\rangle$ with the marginal moments of the second order, say $\langle|F|^2\rangle$ and $\langle|F_p|^2\rangle$.

The need for an accurate σ_A estimate is crucial for any phasing method, particularly also for the *VLD* method (Burla, Caliandro *et al.*, 2010; Burla, Giacovazzo & Polidori, 2010), a new phasing approach using the properties of the Fourier transform for recovering the correct structure from a random model. But, in spite of the wide use of σ_A, not all of the statistical properties of σ_A are well known. In this paper we will:

(i) derive the formula, equivalent to equation (3), valid for centric crystals;

(ii) provide a theoretical justification for the local renormalization of the structure factors by using quasi-Wilson distributions;

(iii) study the effects, on the σ_A estimate, of the deviations of the structure-factor statistics from Wilson distributions; and

(iv) provide additional statistical tools for estimating σ_A.

The conclusive formulas will be applied to some test cases.

3. The calculation of σ_A in $P\bar{1}$

In accordance with point (i) of §2 we extend the approach of Caliandro *et al.* (2005) to centric space groups by calculating the joint probability distribution function $P(E, E_p)$ in $P\bar{1}$ under the following conditions:

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

(ii) The supplementary primitive random variable μ is used (μ is now a real number), arising from the experimental uncertainty of the observed structure-factor amplitude. Accordingly, the mathematical model for the structure factors will be

$$E = \left\{2\sum_{j=1}^{N/2} f_j \cos(2\pi\mathbf{hr}_j) + \mu\right\}\Big/\big(\Sigma_N\big)^{1/2},$$
$$E_p = 2\sum_{j=1}^{p/2} f_j \cos[2\pi\mathbf{h}(\mathbf{r}_j+\Delta\mathbf{r}_j)]\Big/\big(\Sigma_p\big)^{1/2}.$$

Since

$$\langle|F|^2\rangle = \Sigma_N + \langle\mu^2\rangle,\ \langle|F_p|^2\rangle = \Sigma_p \text{ and } \langle FF_p\rangle = D\Sigma_p,$$

we have

$$\langle R^2\rangle = 1+\sigma_R^2 = e, \quad \langle {R_p}^2\rangle = 1$$

and

$$\langle EE_p\rangle = D\big(\Sigma_p/\Sigma_N\big)^{1/2} = \sigma_A.$$

Finally, the characteristic function of the distribution $P(E, E_p)$ is

$$C(u,u_p) = \big\langle\exp i\big(uE+u_pE_p\big)\big\rangle$$
$$= \exp\big(-\tfrac{1}{2}eu^2 - \tfrac{1}{2}u_p^2 - \sigma_A uu_p\big),$$

from which we obtain

$$P(E, E_p) = \frac{1}{(2\pi)[(e-\sigma_A^2)]^{1/2}} \times \exp\left\{-\frac{1}{2(e-\sigma_A^2)}\left[eE_p^2 + E^2 - 2\sigma_A E E_p\right]\right\}.$$

Then

$$\langle R^2 R_p^2\rangle = e + 2\sigma_A^2 \tag{7}$$

or also

$$\sigma_A^2 = (1/2)(\langle R^2 R_p^2\rangle - e). \tag{8}$$

Equation (8) strongly differs from equation (3), the corresponding relation for $P1$: using (3) for centric space groups (or for acentric space groups with a high percentage of centric reflections) may lead to strong overestimates of σ_A. Equation (8) agrees well with Wilson statistics. Indeed, on supposing e very close to unity,

(i) if $\sigma_A = 1$, model and target structure coincide, and $\langle R^2 {R_p}^2\rangle = \langle R^4\rangle = 3$, the value expected by Wilson statistics;

(ii) if $\sigma_A = 0$, then $P(E)$ and $P(E_p)$ are uncorrelated, and $\langle R^2 R_p^2\rangle = \langle R^2\rangle\langle R_p^2\rangle = 1$.

Accordingly, the average $\langle R^2 R_p^2\rangle$ is expected to lie in the range 1–2 for acentric crystals, and in the range 1–3 for centric ones.

4. The role of higher-order moment in the σ_A estimation

So far we have assumed that the atoms are randomly distributed in the unit cell: this is equivalent to assuming that the experimental normalized structure-factor amplitudes fit the Wilson distributions

$$P_1(R) = 2R\exp(-R^2), \tag{9}$$

$$P_{\bar{1}}(R) = (2/\pi)^{1/2}\exp(-R^2/2), \tag{10}$$

for acentric and centric space groups, respectively. The above hypothesis is frequently violated in practice: indeed, because of chemical interactions, atoms are not randomly distributed and often the structure-factor statistics show strong deviations from Wilson equations. The consequent effect is that Wilson plots are never straight lines, but are usually curves wrapped around least-square straight lines: moments of $P(R)$ or of $P(R_p)$ may locally attain values strongly different from those foreseen by Wilson statistics. Such effects may be responsible for local overestimates or underestimates of the σ_A parameter. We will show that σ_A, a resolution-dependent parameter, depends not only on the value of the joint moment $\langle R^2 R_p^2\rangle$ but also on the local amplitude distributions of the target and of the model structures (which may be represented by the set of their moments).

