Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 4 October 1999
Accepted 14 February 2000

On the extrapolation of the magnitudes $|E|$ of the normalized structure factors E

Hongliang Xu* and Herbert A. Hauptman

Hauptman–Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203, USA. Correspondence e-mail: xu@hwi.buffalo.edu

For each fixed reciprocal-lattice vector **H**, the correlation coefficient, $\rho_{\mathbf{H}}$, of the pair $(|E_{\mathbf{K}}|, |E_{\mathbf{H-K}}|)$, where the magnitudes $|E_{\mathbf{K}}|$ and $|E_{\mathbf{H-K}}|$ are presumed to be known, is itself positively correlated with $|E_{\mathbf{H}}|$. Thus, the correlation coefficients $\rho_{\mathbf{H}}$ serve to mediate the transfer of information from one region of reciprocal space to another. In particular, the calculated values of the correlation coefficients $\rho_{\mathbf{H}}$ lead to estimates of the unknown magnitudes $|E_{\mathbf{H}}|$, even when the latter lie outside the sphere of the experimentally observed $|E|$s (extrapolation). The applications show that the procedure described here to carry out this extrapolation is superior to existing methods.

1. Introduction

It is well known that the inability to collect diffraction data to sufficiently high resolution often limits the effectiveness of phase determination by direct methods. It has been demonstrated in direct-methods applications that the adverse effects of missing data may be magnified because of the number of triplet invariant relationships that become unavailable for use (Xu *et al.*, 2000*a*,*b*). It is therefore a matter of some importance to devise methods that will exploit measured magnitudes $|E|$ to estimate the values of other magnitudes $|E|$ not initially obtainable experimentally. In the present paper, a procedure is described that realizes this goal.

For noncentric structures, the distribution for the normalized structure factor is defined by

$$P(|E|) = 2|E| \exp(-|E|^2), \tag{1}$$

and the expected value of $|E|$ is

$$\langle |E| \rangle = \int_0^\infty |E| P(|E|)\, \mathrm{d}|E| = \pi^{1/2}/2. \tag{2}$$

This is a crude estimate of the values of nonmeasured intensities, and this relationship is almost useless for most practical purposes. Van der Putten *et al.* (1982) proposed probabilistic expressions for estimating $|E_{\mathbf{H}}|$ using all the most reliable quartets in which **H** is a cross term. David (1987) suggested an algebraic formula

$$\langle |F_{\mathbf{H}}|^2 \rangle \approx \sum_{\mathbf{K}} |F_{\mathbf{K}}|^2 |F_{\mathbf{H-K}}|^2, \tag{3}$$

based on the fact that the Patterson function $P(\mathbf{u})$ and its square $P^2(\mathbf{u})$ are positive functions. Following up on this earlier work, more recently Cascarano *et al.* (1991) presented an improved formula:

$$\langle |E_{\mathbf{H}}| \rangle = \pi^{1/2}/2[1 + L(2N)^{-1} \langle (|E_{\mathbf{K}}|^2 - 1)(|E_{\mathbf{H-K}}|^2 - 1) \rangle], \tag{4}$$

based on the joint probability distribution of three normalized structure factors (Hauptman & Karle, 1953; Cochran, 1955), where L is a constant determined by the requirement that the average of the $|E_{\mathbf{H}}|^2$s must be unity. While equation (4) proved to be superior to the earlier attempts, its success was still limited and the goal of reliably estimating $|E|$s, given the values of measured $|E|$s, remained elusive. As we shall show in the applications, the formulation given here, while still less than perfect, is nevertheless superior to equation (4).

In this paper, we outline a different mathematical approach that yields estimates of the values of some nonmeasured diffraction amplitudes on the basis of suitable correlation coefficients. We first demonstrate that the values of nonmeasured $|E_{\mathbf{H}}|$s are correlated with the values of the correlation coefficients between $|E_{\mathbf{K}}|$ and $|E_{\mathbf{H-K}}|$ (both presumed to be measured). We then derive an empirical formula (dependent on the correlation coefficient and $\sin\theta/\lambda$ only) to assign a value to $|E_{\mathbf{H}}|$. Experimental tests on several data sets show a significant improvement over existing methods.

