Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 14 November 2002
Accepted 7 March 2003

The estimate of the protein phases for the SAD case: a simplified formula

Carmelo Giacovazzo,[a,b]* Massimo Ladisa[b] and Dritan Siliqi[b,c]

[a]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Istituto di Cristallografia, CNR, c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and [c]Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana, Albania. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

The probabilistic formula derived *via* the rigorous method of the joint probability distribution function to estimate protein phases in the single-wavelength anomalous diffraction (SAD) case [Giacovazzo & Siliqi (2001). *Acta Cryst.* A**57**, 40–46] has been revised. A simple but equally effective formula is provided, allowing an easy interpretation of the role of the structural parameters accessible *via* a diffraction experiment. In particular, the formula is able to simultaneously combine the contribution arising from the anomalous differences with a Sim-like contribution, and also to take the errors into account.

1. Notation

N: number of atoms in the unit cell.
a: number of anomalous scatterers in the unit cell.
$na = N - a$: number of non-anomalous scatterers in the unit cell.
Z: atomic number.
$f_j = f_j^0 + \Delta f_j + if_j'' = f_j' + if_j''$: scattering factor of the jth atom (thermal factor included).
$F_a^+ = |F_a^+| \exp(i\varphi_a^+) = \sum_a f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j)$.
$F_a^- = |F_a^-| \exp(i\varphi_a^-) = \sum_a f_j \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}_j)$.
$F^+ = |F^+| \exp(i\varphi^+) = F_h = F_a^+ + F_{na}^+ + \mu^+$.
$F^- = |F^-| \exp(i\varphi^-) = F_{-h} = F_a^- + F_{na}^- + \mu^-$.
μ^+ and μ^- represent the cumulative errors arising from different sources (*i.e.* the structural model constituted by the located anomalous scatterers and errors in measurements).
$|F| \exp(i\varphi)$: structure factor calculated by taking into account non-anomalous scattering (all the atoms in the unit cell included).
$\sum_a, \sum_{na}, \sum_N = \sum (f_j'^2 + f_j''^2)$, where the summation is extended to a, na and N atoms.
v_a, v_N: $\sum Z_j^2$, where the summation is extended to the anomalous scatterer substructure or to all the atoms in the unit cell.
$\Delta_{\rm ano} = |F^+| - |F^-|$.

The paper by Giacovazzo & Siliqi (2001*a*) will be denoted paper I.

2. Introduction

Owing to recent advances in synchrotron technology, the MAD (multiwavelength anomalous diffraction) techniques become more and more popular as a tool for solving protein structures. In recent years, SAD (single-wavelength anomalous diffraction) techniques have started to play a major role, particularly when the data collection is very accurate (*e.g.* high redundancy of measures). As soon as the anomalous scatterers are located (Weeks & Miller, 1999; Terwilliger & Berendzen, 1999; Grosse-Kunstleve & Brunger, 1999; Burla *et al.*, 2002; Schneider & Sheldrick, 2002), probabilistic formulas are applied to estimate the protein phases.

In paper I, the method of the joint probability distribution functions has been applied to derive a probabilistic formula for the SAD case. The final distributions (I.11) and (I.13) provided the conditional probability distributions $P(\varphi^+|R, G, E_a^+, E_a^-)$ and $P(\varphi^-|R, G, E_a^+, E_a^-)$, respectively, where

$$E^+ = R\exp(i\varphi^+), \qquad E^- = G\exp(i\varphi^-),$$
$$E_a^+ = R_a\exp(i\varphi_a^+), \qquad E_a^- = G_a\exp(i\varphi_a^-)$$

are pseudo-normalized structure factors (*i.e.* normalized with respect to the non-anomalous scatterer substructure); *e.g.* $E^+ = F^+/\left(\sum_{na}\right)^{1/2}$. The above distributions are quite general and have been rigorously derived but suffer from some drawbacks:

(i) the estimation of φ^+ requires two (time-consuming) numerical integrations: *i.e.* the calculation of the ratio

$$\int_{-\pi}^{+\pi} \varphi^+ P(\varphi^+|R, G, E_a^+, E_a^-)\,{\rm d}\varphi^+ \Big/ \int_{-\pi}^{+\pi} P(\varphi^+|R, G, E_a^+, E_a^-)\,{\rm d}\varphi^+;$$

(ii) the same calculations are necessary for the estimation of φ^-;

(iii) the variance of the estimates (information necessary for practical applications) requires analogous numerical integrals;

(iv) last, but not least, the complicated algebraic expressions of (I.11) and (I.13) do not allow the reader to understand the

nature of the terms providing the phase information and their interpretation in terms of parameters accessible *via* the diffraction experiment. Thus the connexion of (I.11) and (I.13) with other approaches described in the literature is difficult.