Such dependence is not evident when one considers equations (3) or (8), just because some fourth-order moments are replaced by their numerical Wilson values. Indeed, a more careful inspection of the results obtained in §3 shows that (8) is the numerical result of the following equation (obtained when the integration on R_p is already performed; of course, one can first integrate over R and then over R_p),

$$\begin{aligned}\langle R^2 R_p^2\rangle &= (e-\sigma_A^2)(2/\pi)^{1/2}\int_0^\infty R^2\exp(-R^2/2)\,\mathrm{d}R \\ &\quad + (2/\pi)^{1/2}\sigma_A^2\int_0^\infty R^4\exp(-R^2/2)\,\mathrm{d}R \\ &= (e-\sigma_A^2)\langle R^2\rangle + \sigma_A^2\langle R^4\rangle.\end{aligned} \tag{11}$$

From (11), equation (8) arises because it is assumed that structure-factor magnitudes obey the Wilson distribution (10): then $\langle R^2\rangle = 1$ and $\langle R^4\rangle = 3$. Such assumptions are no longer valid if the local amplitude distribution does not fit Wilson statistics: then the moments should be replaced by more realistic values. We therefore need a mathematical tool to modify standard Wilson distributions in order to take into account the effects of the chemistry. This is described in §5.

5. Quasi-Wilson distributions

According to Debye (1915) the expected value of $|F_{\mathbf{h}}|^2$ per resolution shell is given by

$$\langle|F_{\mathbf{h}}|^2\rangle = \Sigma_N + \sum_{i\neq j=1}^{N} f_i f_j \frac{\sin 2\pi\mathbf{h}\mathbf{r}_{ij}}{2\pi\mathbf{h}\mathbf{r}_{ij}}, \tag{12}$$

where $h = |\mathbf{h}|$ and $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$. We will refer to the last term on the right-hand side of (12) as the *interference* term. Equation (12), combining atomic scattering and interference terms, gives a full account of the average scattering *versus s*. *Vice versa*, it may be used to estimate, *via* Fourier transform (Cascarano *et al.*, 1992), the shortest interatomic distances. Owing to crystal-chemical reasons, such distances are usually clustered (Hall & Subramanian, 1982*a*,*b*; Morris *et al.*, 2004). The overall effects, frequently occurring in any resolution shell, are:

(*a*) $\langle|F_{\mathbf{h}}|^2\rangle > \Sigma_N$ or $\langle|F_{\mathbf{h}}|^2\rangle < \Sigma_N$, according to the value of the local interference term. In terms of normalized structure factors this condition leads to the relation

$$\langle R_{\mathbf{h}}^2\rangle = 1 + \left(\sum_{i\neq j=1}^{N} f_i f_j \frac{\sin 2\pi\mathbf{h}\mathbf{r}_{ij}}{2\pi\mathbf{h}\mathbf{r}_{ij}}\right)/\Sigma_N = \delta,$$

where δ is a parameter oscillating about unity.

(*b*) The classical Wilson distribution is no longer satisfied: for example, the percentage of strong reflections may be much larger or smaller than that theoretically predicted by Wilson.

To describe such local statistical effects we will continue to consider the atomic positional vectors $\mathbf{r}_j$ as the primitive random variables, $\Delta\mathbf{r}_j$ as local variables, and we will define in acentric space groups

$$E = \delta^{1/2}\left[\sum_{j=1}^{N} f_j\exp(2\pi i\mathbf{h}\mathbf{r}_j) + |\mu|\exp(i\vartheta)\right]/\left(\Sigma_N\right)^{1/2},$$

$$E_p = \delta_p^{1/2}\sum_{j=1}^{p} f_j\exp[2\pi i\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)]/\left(\Sigma_p\right)^{1/2}.$$

The purpose of the above new definitions for E and E_p is to allow integrations like those described in (11) to be performed over moduli that do not strictly obey Wilson distributions. We obtain

$$P(R) = (2/e\delta)R\exp(-R^2/e\delta) \qquad (13a)$$

and

$$P(R_p) = (2/\delta_p)R_p\exp(-R_p^2/\delta_p). \qquad (13b)$$

Similar expressions may be obtained for the centric case: in particular, defining

$$E = 2\delta^{1/2}\left[\sum_{j=1}^{N/2} f_j\cos(2\pi\mathbf{hr}_j) + \mu\right]/\left(\Sigma_N\right)^{1/2}$$

and

$$E_p = 2\delta_p^{1/2}\sum_{j=1}^{p/2} f_j\cos[2\pi\mathbf{h}(\mathbf{r}_j + \Delta\mathbf{r}_j)]/\left(\Sigma_p\right)^{1/2}$$

leads to the distributions

$$P(R) = (2/\pi e\delta)^{1/2}\exp[-R^2/(2e\delta)] \qquad (14a)$$

and

$$P(R_p) = (2/\pi\delta_p)^{1/2}\exp[-R_p^2/(2\delta_p)], \qquad (14b)$$

respectively. If δ, δ_p, e are equal to unity, then (13) and (14) reduce to (9) and (10), respectively. The resolution shells for which $\delta > 1$ or $\delta < 1$ correspond to positive or to negative values of the interference term.