2. Correlation coefficient

Let x and y be any two random variables with means $\mu_x = m[x]$ and $\mu_y = m[y]$ and positive variances $\sigma_x^2 = m[(x - \mu_x)^2]$ and $\sigma_y^2 = m[(y - \mu_y)^2]$. The *correlation coefficient* between x, y is defined by

$$\rho(x, y) = \mathrm{cov}(x, y)/\sigma_x \sigma_y, \tag{5}$$

where

$$\mathrm{cov}(x, y) = m[(x - \mu_x)(y - \mu_y)] \tag{6}$$

is the covariance of x and y. If $\rho(x, y) = 0$, we say that x and y are *uncorrelated* (to distinguish from independent). Clearly, this correlation coefficient is independent of the origins and units of measurements, that is, for any constant a_1, a_2, b_1, b_2, with $a_1 > 0$, $a_2 > 0$, we have

$$\rho(a_1x + b_1, a_2y + b_2) = \rho(x, y). \tag{7}$$

The correlation coefficient provides a measure of how well one may predict the value of one of the random variables on the basis of the observed value of the other. It is well known that

$$-1 \leq \rho(x, y) \leq 1. \tag{8}$$

Further, $\rho^2(x, y) = 1$ if and only if

$$\text{Probability}(y = ax + b) = 1. \tag{9}$$

From (9), it follows that if the correlation coefficient equals 1 or -1 then prediction is perfect: to a given value of one of the random variables, there is one and only one value that the other random variable can assume.

If $\mathbf{H}$ is an arbitrary reciprocal-lattice vector, then the normalized structure factor $E_{\mathbf{H}}$ is defined by

$$E_{\mathbf{H}} = |E_{\mathbf{H}}| \exp(i\varphi_{\mathbf{H}}) = N^{-1/2} \sum_{j=1}^{N} \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \tag{10}$$

where N is the number of atoms, assumed to be identical, in the unit cell and $\mathbf{r}_j$ is the position vector of the atom labeled j. It is assumed that the atomic position vectors $\mathbf{r}_j$ are the primitive random variables, uniformly and independently distributed in the asymmetric unit.

For each fixed reciprocal-lattice vector $\mathbf{H}$ outside a certain limiting sphere, let $\mathbf{K}$ range over all reciprocal-lattice vectors within the limiting sphere such that $|E_{\mathbf{K}}|$ and $|E_{\mathbf{H}-\mathbf{K}}|$ are measured. Then, $|E_{\mathbf{K}}|$ and $|E_{\mathbf{H}-\mathbf{K}}|$, as functions of the random variables $\mathbf{r}_j$, are themselves random variables. Let $x = |E_{\mathbf{K}}|$ and $y = |E_{\mathbf{H}-\mathbf{K}}|$, for each fixed $\mathbf{H}$, the correlation coefficient $\rho_{\mathbf{H}} = \rho(|E_{\mathbf{K}}|, |E_{\mathbf{H}-\mathbf{K}}|)$, in view of equation (5), becomes

$$\rho_{\mathbf{H}} = \frac{\sum_{\mathbf{K}}(|E_{\mathbf{K}}| - \overline{|E_{\mathbf{K}}|})(|E_{\mathbf{H}-\mathbf{K}}| - \overline{|E_{\mathbf{H}-\mathbf{K}}|})}{[\sum_{\mathbf{K}}(|E_{\mathbf{K}}| - \overline{|E_{\mathbf{K}}|})^2 \sum_{\mathbf{K}}(|E_{\mathbf{H}-\mathbf{K}}| - \overline{|E_{\mathbf{H}-\mathbf{K}}|})^2]^{1/2}} \tag{11}$$

$$= \frac{\overline{|E_{\mathbf{K}}||E_{\mathbf{H}-\mathbf{K}}|} - \overline{|E_{\mathbf{K}}|}\,\overline{|E_{\mathbf{H}-\mathbf{K}}|}}{(\overline{|E_{\mathbf{K}}|^2} - \overline{|E_{\mathbf{K}}|}^2)^{1/2}(\overline{|E_{\mathbf{H}-\mathbf{K}}|^2} - \overline{|E_{\mathbf{H}-\mathbf{K}}|}^2)^{1/2}}, \tag{12}$$

