Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Globbic approximation in low-resolution direct-methods phasing

D. Y. Guo,* Robert H. Blessing and David A. Langs

Hauptman–Woodward Institute, 73 High Street, Buffalo, New York 14203, USA

Correspondence e-mail: guo@hwi.buffalo.edu

Received 3 March 2000
Accepted 9 June 2000

Probabilistic direct-methods phasing theory, originally based on a uniform atomic distribution hypothesis, is shown to be adaptable to a non-uniform bulk-solvent-compensated globbic approximation for protein crystals at low resolution. The effective number n_g of non-H protein atoms per polyatomic glob increases with decreasing resolution; low-resolution phases depend on the positions of only $N_g = N_a/n_g$ globs rather than N_a atoms. Test calculations were performed with measured structure-factor data and the refined structural parameters from a protein crystal with ~10 000 non-H protein atoms per molecule and ~60% solvent volume. Low-resolution data sets with $d_{\min}$ ranging from 15 to 5 Å gave $n_g = ad_{\min} + b$, with $a = 1.0$ Å^{-1} and $b = -1.9$ for the test case. Results of tangent-formula phase-estimation trials emphasize that completeness of the low-resolution data is critically important for probabilistic phasing.

1. Introduction

In two respects, protein crystals appear to violate the starting hypothesis of probabilistic direct-methods phasing theory, *viz.* the hypothesis of uniform random distributions of independent atoms. Firstly, protein crystals seldom diffract to atomic resolution and secondly, the crystals are partitioned into more-or-less distinct higher density protein and lower density solvent regions.

In a recent paper (Guo *et al.*, 2000*a*), we showed that since the protein and solvent regions represent Babinet complementary opposite-phase scattering masks (Fig. 1), the scattering by the partitioned protein and solvent electron density is equivalent to scattering by protein-minus-solvent difference electron density in the protein regions and zero density in the solvent regions outside the protein regions. This led to a reformulation of the crystal structure factor in terms of difference atomic scattering factors. Thus,

$$F_{\mathbf{h}} = \sum_{j=1}^{N_a} (f_{j\mathbf{h}} - f_{S\mathbf{h}}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j), \tag{1}$$

where N_a is the number of non-H protein atoms per unit cell and the $f_{S\mathbf{h}}$ are Fourier transforms of protein-atom-centered large-radius spheres of dilute uniform bulk-solvent-compensating electron density. The bulk-solvent-compensating scattering factors are given by

$$f_{S\mathbf{h}} = Z_S \Phi(2\pi |\mathbf{h}| R_S), \tag{2}$$

in which

$$\Phi(u) = 3(\sin u - u \cos u)/u^3, \tag{3}$$

$$|\mathbf{h}| = 1/d_{hkl} = 2(\sin \theta_{hkl})/\lambda \tag{4}$$

and $Z_S \simeq 5.75$ e and $R_S \simeq 5$ Å from empirical calculations in which the bulk solvent was assumed to be liquid-like water. [(2) is a corrected form (Guo *et al.*, 2000*b*) corresponding to equation (12) in our earlier paper (Guo *et al.*, 2000*a*), which as printed incorrectly included a factor $(4/3)R_S^3$ on the right-hand side of (12).]

In another recent paper (Guo *et al.*, 1999), we showed that for low-resolution protein structure analysis Debye–Harker scattering factors for spherically averaged n-atom groups or globs,

$$g_{\mathbf{h}} = \left[\sum_{j=1}^{n}\sum_{k=1}^{n} f_{j\mathbf{h}} f_{k\mathbf{h}} \frac{\sin(2\pi|\mathbf{h}||\mathbf{r}_j - \mathbf{r}_k|)}{2\pi|\mathbf{h}||\mathbf{r}_j - \mathbf{r}_k|}\right]^{1/2}, \qquad (5)$$

in particular the globbic scattering factors for main-chain peptide groups $-C^{\alpha}-C'(=O)-N-$ and amino-acid side-chain groups $-C^{\beta}-R$, are good approximations in structure-factor and electron-density calculations with $d_{hkl} > 3.5$ Å. It was also shown that globbic scattering factors from (5) could be approximated quite well at low resolution by single Gaussian functions, $g_{\mathbf{h}} = a\exp(-b|\mathbf{h}|^2)$.