The distributions (13*a*) and (14*a*) are plotted in Figs. 1 and 2 for the acentric and for the centric case, respectively, for some selected values of δ. Such distributions modify the classical Wilson distributions in order to take into account (i) the experimental shift of $\langle R^2\rangle$ from unity; (ii) an exceptional number of large or small normalized amplitudes, as frequently occurs when pseudo-symmetries are present.

The question is now, how to fix the δ value for which (13) and (14) fit the experimental distributions? The estimate of δ for a local experimental distribution may be obtained just by calculating the marginal moments of (13) and (14), according to the general integration formulas

$$\int_0^\infty x^{2n}\exp(-px^2)\,\mathrm{d}x = \frac{(2n-1)!!}{2(2p)^n}(\pi/p)^{1/2} \quad \text{and}$$
$$\int_0^{+\infty} x^{2n+1}\exp(-px^2)\,\mathrm{d}x = \frac{n!}{2p^{n+1}}. \qquad (15)$$

By considering the low-order moments (up to the order six) of the experimental amplitude distribution we obtain (subscripts $\overline{1}$ and 1 indicate that averages are performed for centric and acentric crystals, respectively)

$$\langle R\rangle_1 = \frac{\pi^{1/2}}{2}(e\delta)^{1/2}, \quad \langle R^2\rangle_1 = e\delta, \quad \langle R^3\rangle_1 = \frac{3}{4}\pi^{1/2}(e\delta)^{3/2},$$
$$\langle R^4\rangle_1 = 2(e\delta)^2, \quad \langle R^5\rangle_1 = \frac{15}{8}\pi^{1/2}(e\delta)^{5/2}, \quad \langle R^6\rangle_1 = 6(e\delta)^3, \qquad (16)$$

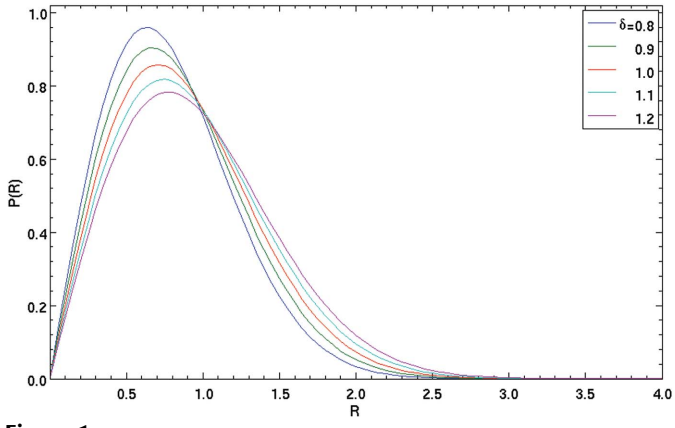

Figure 1
Quasi-Wilson distributions for acentric crystals for different values of δ: they take into account local $\langle R^2\rangle$ oscillations generated by crystal chemistry. The curve for δ = 1 corresponds to the acentric Wilson distribution.

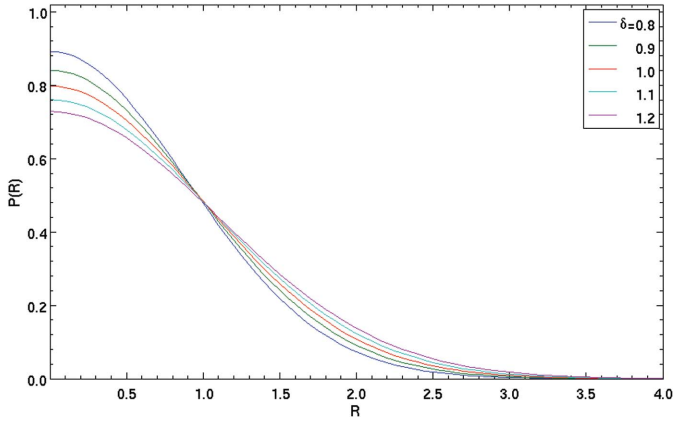

Figure 2
Quasi-Wilson distributions for centric crystals for different values of δ: they take into account local $\langle R^2\rangle$ oscillations generated by crystal chemistry. The curve for δ = 1 corresponds to the centric Wilson distribution.

and

$$\langle R\rangle_{\overline{1}} = (2/\pi)^{1/2}(e\delta)^{1/2}, \quad \langle R^2\rangle_{\overline{1}} = e\delta, \quad \langle R^3\rangle_{\overline{1}} = \frac{2^{3/2}}{\pi^{1/2}}(e\delta)^{3/2},$$
$$\langle R^4\rangle_{\overline{1}} = 3(e\delta)^2, \quad \langle R^5\rangle_{\overline{1}} = 2^{7/2}(e\delta)^{5/2}\pi^{-1/2}, \quad \langle R^6\rangle_{\overline{1}} = 15(e\delta)^3. \qquad (17)$$

The moments of the model amplitude distribution (not written here for the sake of simplicity) may be obtained by replacing in (16) and (17) the value $e\delta$ by δ_p.