where $\overline{|E_{\mathbf{K}}|}$ is the average value of the $|E_{\mathbf{K}}|$s. In view of the known joint probability distribution of the pair $(|E_{\mathbf{K}}|, |E_{\mathbf{H}-\mathbf{K}}|)$ (Hauptman, 1972), one anticipates that $\rho_{\mathbf{H}}$ and $|E_{\mathbf{H}}|$ are positively correlated, *i.e.* the probability is high that $|E_{\mathbf{H}}|$ will be large when $\rho_{\mathbf{H}}$ is large and $|E_{\mathbf{H}}|$ will be small when $\rho_{\mathbf{H}}$ is small. Therefore, the value of $\rho_{\mathbf{H}}$ can be used to predict the value of $|E_{\mathbf{H}}|$.

3. Determination of an empirical formula

If data up to $\sin\theta/\lambda = s$ are available, we shall try to estimate the magnitudes $|E|$ in the shell (s, s'), where $s' > s$. For each reciprocal-lattice vector $\mathbf{H}$ with $s < \sin\theta/\lambda \leq s'$, the correlation coefficient $\rho_{\mathbf{H}} = \rho(|E_{\mathbf{K}}|, |E_{\mathbf{H}-\mathbf{K}}|)$ can be calculated using (12). Of course, the value of the correlation coefficient depends on the resolution, which determines the number of contributors on the right-hand side of (12). In order to reduce the effects of resolution, we define the so-called 'normalized correlation coefficient $\hat{\rho}_{\mathbf{H}}$': If the minimum and maximum values of the correlation coefficients in the spherical shell (s, s') are denoted by $\rho_{\min}$ and $\rho_{\max}$, respectively, then the normalized correlation coefficient $\hat{\rho}_{\mathbf{H}}$ is defined by

$$\hat{\rho}_{\mathbf{H}} = (2\rho_{\mathbf{H}} - \rho_{\min} - \rho_{\max})/(\rho_{\max} - \rho_{\min}). \tag{13}$$

Since the smallest value of $\hat{\rho}_{\mathbf{H}}$ occurs when $\rho_{\mathbf{H}} = \rho_{\min}$, in which case $\hat{\rho}_{\mathbf{H}} = -1$, and the largest value of $\hat{\rho}_{\mathbf{H}}$ occurs when $\rho_{\mathbf{H}} = \rho_{\max}$, in which case $\hat{\rho}_{\mathbf{H}} = 1$, it follows that

$$-1 \leq \hat{\rho}_{\mathbf{H}} \leq 1. \tag{14}$$

Furthermore, it is clear that the relationship between $\rho_{\mathbf{H}}$ and $\hat{\rho}_{\mathbf{H}}$ is order preserving in the sense that $\rho_{\mathbf{H}_1} < \rho_{\mathbf{H}_2}$ implies $\hat{\rho}_{\mathbf{H}_1} < \hat{\rho}_{\mathbf{H}_2}$.

We describe our procedure using error-free data for alpha-1 peptide (508 non-H atoms, space group $P1$). There are 52410 reflections available to $s = 0.685\ \text{Å}^{-1}$ and 7018 reflections in the $\sin\theta/\lambda$ range $(0.685, 0.714\ \text{Å}^{-1})$. All 7018 correlation coefficients are calculated by (12) and normalized by (13). The $\hat{\rho}_{\mathbf{H}}$ are then sorted in decreasing order and divided into 140 groups with 50 reflections in each group. For each of these 140 groups, N, the percentage of reflections with $|E| \geq 1.0$ is calculated and $\overline{\hat{\rho}}$ and $\overline{|E|}$, the average values of $\hat{\rho}_{\mathbf{H}}$s and $|E_{\mathbf{H}}|$s, respectively, in each of these 140 groups, are also calculated. The 140 pairs $(\overline{\hat{\rho}}, N)$ are plotted in Fig. 1 and the 140 pairs $(\overline{\hat{\rho}}, \overline{|E|})$ are plotted in Fig. 2. It is observed from Fig. 1 that the percentage of $|E| \geq 1.0$ in each group decreases as $\overline{\hat{\rho}}$ decreases. Again, from Fig. 2, it is clear that $\overline{|E|}$, the average value of $|E|$ in each group, decreases as $\overline{\hat{\rho}}$ decreases. Note, in particular, for the first group, where $\overline{\hat{\rho}} \approx 0.737$, 94% of $|E|$s are greater than 1.0 (Fig. 1) and $\overline{|E|} \approx 1.805$ (Fig. 2). For the