In the present paper, we use bulk-solvent-compensated globbic scattering factors to adapt probabilistic phasing theory from the uniform atomic distribution hypothesis to a non-uniform globbic distribution approximation to low-resolution protein crystal structure.

2. Probabilistic theory in the globbic approximation

Some main points of probabilistic phasing theory are summarized in *Appendix A*. To adapt the theory from atoms to globs, we replace the N_a atomic factors $(f_{j\mathbf{h}} - f_{S\mathbf{h}})$ in (1) by $N_g < N_a$ polyatomic globbic factors $(g_{j\mathbf{h}} - g_{S\mathbf{h}})$. The normalized crystal structure factors defined by (21) and (22) in *Appendix A* are then redefined by

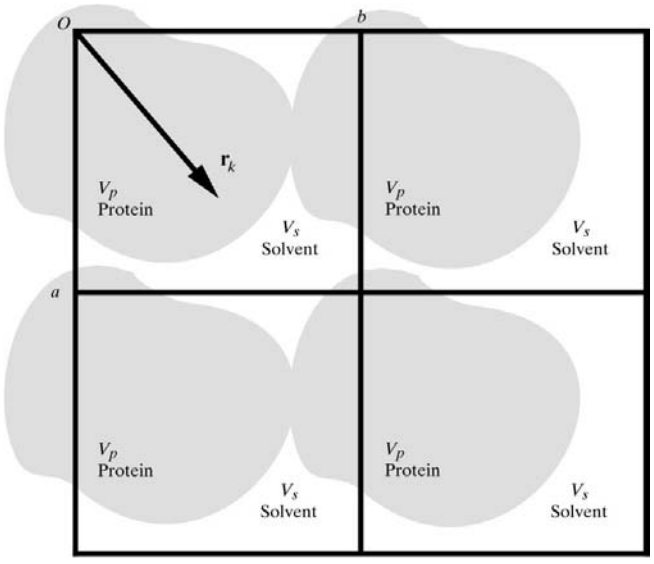

Figure 1
Schematic illustration of the protein–solvent partitioning that results in non-uniform atomic position distributions in protein crystals.

$$E_{\mathbf{h}} = \frac{F_{\mathbf{h}}}{\langle|F|^2\rangle_{\mathbf{h}}^{1/2}} = \sum_{j=1}^{N_g} \frac{g_{j\mathbf{h}} - g_{S\mathbf{h}}}{\gamma_{\mathbf{h}}^{1/2}} \exp(2\pi i\mathbf{h}\cdot\mathbf{r}_j) = |E_{\mathbf{h}}|\exp(i\varphi_{\mathbf{h}}), \qquad (6)$$

where $\langle|E|^2\rangle_{\mathbf{h}} = 1$, since

$$\gamma_{\mathbf{h}} = \langle|F|^2\rangle_{\mathbf{h}} = \sum_{j=1}^{N_g}(g_{j\mathbf{h}} - g_{S\mathbf{h}})^2 \qquad (7)$$

is the Wilson (1949) expectation value for the intensity of scattering by N_g bulk-solvent-compensated globs distributed independently and uniformly within the protein molecular volume but excluded from the bulk-solvent volume, *i.e.*

$$\mathbf{r}_j \subseteq V_P = V_{\text{cell}} - V_S, \quad \mathbf{r}_j \not\subseteq V_S. \qquad (8)$$

As mentioned in *Appendix A*, key steps in the probabilistic derivations involve simplifications based on averages $\langle\cos x\rangle = 0$ and $\langle\cos^2 x\rangle = \frac{1}{2}$ when x spans the whole interval $0 \le x \pmod{2\pi} < 2\pi$. Under the non-uniform globbic distribution approximation we cannot assume the same simplifications, because in averages such as

$$c_{1,\mathbf{h}} = \langle\cos(2\pi\mathbf{h}\cdot\mathbf{r})\rangle_{\mathbf{r}} \text{ and } c_{2,\mathbf{h}} = \langle\cos^2(2\pi\mathbf{h}\cdot\mathbf{r})\rangle_{\mathbf{r}} \qquad (9)$$

the arguments may not span the whole interval $(0, 2\pi)$. The $\mathbf{h}$ are restricted to low-resolution reflections with small $|\mathbf{h}|$ and the $\mathbf{r}$ are restricted by (8) to the protein molecular subvolume of the unit cell.