6. The σ_A estimate *via* quasi-Wilson distributions

Let us suppose that, per resolution shell, we have used the experimental distributions $P(R)$ and $P(R_p)$ given by (13) and (14), and that for each shell we have estimated δ and δ_p according to one of the relationships (16) and (17). Then the local joint probability distribution $P(E, E_p)$ may be studied by first calculating its characteristic function C, and then calculating its Fourier transform. For the acentric case we obtain

$$C(u, u_p, v, v_p) = \exp\left\{-\frac{1}{4}\left[\delta e(u^2+v^2) + \delta_p(u_p^2+v_p^2) + 2(\delta\delta_p)^{1/2}\sigma_A(uu_p+vv_p)\right]\right\}$$

and for the centric case we obtain

$$C(u, u_p) = \exp\left[-\frac{1}{2}\delta e u^2 - \frac{1}{2}\delta_p u_p^2 - (\delta\delta_p)^{1/2}\sigma_A u u_p\right].$$

In the acentric case u, u_p, v, v_p are carrying variables associated with A, A_p, B, B_p, respectively; in the centric case u, u_p are carrying variables associated with E, E_p. By Fourier transform we derived for the acentric case

$$P(R, R_p, \varphi, \varphi_p) = RR_p\frac{\pi^{-2}}{\delta\delta_p(e-\sigma_A^2)}\exp\left\{-\frac{1}{\delta\delta_p(e-\sigma_A^2)} \times\left[\delta_p R^2 + e\delta R_p^2 - 2(\delta\delta_p)^{1/2}\sigma_A RR_p\cos(\varphi-\varphi_p)\right]\right\},$$

from which

$$\langle R^2R_p^2\rangle = \delta\delta_p(e+\sigma_A^2)$$

and

$$\sigma_A^2 = \left(\frac{\langle R^2R_p^2\rangle}{\delta\delta_p} - e\right). \tag{18}$$

The conditional distribution $P(\varphi|R, R_p, \varphi_p)$ is then given by

$$P(\varphi|R, R_p, \varphi_p) = [2\pi I_0(X)]^{-1}\exp\left[X\cos(\varphi-\varphi_p)\right],$$

where

$$X = 2\frac{\sigma_A}{(e-\sigma_A^2)}\frac{RR_p}{(\delta\delta_p)^{1/2}}. \tag{19}$$

For the centric case we obtain

$$P(E, E_p) = \frac{1}{(2\pi)[\delta\delta_p(e-\sigma_A^2)]^{1/2}}\exp\left\{-\frac{1}{2\delta\delta_p(e-\sigma_A^2)} \times\left[e\delta E_p^2 + \delta_p E^2 - 2(\delta\delta_p)^{1/2}\sigma_A EE_p\right]\right\}$$

from which

$$\langle R^2R_p^2\rangle = \delta\delta_p\left(e+2\sigma_A^2\right)$$

and

$$\sigma_A^2 = \frac{1}{2}\left(\frac{\langle R^2R_p^2\rangle}{\delta\delta_p} - e\right). \tag{20}$$

Additional calculations show that the phase indication $\varphi \simeq \varphi_p$ will then depend on the value of

$$\tanh\frac{\sigma_A}{(e-\sigma_A^2)}\frac{RR_p}{(\delta\delta_p)^{1/2}}. \tag{21}$$

Both equations (18) and (20) satisfy the asymptotic σ_A features: when R and R_p are uncorrelated, σ_A vanishes; when the partial and the target structures coincide, σ_A attains unity.

In order to apply equations (18)–(20), prior estimates of δ_p and δ are needed. In accordance with (16) and (17) we may estimate them *via* moments of different order: then different formulas arise according to the chosen moment order. It is worthwhile stressing that modeling a distribution *via* equations (13) or (14) by using the δ value suggested by a moment of a given order does not allow all the features of the experimental data to be captured, even if the chosen δ value allows the experiment to fit better than the corresponding Wilson distribution. As a numerical example, according to (16) the value of δ to use for modeling the experimental distribution may be obtained both as $\delta = \langle R^2\rangle_1$ or as $\delta = (\langle R^4\rangle_1/2)^{1/2}$. In practice, these two values seldom coincide [for example, because the percentage of measured reflections with very large amplitude does not coincide with that foreseen by (13*a*)].