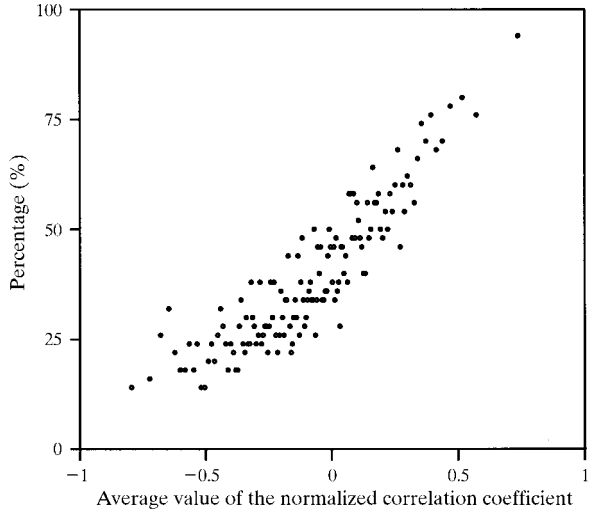

Figure 1
Percentage of $|E| \geq 1.0$ as a discrete function of $\overline{\hat{\rho}}$, the average value of the normalized correlation coefficient $\hat{\rho}$ within 140 groups each consisting of 50 $|E|$s.

Table 1
Experimental values of α, β and γ using error-free data sets for alpha-1 peptide, crambin and gramicidin A at several $\sin\theta/\lambda$ ranges.

	Alpha-1			Crambin			Gramicidin A		
$\sin\theta/\lambda$ (Å^{-1})	α	β	γ	α	β	γ	α	β	γ
(0.685, 0.714)	0.280	2.10	0.900	0.283	2.08	0.950	0.270	2.05	0.930
(0.641, 0.667)	0.245	2.00	0.853	0.242	2.03	0.873	0.230	1.94	0.827
(0.602, 0.625)	0.226	1.90	0.837	0.221	1.97	0.826	0.195	1.85	0.815
(0.568, 0.588)	0.198	1.85	0.828	0.191	1.83	0.810	0.186	1.80	0.810
(0.538, 0.556)	0.186	1.80	0.813	0.182	1.80	0.800	0.180	1.77	0.807
(0.510, 0.526)	0.179	1.75	0.807	0.178	1.77	0.793	0.175	1.68	0.800

Table 2
Comparison of the residuals, RES_P and RES_C using error-free data sets for alpha-1 peptide, crambin and gramicidin A.

	Alpha-1		Crambin		Gramicidin A	
$\sin\theta/\lambda$ (Å^{-1})	RES_P	RES_C	RES_P	RES_C	RES_P	RES_C
(0.685, 0.714)	0.459	0.392	0.448	0.409	0.490	0.423
(0.641, 0.667)	0.500	0.396	0.528	0.410	0.528	0.438
(0.602, 0.625)	0.532	0.412	0.552	0.414	0.533	0.432
(0.568, 0.588)	0.528	0.414	0.533	0.417	0.514	0.430
(0.538, 0.556)	0.538	0.409	0.535	0.428	0.502	0.460
(0.510, 0.526)	0.497	0.412	0.533	0.424	0.483	0.430

last group, on the other hand, where $\overline{\hat{\rho}} \approx -0.796$, only 14% of $|E|$s are greater than 1.0 (Fig. 1) and $\overline{|E|} \approx 0.693$ (Fig. 2). An empirical function of the form

$$|E_{\rm cal}| = \alpha[\exp(\beta\hat{\rho}) - 1] + \gamma, \tag{15}$$

with $\alpha = 0.28$, $\beta = 2.1$ and $\gamma = 0.90$, is then determined by least squares using the data shown in Fig. 2. The same procedure has been repeated using error-free data for alpha-1 peptide in the different $s = \sin\theta/\lambda$ ranges, and the corresponding values of α, β and γ are obtained and listed in Table 1.