This paper aims at overcoming the above drawbacks by deriving probabilistic conditional distributions for φ^+ and for φ^- easily understandable and of immediate use.

We will focus our attention on the joint probability distribution (I.8), from which (see Appendix *A*) the following conditional distribution may be derived:

$$\begin{aligned} P(\varphi^+|R, G, E_a^+, E_a^-) &\approx S\exp\{-(2/e)[R_a(G - e^-R)\cos(\varphi^+ - \varphi_a^+) \\ &\quad + G_a(R - e^+G)\cos(\varphi^+ + \varphi_a^-)]\}, \end{aligned} \tag{1}$$

where S is a suitable scale factor,

$$e^+ = 1 + \frac{\langle|\mu^+|^2\rangle}{\sum_{na}}, \quad e^- = 1 + \frac{\langle|\mu^-|^2\rangle}{\sum_{na}}, \tag{2}$$

$$e = (e^+e^- - 1) = \frac{\langle|\mu^+|^2\rangle}{\sum_{na}} + \frac{|\langle|\mu^-|^2\rangle}{\sum_{na}} + \frac{\langle|\mu^+|^2\rangle\langle|\mu^-|^2\rangle}{\left(\sum_{na}\right)^2}. \tag{3}$$

Since $\langle|\mu^+|^2\rangle$ and $\langle|\mu^-|^2\rangle$ are usually negligible with respect to $\sum_{na}$, we can approximate (3) as

$$e = (\langle|\mu^+|^2\rangle + \langle|\mu^-|^2\rangle)/\textstyle\sum_{na}. \tag{4}$$

Accordingly, (1) reduces to

$$\begin{aligned} P(\varphi^+|R, G, E_a^+, E_a^-) &\approx S\exp\{2q^+R_aR\cos(\varphi^+ - \varphi_a^+) + 2q^-G_aG\cos(\varphi^+ + \varphi_a^-) \\ &\quad + 2(R - G)[R_a\cos(\varphi^+ - \varphi_a^+) - G_a\cos(\varphi^+ + \varphi_a^-)]/e\}, \end{aligned} \tag{5}$$

where

$$q^+ = \langle|\mu^-|^2\rangle/(\langle|\mu^+|^2\rangle + \langle|\mu^-|^2\rangle)$$

and

$$q^- = \langle|\mu^+|^2\rangle/(\langle|\mu^+|^2\rangle + \langle|\mu^-|^2\rangle).$$

Factorizing the terms containing $\cos(\varphi^+ - \varphi_a^+)$ and the terms containing $\cos(\varphi^+ + \varphi_a^-)$ gives

$$P(\varphi^+|R, G, E_a^+, E_a^-) \approx [2\pi I_o(X)]^{-1}\exp\{X\cos(\varphi^+ - \theta^+)\}, \tag{6}$$

where

$$\tan\theta^+ = P/Q, \tag{7}$$

$$\begin{aligned} P &= 2(q^+RR_a\sin\varphi_a^+ - q^-GG_a\sin\varphi_a^-) \\ &\quad + 2[(R - G)/e][R_a\sin\varphi_a^+ + G_a\sin\varphi_a^-], \\ Q &= 2(q^+RR_a\cos\varphi_a^+ + q^-GG_a\cos\varphi_a^-) \\ &\quad + 2[(R - G)/e][R_a\cos\varphi_a^+ - G_a\cos\varphi_a^-], \\ X &= (P^2 + Q^2)^{1/2}, \end{aligned} \tag{8}$$

and e is given by (4).

Let us now characterize the nature of the terms in P and Q. Both P and Q are constituted by two contributors: the first is a Sim-like term (Sim, 1959, 1960), the second depends on the $\Delta_{\rm ano}$ experimental measurements.