The above theoretical results enable us to suggest the following general expression for the σ_A estimate,

$$\sigma_A^2 = qe\left(\frac{\langle R^2R_p^2\rangle}{\langle R^m\rangle^{2/m}\langle R_p^m\rangle^{2/m}}\, w_m^{2/m}\, w_{pm}^{2/m} - 1\right), \tag{22}$$

where $q = 1$ or 0.5 according to whether the crystal is acentric or centric; $\langle R^m\rangle$ and $\langle R_p^m\rangle$ are the experimental m-order moments for the target and the model structure, respectively; $w_m = \langle R^m\rangle_W$ and $w_{pm} = \langle R_p^m\rangle_W$ are the m-order moment values according to Wilson distributions;

$$w_m = \pi^{1/2}/2,\ 1,\ (3/4)\pi^{1/2},\ 2,\ (15/8)\pi^{1/2},\ 6$$

for $m = 1, 2, \ldots, 6$ for acentric crystals; and

$$w_m = (2/\pi)^{1/2},\ 1,\ 2^{3/2}/\pi^{1/2},\ 3,\ 2^{7/2}/\pi^{1/2},\ 15$$

for $m = 1, 2, \ldots, 6$ for centric crystals.

Equation (22) encompasses previous formulas for the σ_A estimation: in particular, when $m = 2$ is selected, it justifies the practice of using the locally renormalized structure factors [see equation (6)]. Furthermore, equations (19) and (21) suggest that the values of δ_p, δ and e should also be taken into account to estimate the phase reliability. This by no means implies that locally renormalized structure-factor moduli are the best coefficients for the calculation of the observed electron-density maps. Indeed, local renormalization strongly reduces the correlation between structure-factor moduli and dominant interatomic distances in the structure. Observed amplitudes, normalized according to the Wilson plot, and phases estimated according to (19) and (21) should be preferable.

7. About the σ_A estimate from joint moment $\langle R^\mu R_p^\mu\rangle$

So far, σ_A estimates have been derived *via* the use of $\langle R^2R_p^2\rangle$: obviously, any other joint moment may be employed for the same purpose. First we focus our attention on the joint moment $\langle RR_p\rangle$. For acentric crystals, under the hypothesis that the experimental amplitude distributions for the target and the model structures satisfy Wilson statistics, the following relation holds (Caliandro *et al.*, 2005),

$$\langle RR_p\rangle = \frac{\pi}{4}e^{1/2}\,F\left(\frac{-1}{2}, \frac{-1}{2}; 1; \frac{\sigma_A^2}{e}\right), \tag{23}$$

where F is the Gaussian hypergeometric function. For application purposes $F(-1/2, -1/2; 1; \sigma_A^2/e)$ was numerically approximated by the function $[1 + (\pi/12)(\sigma_A^2/e)]$, so establishing the following relation,

$$\langle RR_p\rangle = \frac{\pi}{4}e^{1/2}\left[1 + \frac{\pi}{12}\frac{\sigma_A^2}{e}\right]. \quad (24)$$

Even if (22) and (23) are very close to each other when the argument of the hypergeometric function varies in the range (0, 1), equation (24) does not perfectly satisfy the asymptotic σ_A properties. Indeed, when R and R_p are uncorrelated, σ_A vanishes, as it should; but if the partial and the target structures coincide, $\sigma_A = (12/\pi)[(4/\pi) - 1] = 0.82$, instead of attaining unity. We suggest the better approximation

$$\langle RR_p\rangle = \frac{\pi}{4}e^{1/2}\left[1 + \left(\frac{4}{\pi} - 1\right)\frac{\sigma_A^2}{e}\right], \quad (25)$$

which perfectly satisfies the required asymptotic σ_A properties [in practice, the new approximation replaces $\pi/12 = 0.262$ by $(4/\pi) - 1 = 0.273$]. According to (25), $\langle RR_p\rangle$ is expected to lie in the range (0, 1).

In the case where the experimental amplitude distributions for the target and the model structures satisfy quasi-Wilson statistics, we obtain the following relationship,

$$\sigma_A^2 = \frac{\pi e}{4 - \pi}\left[\frac{\langle RR_p\rangle}{\delta^{1/2}\delta_p^{1/2}}\frac{4}{\pi e^{1/2}} - 1\right].$$

As for (18), the values of δ and δ_p may be estimated *via* any of the equations (16): the corresponding formulas may then be written down as

$$\sigma_A^2 = \frac{\pi e}{4 - \pi}\left(\frac{\langle RR_p\rangle}{\langle R^m\rangle^{1/m}\langle R_p^m\rangle^{1/m}}\frac{w_m^{1/m}w_{pm}^{1/m}}{w_1 w_{p1}} - 1\right). \quad (26)$$

Let us now consider the use of the general moment $\langle R^\mu R_p^\mu\rangle$, where μ is an even number. By definition,

$$\begin{aligned}\langle R^\mu R_p^\mu\rangle = {} & \frac{4}{\delta\delta_p(e - \sigma_A^2)}\int_0^\infty R_p^{\mu+1}\exp\left[-\frac{e}{\delta_p(e - \sigma_A^2)}R_p^2\right]\mathrm{d}R_p \\ & \times \int_0^\infty R^{\mu+1}\exp\left[-\frac{1}{\delta(e - \sigma_A^2)}R^2\right] \\ & \times I_0\left[\frac{2\sigma_A}{(\delta\delta_p)^{1/2}(e - \sigma_A^2)}RR_p\right]\mathrm{d}R.\end{aligned}$$