We saw no route to general results for restricted averages (9), so we examined their behavior by numerical calculations for a hypothetical model structure in space group $P1$ with unit-cell parameters $a = b = c = 100$ Å, $\alpha = \beta = \gamma = 90^\circ$ and $Z = 1$ spherical molecule centered on the origin. Restricted averages of the cosine and squared-cosine functions were computed for molecular-volume fractions $V_P/V_{\text{cell}} = 0.1$, 0.3 and 0.5 corresponding to spherical molecular volumes $V_P = (4/3)\pi R_P^3$ with radii $R_P = 28.8$, 41.5 and 49.2 Å, respectively. Each spherical molecular volume was subdivided by a $1 \times 1 \times 1$ Å grid and the averages (9) were computed over all the grid points with $|\mathbf{r}| \le R_P$ for each of 30 reflections chosen such that their $|\mathbf{h}|$ were approximately equally spaced in the $1/d_{hkl}$ interval corresponding to $100 > d_{hkl} > 10$ Å.

Fig. 2 shows that the restricted averages (9) fall close to the unrestricted average values 0 and $\frac{1}{2}$, except for small numbers of very low resolution reflections with $d_{hkl} \gtrsim 0.5\,V_{\text{cell}}^{1/3}$. We can therefore expect that the averages of cosine products in (26) for non-uniform random globbic distributions would be similar to the averages for uniform random atomic distributions. If that is the case, the probabilistic theory applies at least approximately, *mutatis mutandis*, under the bulk-solvent-compensated globbic approximation. The necessary changes for (28) through (34) are

$$N_a \to N_g, \quad f_{j\mathbf{h}} \to (g_{j\mathbf{h}} - g_{S\mathbf{h}}), \quad \alpha_{\mathbf{h}} \to \gamma_{\mathbf{h}}. \qquad (10)$$

With these changes, the $A_{\mathbf{hk}}$ values (30) and (31) increase because they are inversely proportional to $N_g^{1/2} < N_a^{1/2}$. Since the variance of the distribution (28) decreases with increasing $A_{\mathbf{hk}}$, the bulk-solvent-compensated globbic approximation

reduces the uncertainty of probabilistic estimation of low-resolution phases for large structures.

3. Single-Gaussian large globs

For test calculations at lower resolution with larger globs, we adopted scattering factors for spherical globs represented as single Gaussian functions. Thus, the forms of the $g_{j\mathbf{h}}$, $g_{S\mathbf{h}}$ and $f_{S\mathbf{h}}$ globbic scattering factors were simplified from (5) and (2) to

$$g_{\mathbf{h}} = Z[3(\sin u - u\cos u)/u^3] \simeq Z\exp(-0.1u^2), \qquad (11)$$

where $Z = Z_P$ or Z_S is the number of electrons per glob,

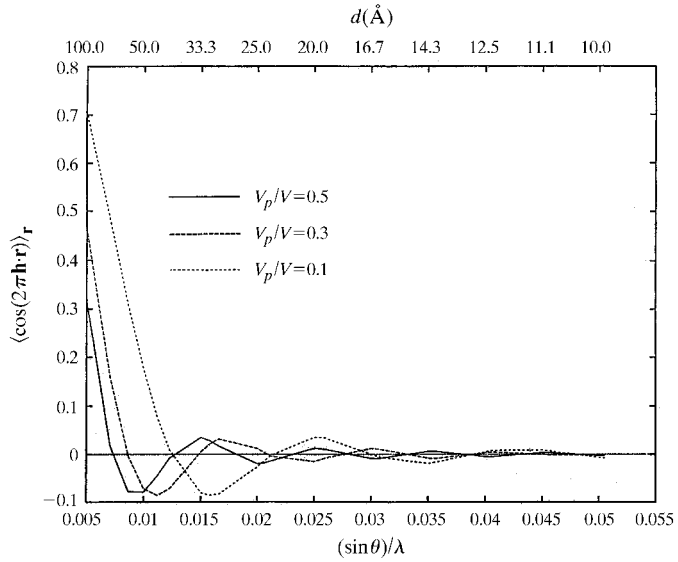

Figure 2
Plots of numerical averages of (*a*) the cosine and (*b*) the squared-cosine functions (9) for low-diffraction resolution and large solvent-volume fractions.

$$u = 2\pi|\mathbf{h}|R, \qquad (12)$$

and $R = R_P$ or R_S is the glob radius. As illustrated in Fig. 3, a single Gaussian represents a uniform density sphere very well at low resolution and does not have the disadvantage of the high-resolution nodes and ripples present in the Fourier transform of a uniform density or 'hard' sphere (Guinier, 1994*a*,*b*; Patterson, 1967).