The non-Sim terms in the P and Q expressions, say

$$2(R - G)[R_a\sin\varphi_a^+ + G_a\sin(\pi - \varphi_a^-)]/e$$

and

$$2(R - G)[R_a\cos\varphi_a^+ + G_a\cos(\pi - \varphi_a^-)]/e,$$

are, in the Argand plane, nothing else but the components of the vector

$$2(R - G)\{R_a\exp(i\varphi_a^+) + G_\alpha\exp[i(\pi - \varphi_a^-)]\}/e.$$

Let us now denote by E_a^{-*} the complex conjugate of E_a^-. Then,

$$\begin{aligned} E_a^+ - E_a^{-*} &= R_a\exp(i\varphi_a^+) - G_a\exp(-i\varphi_a^-) \\ &= R_a\exp(i\varphi_a^+) + G_a\exp[i(\pi - \varphi_a^-)] \\ &= 2iE_a''^+, \end{aligned}$$

where

$$\begin{aligned} iE_a''^+ &= i\left(\textstyle\sum_{na}\right)^{-1/2}\sum_j f_j''\exp(2\pi i\mathbf{h}\cdot\mathbf{r}_j) \\ &= iR_a''\exp(i\varphi_a''^+) \\ &= R_a''\exp[i(\varphi_a'' + \pi/2)]. \end{aligned}$$

Accordingly, P and Q may be rewritten as

$$\begin{aligned} P &= 2(q^+RR_a\sin\varphi_a^+ - q^-GG_a\sin\varphi_a^-) \\ &\quad + 4[(R - G)/e]R_a''\sin(\varphi_a''^+ + \pi/2), \end{aligned} \tag{9}$$

$$\begin{aligned} Q &= 2(q^+RR_a\cos\varphi_a^+ + q^-GG_a\cos\varphi_a^-) \\ &\quad + 4[(R - G)/e]R_a''\cos(\varphi_a'' + \pi/2). \end{aligned} \tag{10}$$

For sufficiently large proteins, the Sim contribution may be negligible (this occurs when the scattering power of the anomalous scatterers is negligible with respect to the total scattering power of the protein). Then P and Q reduce to

$$P = 4[(R - G)/e]R_a''\sin(\varphi_a'' + \pi/2), \tag{11}$$

$$Q = 4[(R - G)/e]R_a''\cos(\varphi_a'' + \pi/2). \tag{12}$$

In such a situation,

$$\theta^+ \approx \begin{cases} \varphi_a''^+ + \pi/2 & \text{if } (R - G) > 0, \\ \varphi_a''^+ - \pi/2 & \text{if } (R - G) < 0 \end{cases}$$

and

$$X = 4(R - G)R_a''/e. \tag{13}$$

X may also be written as

$$X = \frac{(|F^+| - |F^-|)|F_a''^+|}{\langle|\mu^+|^2\rangle + \langle|\mu^-|^2\rangle} = \frac{\Delta_{\rm ano}|F_a''^+|}{\langle|\mu^+|^2\rangle + \langle|\mu^-|^2\rangle}. \tag{14}$$

In the next section, we will apply (11) and (12) to a case in which they constitute a useful approximation of (9) and (10). However, a second application will be described for which the partial information exploited by (11) and (12) is not sufficient for obtaining good phase estimates. As a conclusion, from our tests the use of (9) and (10) is always advised.

Table 1
JIA experimental data: cumulative average phase error ($\Delta\varphi^\circ$) for the estimates provided by equations (I.11) and (I.13) and (9) and (10) (the weighted phase error is in parentheses); NREF is the number of phase estimates with reliability factor larger than w.

w	NREF	(I.11)–(I.13) $\Delta\varphi^\circ$	(9)–(10) $\Delta\varphi^\circ$
0.1	26142	59 (53)	60 (52)
0.2	23104	56 (52)	57 (51)
0.3	20569	54 (51)	53 (50)
0.4	18424	52 (50)	51 (48)
0.5	16464	50 (49)	48 (46)
0.6	14619	49 (48)	45 (44)
0.7	12677	47 (47)	43 (42)
0.8	10114	46 (46)	40 (40)
0.9	6277	44 (44)	38 (38)

Table 2
CAUFD experimental data: cumulative average phase error ($\Delta\varphi^\circ$) for the estimates provided by equations (I.11) and (I.13), (9) and (10) and (11) and (12) (the weighted phase error is in parentheses); NREF is the number of phase estimates with reliability factor larger than w.