After some calculations we obtain

$$\begin{aligned}\langle R^\mu R_p^\mu\rangle = {} & 2\Gamma\left(\frac{\mu}{2} + 1\right)\frac{\delta^{\mu/2}}{\delta_p}(e - \sigma_A^2)^{\mu/2} \\ & \times \int_0^\infty R_p^{\mu+1}\exp\left(-R_p^2/\delta_p\right)L_{\mu/2}\left[-\frac{\sigma_A^2}{\delta_p(e - \sigma_A^2)}{R_p}^2\right]\mathrm{d}R_p,\end{aligned} \quad (27)$$

where $L_{\mu/2}$ are Laguerre polynomials of order μ in R_p.

For $\mu = 2$, equation (27) reduces to [to be compared with equation (11)]

$$\langle R^2R_p^2\rangle = \delta(e - \sigma_A^2)\langle R_p^2\rangle + \sigma_A^2(\delta/\delta_p)\langle R_p^4\rangle,$$

showing, for a quasi-Wilson distribution, the relation between $\langle R^2R_p^2\rangle$ (and therefore σ_A) and the fourth-order moments.

The integration on the right-hand side of (27) may be accomplished by using the second of the equations (15): we obtain

$$\langle R^4R_p^4\rangle = 4\delta^2\delta_p^2\left(e^2 + 4e\sigma_A^2 + \sigma_A^4\right).$$

Choosing δ and δ_p according to (16) leads to the following general formula,

$$\frac{\langle R^4R_p^4\rangle}{\langle R^m\rangle^{4/m}\langle R_p^m\rangle^{4/m}}\frac{w_m^{4/m}w_{pm}^{4/m}}{w_4 w_{p4}} - 1 = \frac{1}{e^2}\left(4e\sigma_A^2 + \sigma_A^4\right). \quad (28)$$

Equation (28) may be easily solved with respect to σ_A^2 per resolution shell. For higher (and even) μ values an equation similar to (28) is expected, with, at the left-hand side, the term

$$\frac{\langle R^\mu R_p^\mu\rangle}{\langle R^m\rangle^{\mu/m}\langle R_p^m\rangle^{\mu/m}}\frac{w_m^{\mu/m}w_{pm}^{\mu/m}}{w_\mu w_{p\mu}} - 1,$$

and a suitable polynomial arising from the integral of $L_{\mu/2}$ on the right-hand side. However, using the moment $\langle R^\mu R_p^\mu\rangle$ involves powers up to order $2\mu + 1$, or, equivalently, moments up to order 2μ. The use of μ values larger than four is discouraged, because the moments of order eight or larger are too sensitive to small anomalies in the normalized structure-factor distributions.

Finally we notice that, whatever the value of μ, it is expected that σ_A decreases with increasing s values. To guarantee a soft trend, a least-squares straight line (say, LS σ_A) should be calculated for any pair (μ, m).

8. Applications

The following applications aim at checking the correctness of the theory described in the preceding sections. We applied it to eight proteins: the model electron-density maps were obtained by molecular replacement *via* the program *REMO09* (Caliandro *et al.*, 2009). We only show the results corresponding to two extreme cases: in the first, with Protein Data Bank (PDB) code 1xyg (Center for Eukaryotic Structural Genomics, 2011), the model and the observed normalized moduli satisfy Wilson statistics well; in the second, with PDB code 1lat (Gewirth & Sigler, 1995), the observed and model distributions are very far from the ideal ones. The other six test structures have intermediate features and are not mentioned further here.

In Table 1 the column 'Code' defines the PDB codes of the two test proteins and 'RES' is the experimental data resolution. In accordance with the theoretical results of §7 we explore the potential usefulness of equations (22), (26) and (28). In Fig. 3 we show, for 1xyg, LS σ_A *versus* s for $m = 1, \ldots, 6$ according to (22): the LS σ_A corresponding to a given m value

Table 1
For each test structure RES is the data resolution (Å) and $C(\mu, m)$ is the correlation between the published electron-density map and the model map calculated by using weights according to equation (19), where σ_A is estimated *via* different values of μ and m.

Code	RES	μ	$C(\mu, 1)$	$C(\mu, 2)$	$C(\mu, 3)$	$C(\mu, 4)$	$C(\mu, 5)$	$C(\mu, 6)$
1xyg	2.19	1	0.63	0.62	0.58	0.42	–	–
		2	0.63	0.63	0..62	0.61	0.59	0.55
		4	0.63	0.63	0.62	0.62	0.61	0.60
1lat	1.90	1	0.53	0.14	–	–	–	–
		2	0.52	0.52	0.42	–	–	–
		4	0.50	0.52	0.51	0.50	0.48	0.47

lies below that corresponding to $m - 1$, but preserves the same negative slope. In Fig. 4 we show LS σ_A for 1lat: they have different slopes, and for $m > 3$ the σ_A values constantly go to zero or become negative for most of the resolution shells (in these cases LS σ_A are not calculated).