4. Effective glob size *versus* resolution

Numerical tests of the bulk-solvent-compensated globbic approximation were carried out using the measured structure-factor amplitudes and the refined atomic positional mean-square displacement and site-occupancy parameters for crystals of glyceraldehyde-3-phosphate dehydrogenase or GPD (Murthy *et al.*, 1980), space group $P1$, unit-cell parameters a = 83.0, b = 81.0, c = 82.5 Å, α = 110.8, β = 71.5, γ = 116.9°, Z = 1 GPD molecule with N_a = 10 028 independent non-H protein atoms, M_r = 143 000, V_m = 3.18 Å^3 Da^{-1} and $V_P/V_{\rm cell} = 1 - (V_S/V_{\rm cell})$ = 38.7%. The measured data set of 37 655 reflections is 63% complete, with $d_{\rm max}$ = 45.9 Å and $d_{\rm min}$ = 2.5 Å. Assuming the bulk solvent to be liquid-like water with $V_{\rm H_2O}$ = 29.9 Å^3 per molecule, the 61.3% solvent volume in GPD crystals corresponds to 10 300 H_2O molecules.

The effect of the bulk-solvent-compensated globbic approximation on intensity-data normalization was assessed by calculation results summarized in Table 1 and Fig. 4. For the 5644 measured reflections with $d > 5$ Å in the GPD data set, a scale factor k and a radius R_S in the bulk-solvent-compensating scattering factors $f_{S\mathbf{h}}$ defined by (2) and (11) were determined by a two-parameter fit of $|F_c|$ to $|F_o|$, where

$$F_c = k^{-1}\sum_{j=1}^{N_a} p_j[f_{j\mathbf{h}}\exp(-2\pi^2\langle u_j^2\rangle/d_{\mathbf{h}}^2) - f_{S\mathbf{h}}]\exp(2\pi i\mathbf{h}\cdot\mathbf{r}_j), \qquad (13)$$

in which the $p_j \leq 1$ are atomic site-occupation parameters and the $\langle u_j^2\rangle = B_j/(8\pi^2)$ are atomic mean-square displacement

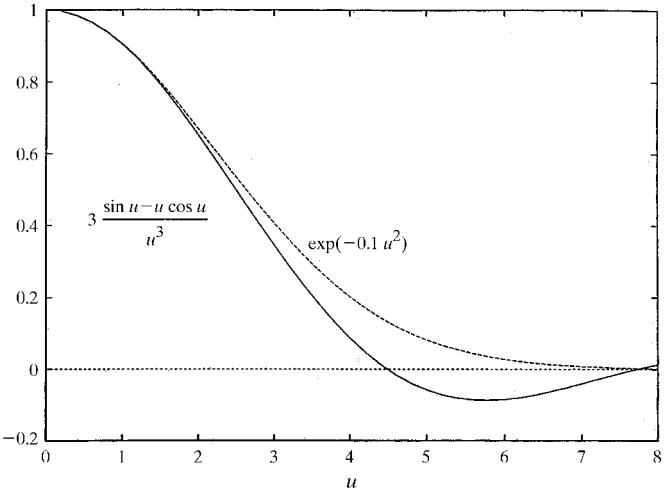

Figure 3
Scattering factors (11) for a sphere of uniform electron density and for a single-Gaussian electron-density distribution.

Table 1
Low-resolution intensity normalization and effective glob sizes.

The normalizing factors $\alpha_{\mathbf{h}}$ defined by (22) were computed for N_a = 10 028 non-H protein atoms of the GPD structure and the normalizing factors $\gamma_{\mathbf{h}}$ defined by (7) were computed for $N_g = N_a/n_g$ protein globs.