w	NREF	(I.11)–(I.13) $\Delta\varphi^\circ$	(9)–(10) $\Delta\varphi^\circ$	(11)–(12) $\Delta\varphi^\circ$
0.1	20556	45 (37)	45 (33)	62 (57)
0.2	14640	40 (35)	32 (28)	56 (53)
0.3	10135	35 (33)	27 (26)	52 (50)
0.4	6943	32 (30)	24 (23)	49 (48)
0.5	4530	28 (28)	22 (22)	47 (46)
0.6	2789	26 (26)	21 (21)	45 (45)
0.7	1479	25 (25)	21 (21)	42 (42)
0.8	496	24 (24)	22 (22)	39 (39)
0.9	84	25 (25)	33 (33)	39 (39)

3. Applications

We first check whether the simple formulas (9) and (10) lose information with respect to the use of the complicated conditional distributions (I.11) and (I.13). We apply them to the following experimental data.

(*a*) JIA (Li *et al.*, 2000), space group $C222_1$, $a = 95.90$, $b = 119.80$, $c = 165.48$ Å, 570 residues and 8 Se atoms in the asymmetric unit. The data correspond to the wavelength $\lambda = 0.9793$ Å, where $\Delta f' \simeq -5.6$, $f'' \simeq 3.3$.

(*b*) A smaller protein, CAUFD (Dauter *et al.*, 1997), space group $P4_32_12$, $a = 33.95$, $c = 74.82$ Å, 94 residues. Eight Fe atoms (per asymmetric unit) are the anomalous scatterers, with $\Delta f' \simeq 0.26$ and $f'' \simeq 1.25$ at $\lambda = 0.88$ Å.

The results are shown in Tables 1 and 2. The phase estimates *via* (I.11) and (I.13) are obtained by calculating the centroids of the distributions and the corresponding variances (*via* the figure of merit $w = m$, see paper I). The estimates *via* (9) and (10) are ordered according to $w = D_1(X) = I_1(X)/I_0(X)$, where I_i is the modified Bessel function of order i. The efficiencies of the old and the new formulations are nearly equivalent. The correlation factors (CORR) between the electron-density maps calculated *via* the estimated phases and the published refined maps are the following:

(i) for JIA, CORR = 0.49 if (I.11) and (I.13) are used, 0.48 if the estimates are provided by (9) and (10);

(ii) for CAUFD, CORR = 0.66 both for (I.11) and (I.13) and for (9) and (10).

Let us now check if the efficiency of (11) and (12) is comparable with that shown by formulas (9) and (10). For JIA, the phase error (and the CORR value) does not change if (11) and (12) replace (9) and (10).

Apply now (11) and (12) and (9) and (10) to CAUFD (see Table 2). The necessity of using the full equations (9) and (10) is evident from comparison of the phase errors. The value of CORR is 0.66 when (9) and (10) are used, but is only 0.29 if the estimates are provided by (11) and (12). The role of the Sim component is therefore central for CAUFD, while it is negligible for JIA. This different feature is due to the different scattering powers of the anomalous scatterers in the two structures: indeed, $v_a/v_N = 0.05$ for JIA, $v_a/v_N = 0.15$ for CAUFD.

The phases obtained *via* (9) and (10) were submitted to an automatic solvent-flattening procedure (Giacovazzo & Siliqi, 1997): the resulting values of CORR were 0.83 for JIA and 0.84 for CAUFD.

We have also applied the widely used program *MLPHARE* from CCP4 (Collaborative Computational Project Number 4, 1994) to the experimental data of JIA and CAUFD. While for JIA the value of CORR is smaller but comparable with that obtained by (9) and (10) (0.47 against 0.49), the situation is much worse for CAUFD, for which CORR = 0.32. The lower efficiency of *MLPHARE* is because such a program is unable (as most of the current packages) to simultaneously use the SIM contribution.

4. Conclusions

A simplified probabilistic formula has been obtained that provides a simple tool for assigning phases in the SAD case, allows an easy interpretation of the phase distribution in terms of parameters accessible *via* the diffraction experiment, and eliminates the necessity of calculating the centroid of the phase distribution and the variance of the estimate *via* numerical methods. The new formula contains two contributors: a Sim-like term and a term arising from the measured anomalous differences. The first term is not straightforwardly used in the usual SAD procedures, but its potential role was not ignored in the literature. To give an example, let us consider the basic algebraic equation on which traditional SAD techniques are based:

$$\varphi = \varphi'' \pm \Delta\varphi, \tag{15}$$

where

$$\Delta\varphi = \cos^{-1}(\Delta_{\rm ano}/2|F''|).$$

In the Bijvoet–Ramachandran–Raman method (Ramachandran & Raman, 1956; Raman, 1959; Moncrief & Lipscomb, 1966), the ambiguity on the phase estimate provided by (15) is solved *a posteriori via* the additional use of the Sim contribution, provided the anomalous scatterers have a non-negligible influence. The approach has been used by Hendrickson