In order to study the effects of space groups on the values of α, β and γ, we repeated the same procedure using error-free data sets for crambin (327 non-H atoms, space group $P2_1$) and gramicidin A (317 non-H atoms, space group $P2_12_12_1$). The corresponding values of α, β and γ are also listed in Table 1. It is clear that the values of α, β and γ are resolution-dependent and structure-dependent, but the space groups have little impact on the values of α, β and γ (at least for space groups $P1$, $P2_1$ and $P2_12_12_1$). So, we can seek an empirical function of the form (15) and assume that α, β and γ are resolution-dependent only. For each of α, β and γ, we apply the least-squares method on the three data sets listed in Table 1 and obtain

$$\alpha = 0.16 + \exp(-1/s^2), \tag{16}$$

$$\beta = 1.68 + 3.50\exp(-1/s^2), \tag{17}$$

$$\gamma = 0.77 + \exp(-1/s^2). \tag{18}$$

The empirical formula (15) is now employed to estimate the values of nonmeasured $|E|$, where $\hat{\rho}$ is defined by (13), and α, β and γ are defined by (16)–(18). Several issues, such as the validation of the formula and the effects of real data *etc*., are addressed in the next section.

4. Results and conclusions

The general effectiveness of the method may be assessed by the values of the residual

$$\mathrm{RES} = \sum_{\mathbf{H}} \big||E_{\rm obs}| - |E_{\rm cal}|\big| \Big/ \sum_{\mathbf{H}} |E_{\rm obs}|. \tag{19}$$

The notations RES_P and RES_C denote the residuals corresponding to equations (4) and (15), respectively, which in turn are based on the probabilistic approach and the correlation-coefficient method described here.

The residuals RES_P and RES_C using error-free data sets for alpha-1 peptide, crambin and gramicidin A are listed in Table 2. The estimates based on the correlation coefficient are in all cases better than the estimates based on the probabilistic formula (4). The effectiveness of equation (15) is slightly reduced with real data. For example, the $|E|$ values of 5705 reflections in the $\sin\theta/\lambda$ range 0.568–0.588 Å^{-1} have been estimated by equation (15) using both real data and error-free data as prior information and $\mathrm{RES}_C = 0.426$ for real data, $\mathrm{RES}_C = 0.414$ for error-free data.

As an application of (15), we have used experimental data sets from 12 structures crystallizing in space groups $P1$, $P2_1$

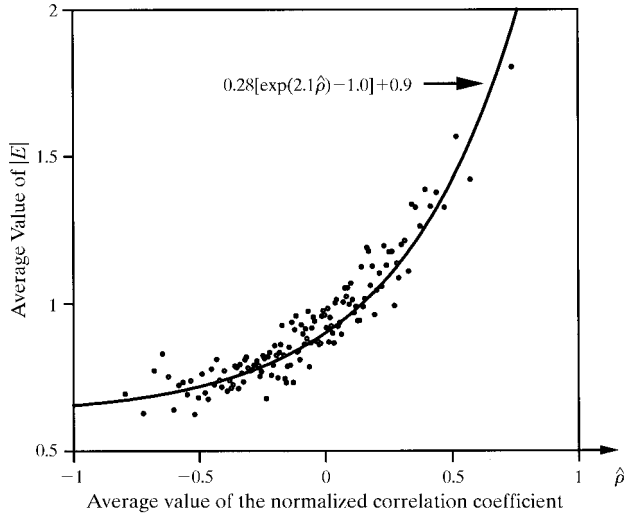

Figure 2
Average value of $|E|$ as a discrete function of $\overline{\hat{\rho}}$, the average value of the normalized correlation coefficient $\hat{\rho}$ within 140 groups each consisting of 50 $|E|$s.

Table 3
Structure name, number of atoms, space group and resolution for the test structures.