$d_{\min}$ (Å)	N_{refln}	$\langle k^2\lvert F_o\rvert^2\rangle \times 10^{-3}$	$\langle\lvert F_c\rvert^2\rangle / \langle k^2\lvert F_o\rvert^2\rangle$	$\langle\alpha\rangle / \langle k^2\lvert F_o\rvert^2\rangle$	$\langle\gamma\rangle / \langle k^2\lvert F_o\rvert^2\rangle$	n_g
15	135	307	1.22	1.42	0.96	13.08
14	175	312	1.14	1.39	1.00	12.12
13	230	319	1.07	1.35	1.04	11.07
12	307	321	1.04	1.33	1.11	10.07
11	415	356	0.99	1.19	1.08	9.06
10	564	436	0.91	0.96	0.94	8.08
9	793	505	0.84	0.82	0.87	7.07
8	1204	498	0.84	0.82	0.93	6.07
7	1885	442	0.92	0.89	1.05	5.06
6	3135	375	0.98	1.01	1.14	4.06
5	5644	357	1.00	0.99	0.95	3.06

parameters. The corresponding atomic normalization factors $\alpha_{\mathbf{h}}$ defined by (22) were computed as

$$\alpha_{\mathbf{h}} = \exp(-4\pi^2\langle u_{\text{iso}}^2\rangle/d_{\mathbf{h}}^2)\sum_{j=1}^{N_a} f_{j\mathbf{h}}^2, \tag{14}$$

where N_a = 10 028 non-H protein atoms and the overall mean-square atomic displacement parameter $\langle u_{\text{iso}}^2\rangle$ = 0.377 Å^2 (B_{iso} = 29.8 Å^2) had been obtained by a Wilson normalization analysis (Blessing *et al.*, 1996, 1998) of the measured $|F_o|$ data set. Low-resolution data-normalization ratios were compiled as averages $\langle\alpha\rangle/\langle k^2|F_o|^2\rangle$ over 11 data subsets each with $d_{\max}$ = 45.9 Å and with $d_{\min}$ ranging from 15 to 5 Å as listed in Table 1.

To compute the globbic normalization ratios in Table 1, we assumed equal globs and guessed that the effective number n_g of non-H protein atoms per glob would increase linearly with decreasing resolution limit, *i.e.*

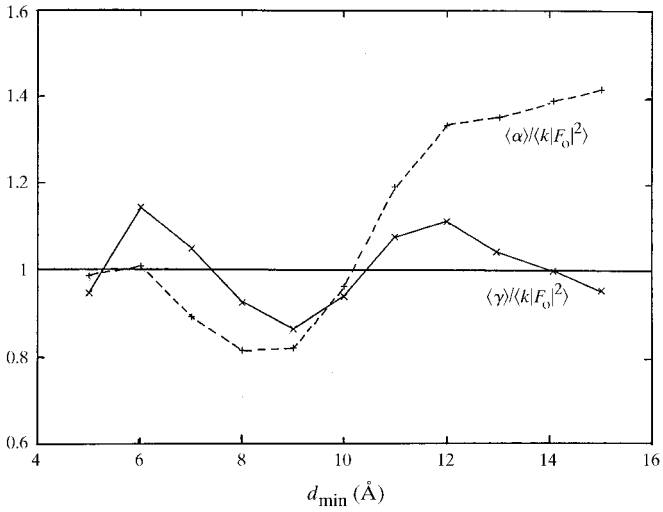

Figure 4
Plots of the atomic and globbic normalization ratios, $\langle\alpha\rangle/\langle k|F_o|^2\rangle$ and $\langle\gamma\rangle/\langle k|F_o|^2\rangle$, respectively, against resolution limit for cumulative low-resolution subsets of the measured GPD data set.