& Teeter (1981) to solve the crystal structure of crambin at 1.5 Å resolution. Our probabilistic equations (9) and (10) first state the most rigorous way of *simultaneously* combining the contribution arising from the anomalous differences with the Sim contribution, also taking the errors into account. The superior efficiency of the formulas has been checked by application to experimental data.

APPENDIX *A*

If all the anomalous atoms have been located, equation (I.8) may be written as follows:

$$\begin{aligned}P(R, G, \varphi^+, \varphi^- | E_a^+, E_a^-) &= [RG/(\pi^2 e^+ e^- c)] \\ &\times \exp\left\{-\frac{1}{c}\left[\frac{R^2 + R_a^2 - 2RR_a\cos(\varphi^+ - \varphi_a^+)}{e^+} + \frac{G^2 + G_a^2 - 2GG_a\cos(\varphi^- - \varphi_a^-)}{e^-}\right]\right. \\ &+ \frac{2c_3}{c}\frac{1}{(e^+e^-)^{1/2}}[RG\cos(\varphi^+ + \varphi^-) + R_aG_a\cos(\varphi_a^+ + \varphi_a^-) \\ &\left. - RG_a\cos(\varphi^+ + \varphi_a^-) - R_aG\cos(\varphi^- + \varphi_a^+)]\right\}, \qquad (16)\end{aligned}$$

where

$$\begin{aligned}c^2 &= [1 - (c_1^2 + c_2^2)]^2, \quad c_1 = c_1'(e^+e^-)^{-1/2}, \\ c_2 &= c_2'(e^+e^-)^{-1/2}, \quad c_3^2 = c_1^2 + c_2^2, \\ c_1' &= \textstyle\sum_{na}(f_j'^2 - f_j''^2)\Big/\sum_{na}(f_j'^2 + f_j''^2), \\ c_2' &= \big(2\textstyle\sum_{na} f_j' f_j''\big)\Big/\sum_{na}(f_j'^2 + f_j''^2).\end{aligned}$$

Let us now introduce in (16) the approximation $\varphi^+ \simeq -\varphi^-$, as suggested for the two-wavelength case by Giacovazzo & Siliqi (2001*b*). We obtain

$$P(R, G, \varphi^+ | E_a^+, E_a^-) = 2\pi P(R, G, \varphi^+, -\varphi^+ | E_a^+, E_a^-).$$

Then,

$$\begin{aligned}&P(\varphi^+ | R, G, E_a^+, E_a^-) \\ &\quad = P(R, G, \varphi^+ | E_a^+, E_a^-)\Big/\int_{-\pi}^{+\pi} P(R, G, \varphi^+ | E_a^+, E_a^-)\,\mathrm{d}\varphi^+\end{aligned}$$

may be calculated, which coincides with equation (1).

The authors thank the referees for their very useful suggestions.

References

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.
Collaborative Computational Project Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z., Wilson, K. S., Sieker, L. C., Meyer, J. & Moulis, J.-M. (1997). *Biochemistry*, **36**, 16065–16073.
Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.
Giacovazzo, C. & Siliqi, D. (2001*a*). *Acta Cryst.* A**57**, 40–46.
Giacovazzo, C. & Siliqi, D. (2001*b*). *Acta Cryst.* A**57**, 700–707.
Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *Acta Cryst.* D**55**, 1568–1577.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Li, J., Derewenda, U., Dauter, Z., Smith, S. & Derewenda, Z. S. (2000). *Nature Struct. Biol.* **7**, 555–559.
Moncrief, J. W. & Lipscomb, W. N. (1966). *Acta Cryst.* **21**, 322–331.
Ramachandran, G. N. & Raman, S. (1956). *Curr. Sci. India*, **25**, 348.
Raman, S. (1959). *Acta Cryst.* **12**, 964–975.
Schneider, R. & Scheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–816.
Sim, G. A. (1960). *Acta Cryst.* **13**, 511–512.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.