In Fig. 5 we plot, for 1xyg and for $m = 1, \ldots 4$, LS σ_A values corresponding to (26): again LS σ_A preserve the same negative slope, but for $m > 4$ they are not calculated because σ_A is negative for most of the resolution shells. For 1lat only the LS σ_A corresponding to $m = 1$ may be calculated, but for simplicity it is not shown here.

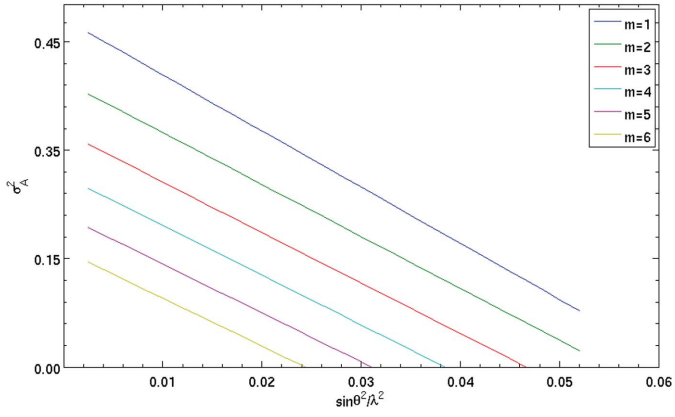

Figure 3
1xyg: LS σ_A *versus* s for $m = 1, \ldots, 6$ according to equation (22).

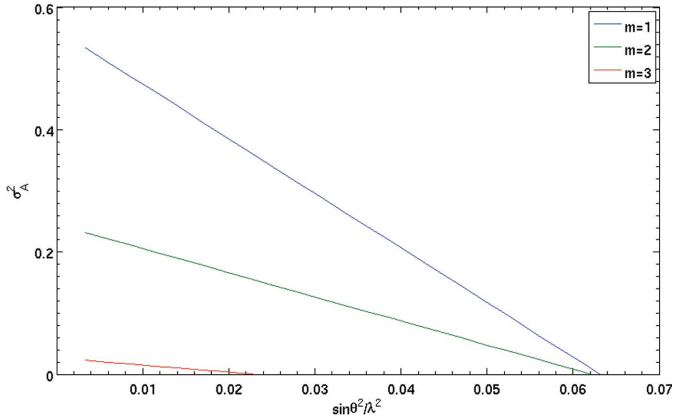

Figure 4
1lat: LS σ_A *versus* s for $m = 1, \ldots, 3$ according to equation (22).

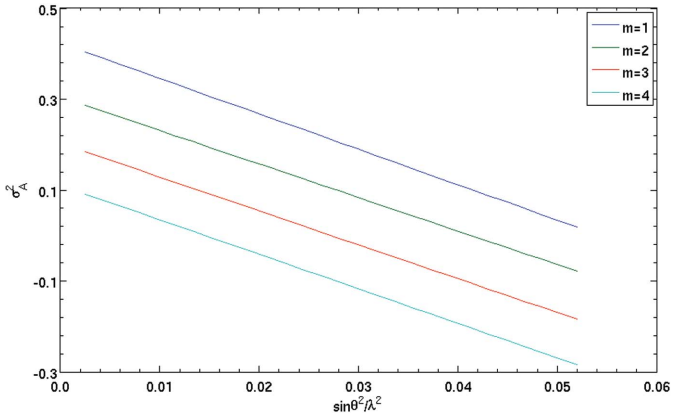

Figure 5
1xyg: LS σ_A *versus* s for $m = 1, \ldots, 4$ according to equation (26).

In Figs. 6 and 7, LS σ_A values corresponding to (28) for $m = 1, \ldots, 4$ for 1xyg and 1lat, respectively, are plotted: again LS σ_A preserve the same negative slope for 1xyg and have different slopes for 1lat.

The basic reasons for the above features are the following. The σ_A values depend on the value of the joint moment $\langle R^{\mu} R_p^{\mu} \rangle$ and on the values of the parameters δ and δ_p. These

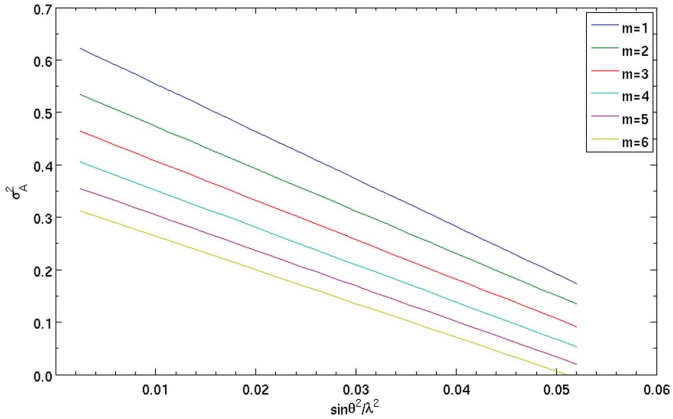

Figure 6
1xyg: LS σ_A *versus* s for $m = 1, \ldots, 6$ according to equation (28).