$$n_g = ad_{\min} + b. \tag{15}$$

Starting from a = 1 Å^{-1} and b = 0, the glob-size parameters a and b were refined by a simplex procedure (Press *et al.*, 1986) to adjust n_g, $N_g = N_a/n_g$ and the globbic normalization factors $\gamma_{\mathbf{h}}$, defined by (7) and computed as

$$\gamma_{\mathbf{h}} = N_g[g_{\mathbf{h}}\exp(-2\pi^2\langle u_{\text{iso}}^2\rangle/d_{\mathbf{h}}^2) - n_g f_{S\mathbf{h}}]^2, \tag{16}$$

with the $g_{\mathbf{h}}$ and $f_{S\mathbf{h}}$ from (11) and (2). The refinement minimized the normalized root-mean-square residual,

$$r = \left[\sum_{i=1}^{n}(\langle|F_o|^2\rangle_i - k^{-2}\langle\gamma\rangle_i)^2 / \sum_{i=1}^{n}\langle|F_o|^2\rangle_i^2\right]^{1/2}, \tag{17}$$

where i indexed the n = 11 low-resolution subsets with $d_{\max}$ = 45.9 Å and 15 > $d_{\min}$ > 5 Å. In each refinement cycle, the new a and b gave a new n_g for each resolution subset; then new N_g, Z_P, R_P and, through (11) and (16), new $g_{\mathbf{h}}$ and $\gamma_{\mathbf{h}}$ values were calculated. Constants of those of the calculations were from our earlier protein and bulk-solvent calculations:

$$\begin{aligned}
Z_{\text{H}_2\text{O}} &= 10 \text{ e} \\
V_{\text{H}_2\text{O}} &= 22.9 \text{ Å}^3 \\
\langle Z_{\text{nonH}}\rangle &= 6.61 \text{ e} \\
\langle V_{\text{nonH}}\rangle &= 17.2 \text{ Å}^3 \\
Z_P &= n_g\langle Z_{\text{nonH}}\rangle \\
R_P &= [(n_g\langle V_{\text{nonH}}\rangle)/(4\pi/3)]^{1/3} \\
Z_S &= (Z_{\text{H}_2\text{O}}/V_{\text{H}_2\text{O}})\langle V_{\text{nonH}}\rangle = 5.75 \text{ e} \\
R_S &= 5.78 \text{ Å},
\end{aligned} \tag{18}$$

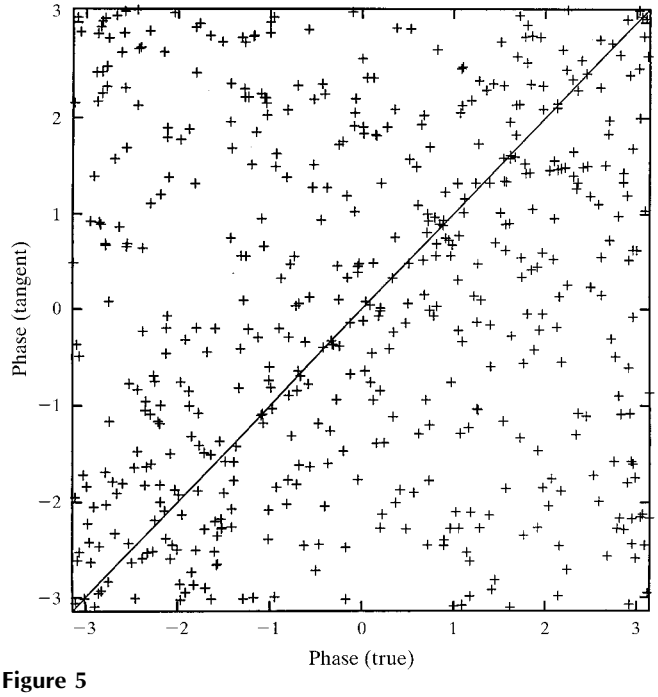

Figure 5
Tangent-formula-estimated *versus* structure-model-calculated phases (radians) for the 521 reflections with $|E_o|$ > 1.0 and d_{hkl} > 8 Å in the measured GPD data set.

where $\langle Z_{\rm nonH}\rangle$ and $\langle V_{\rm nonH}\rangle$ are the average atomic number and average volume per non-H protein atom (Guo *et al.*, 2000*a*). The scale factor k and the Z_S and R_S parameters of the bulk-solvent-compensating scattering factors $f_{S\mathbf{h}}$ were kept fixed at their values from the initial fit of (13) to the measured $|F_o|$ data. The refinement of the glob-size parameters converged to $r = 8.7\%$ with $a = 1.00$ Å^{-1} and $b = -1.94$. Table 1 and Fig. 4 show that the resolution-optimized globbic normalization ratios $\langle\gamma\rangle/\langle k^2|F_o|^2\rangle$ based on (15) and (16) are consistently closer to unity than are the atomic normalization ratios $\langle\alpha\rangle/\langle k^2|F_o|^2\rangle$ based on (14).