	Atoms	Space group	Resolution (Å)	Reference
Emerimycin	74	$P1$	0.91	Marshall *et al.* (1990)
Isoleucinomycin	84	$P2_12_12_1$	0.94	Pletnev *et al.* (1980)
Enkephalin analog	96	$P1$	0.83	Krstenansky (unpublished)
Ternatin	104	$P2_12_12_1$	0.94	Miller *et al.* (1993)
Hexaisoleucinomycin	113	$P2_12_12_1$	1.00	Pletnev *et al.* (1992)
Gramicidin A	317	$P2_12_12_1$	0.86	Langs (1988)
Crambin	327	$P2_1$	0.83	Hendrickson & Teeter (1981)
Rubredoxin	395	$P2_1$	1.00	Dauter *et al.* (1992)
Vancomycin	404	$P1$	0.97	Loll *et al.* (1998)
Alpha-1 peptide	408	$P1$	0.90	Prive *et al.* (1999)
Scorpion toxin II	508	$P2_12_12_1$	0.96	Smith *et al.* (1997)
Triclinic lysozyme	1001	$P1$	0.85	Deacon *et al.* (1998)

Table 4
Comparison of the residuals, RES_P and RES_C, using real data.

	$\sin\theta/\lambda$ (Å^{-1})	RES_P	RES_C
Emerimycin	(0.526, 0.546)	0.469	0.395
Isoleucinomycin	(0.510, 0.532)	0.462	0.358
Enkephalin analog	(0.556, 0.588)	0.393	0.324
Ternatin	(0.516, 0.533)	0.562	0.475
Hexaisoleucinomycin	(0.472, 0.500)	0.467	0.315
Gramicidin A	(0.556, 0.580)	0.434	0.367
Crambin	(0.556, 0.603)	0.504	0.361
Rubredoxin	(0.476, 0.500)	0.483	0.355
Vancomycin	(0.513, 0.532)	0.445	0.417
Alpha-1 peptide	(0.556, 0.575)	0.552	0.426
Scorpion toxin II	(0.500, 0.518)	0.419	0.332
Triclinic lysozyme	(0.556, 0.588)	0.489	0.414

and $P2_12_12_1$. These are shown in Table 3. For each test structure, the range of $\sin\theta/\lambda$ and the residuals, RES_P and RES_C, are given in Table 4. Analysis of Table 4 shows that the correlation-coefficient estimation [equation (15)] is always better than the probabilistic estimate [equation (4)]. Clearly, equation (15) constitutes a significant improvement over the existing methods. It is worth pointing out that (15), together with (16)–(18), is obtained empirically based on the high-resolution data. For the low-resolution data, the correlation coefficient estimation, like all other existing methods, is no longer applicable.

We would like to express our appreciation to Drs C. M. Weeks and R. Blessing for valuable discussions. This research was supported by NIH grant GM-46733 and NSF grant ACI-9721373.

References

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1991). *Acta Cryst.* A**47**, 480–484.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* B**48**, 42–59.
David, W. I. F. (1987). *J. Appl. Cryst.* **20**, 316–319.
Deacon, A. M., Weeks, C. M., Miller, R. & Ealick S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.
Hauptman, H. A. (1972). *Crystal Structure Determination.* New York: Plenum Press.
Hauptman, H. A. & Karle, J. (1953). *The Solution of the Phase Problem I. The Centrosymmetric Crystal. Am. Crystallogr. Assoc.* Monogr. No 3. New York: Polycrystal Book Service.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Langs, D. A. (1988). *Science*, **241**, 188–191.
Loll, P. J., Miller, R., Weeks, C. M. & Axelsen, P. H. (1998). *Chem. Biol.* **5**, 293–298.
Marshall, G. R., Hodgkin, E. E., Langs, D. A., Smith, G. D., Zabrocki, J. & Leplawy, M. T. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 487–491.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Pletnev, V. Z., Galitskii, N. M., Smith, G. D., Weeks, C. M. & Duax, W. L. (1980). *Biopolymers*, **19**, 1517–1534.
Pletnev, V. Z., Ivanov, V. T., Langs, D. A., Strong, P. & Duax, W. L. (1992). *Biopolymers*, **32**, 819–827.
Prive, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Protein Sci.* **8**, 1–9.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.
Van der Putten, N., Schenk, H. & Tsoucaris, G. (1982). *Acta Cryst.* A**38**, 98–102.
Xu, H., Hauptman, H. A., Weeks, C. M. & Miller, R. (2000*a*). *Acta Cryst.* D**56**, 238–240.
Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. A. (2000*b*). *Acta Cryst.* A**56**, 112–118.