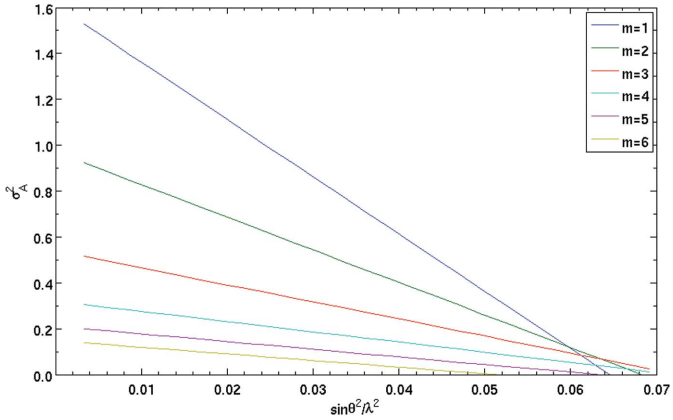

Figure 7
1lat: LS σ_A *versus* s for $m = 1, \ldots, 6$ according to equation (28).

Table 2
1lat: marginal moments for $m = 1, \ldots, 6$ of the observed normalized amplitudes for different resolution shells; n in column 1 is the order of the shell (total number 25) and d (Å) is the corresponding resolution.

n	d	$\langle R \rangle$	$\langle R^2 \rangle$	$\langle R^3 \rangle$	$\langle R^4 \rangle$	$\langle R^5 \rangle$	$\langle R^6 \rangle$
1	8.62	0.74	0.72	0.81	1.02	1.41	2.16
2	6.52	0.64	0.55	0.56	0.66	0.85	1.19
7	3.57	1.01	1.37	2.25	4.24	8.86	19.96
8	3.35	1.04	1.56	3.22	8.90	31.05	126.77
9	3.16	1.01	1.99	7.95	49.74	382.39	3258.13
10	3.00	0.86	1.18	2.52	7.74	29.67	128.40
24	1.94	0.93	1.07	1.53	2.75	6.06	15.95
25	1.90	0.98	1.21	2.04	4.98	17.38	77.53

two parameters may be estimated *via* marginal moments of different order m. Since such moments are, for the two test structures, progressively larger than those foreseen by Wilson statistics, the σ_A estimate progressively decreases with m. The decrement is regular for 1xyg (*i.e.* for all the resolution shells a similar decrement is observed), and rather wild for 1lat. We show in Table 2, for $m = 1, \ldots, 6$, the marginal moments of 1lat for some resolution shells (for Wilson distributions the expected values in the order are about 0.886, 1, 1.329, 2.00, 3.323, 6.00).

The above results are not unexpected: indeed a structure-factor amplitude distribution is defined if all its moments are known; a single moment cannot capture all the features of the distribution. As a consequence the σ_A estimates will be distribution dependent, and will vary according to the moment order we choose for the estimate: variations of the local amplitude distributions will influence the σ_A estimates according to the values of μ and m, and will lead to displaced LS σ_A or to different LS σ_A slopes according to circumstances.

To check the effects of the different σ_A estimates on the electron-density maps, in Table 1 we show the correlation factors $C(\mu, m)$ between electron-density maps calculated *via* the pair (R, φ) (φ are the published phases) and the maps calculated *via* the pair (R, φ_p), by using $\mu = 1$, 2, 4 and $m = 1, 2, \ldots, 6$. Favorable values of $C(\mu, m)$ are always obtained for $\mu = m$, no matter the value of μ.

9. Conclusions

A theoretical study of the statistical properties of σ_A has been accomplished. Quasi-Wilson distributions have been introduced to describe the local deviations of the structure-factor moduli from Wilson statistics. In this way a theoretical justification for the local renormalization of the structure factors has been provided. A wider class of new probabilistic formulas has been proposed to estimate the σ_A parameter: such new tools have been applied to experimental cases and their mean statistical features have been studied.

References

Burla, M. C., Caliandro, R., Giacovazzo, C. & Polidori, G. (2010). *Acta Cryst.* A**66**, 347–361.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2010). *J. Appl. Cryst.* **43**, 825–836.
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., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2009). *Acta Cryst.* A**65**, 512–527.
Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992). *Z. Kristallogr.* **200**, 63–71.
Center for Eukaryotic Structural Genomics (2011). In preparation.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Debye, P. (1915). *Ann. Phys. (Leipzig)*, **46**, 809–823.
Gewirth, D. T. & Sigler, P. B. (1995). *Nat. Struct. Biol.* **2**, 386–394.
Hall, S. R. & Subramanian, V. (1982*a*). *Acta Cryst.* A**38**, 590–598.
Hall, S. R. & Subramanian, V. (1982*b*). *Acta Cryst.* A**38**, 598–608.
Lunin, V. Yu. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.
Morris, R. J., Blanc, E. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 227–240.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.