5. Tangent-formula phase estimation

In our earlier work concerned with bulk-solvent correction in direct-methods phasing (Guo *et al.*, 2000*a*) we showed that although the $|F|$-weighted tangent formula based on the Sayre equation gives essentially random phase estimates for protein data with $d < 8$ Å, the formula does give useful phase estimates for *complete* low-resolution data subsets with $\infty > d > 8$ Å. Now we have also tested low-resolution phase estimation by the probabilistic $|E|$-weighted tangent formula under the bulk-solvent-compensated globbic approximation.

Results from the protein GPD are summarized in Table 2 and Figs. 5, 6, 7 and 8. Using the tangent formula (34), phases $\varphi_{\mathbf{h}}$ were estimated from phases $\varphi_{\mathbf{k}}$ and $\varphi_{\mathbf{h}-\mathbf{k}}$ calculated from (13); unit-weighted and $|E|$-weighted average phase errors were calculated as

$$\langle|\Delta\varphi|\rangle = \sum_{\mathbf{h}} w_{\mathbf{h}}|\varphi_{\mathbf{h},\rm est} - \varphi_{\mathbf{h},\rm calc}|/\sum_{\mathbf{h}} w_{\mathbf{h}}, \quad w_{\mathbf{h}} = 1 \text{ or } |E_{\mathbf{h}}| \quad (19)$$

to compare estimated phases from (34) with calculated phases from (13).

The first line of Table 2 shows that for the 521 measured low-resolution reflections with $d > 8$ Å and $|E_o| = k|F_o|/\gamma^{1/2} > 1.0$ the average phase error is about 70° and the estimated phases are not much better than random phases, with an average phase error of 90°. However, there were 307 low-resolution reflections with $d > 8$ Å and $|E_c| = |F_c|/\gamma^{1/2} > 1.0$ that were missing from the measured data set. In stepwise increasing resolution shells, calculated data were supplied for the missing reflections to complete the $|E| > 1.0$ data in the lowest resolution shells and tangent-formula phase estimation trials were repeated as indicated in Table 2.

The effect of even a very few missing low-order reflections is dramatic. Supplying only the 11 missing near-origin reflections indicated in Fig. 8 reduced the average phase error from a practically useless 70° to a practically useful 45°. The phase improvement this represents is indicated clearly by the scatter plots shown in Figs. 5 and 6. The further improvement on supplying all 307 missing reflections is shown in Fig. 7. It is worth noting that the effect of adding the lowest order missing reflections was very beneficial even though the added reflections were in the near-origin region of fluctuating averages of the cosine and squared-cosine functions shown in Fig. 2.

6. Low-resolution data completeness

We think that a principal reason for the critical importance of even a very few low-resolution strong reflections is the common tendency for strong large-$|E|$ reflections to occur in clusters corresponding to reciprocal-lattice points separated

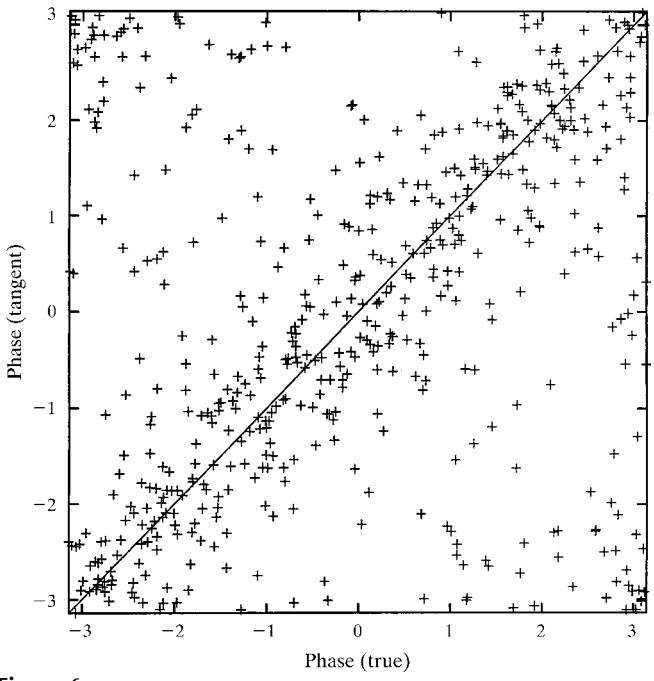

Figure 6
As Fig. 5, but with the 521 measured reflections supplemented by calculated data for the 11 reflections with $d_{hkl} > 40$ Å that were missing from the measured data set. See also Fig. 8.

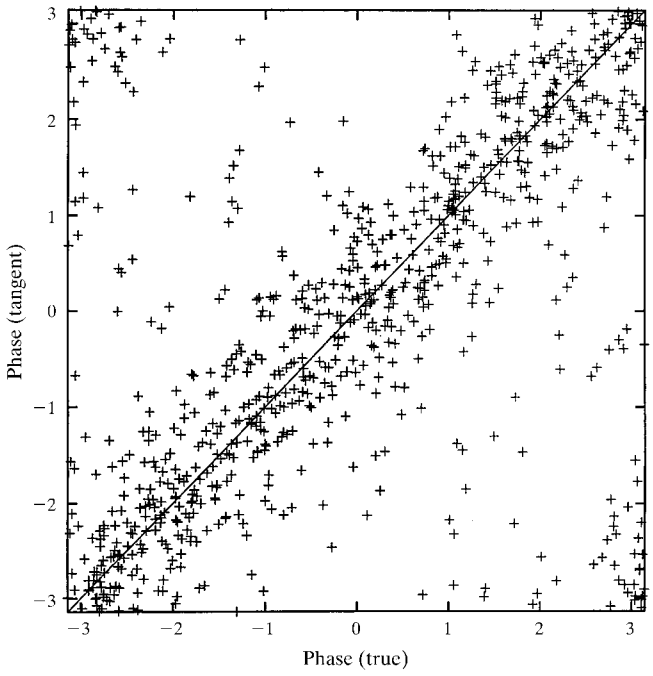

Figure 7
As Fig. 5, but with the 521 measured reflections supplemented by calculated data for all 307 reflections with $d_{hkl} > 8$ Å that were missing from the measured data set.

Table 2
Tangent-formula phase estimation *versus* low-resolution completeness using triplets formed of reflections with $d > 8$ Å and $|E| > 1.0$.

Calculated data for the lowest order reflections missing from the measured GPD data set were added shell-wise to complete the $|E| > 1.0$ data in the lowest resolution shells.

Missing reflections added	$N_{\rm reflns}$	Completeness (%)	$N_{\rm triplets}$	Unit-weighted $\langle\|\Delta\varphi\|\rangle$ (°)	$\|E\|$-weighted $\langle\|\Delta\varphi\|\rangle$ (°)
0	521	62.9	33868	72	71
2, $d > 70$ Å	523	63.2	35084	65	62
4, $d > 60$	525	63.4	36281	58	54
6, $d > 50$	527	63.6	37439	55	52
11, $d > 40$	532	64.3	40089	49	45
24, $d > 35$	545	65.8	46549	47	42
36, $d > 25$	557	67.3	52324	46	42
51, $d > 20$	572	69.1	59145	46	42
90, $d > 15$	611	73.8	76987	46	42
191, $d > 10$	721	87.1	117863	43	39
307, $d > 8$	828	100.0	161549	37	34

from one another by only one or two or a few reciprocal-lattice translations, as illustrated in Fig. 9. Thus, to form amplitude triplets $|E_{\mathbf{h}}|$, $|E_{\mathbf{k}}|$, $|E_{-\mathbf{h}-\mathbf{k}}|$ for three-phase structure invariants $\varphi_{\mathbf{hk}} = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}$ that involve neighboring strong reflections, one triplet component must correspond to a short reciprocal-lattice vector, as illustrated in Fig. 10. Table 2 shows that the effect of adding only the 11 lowest order reflections missing from the measured GPD data set was to generate more than 6000 additional strong triplets for tangent-formula phase estimation.

In terms of a crystal space interpretation, the strongest Fourier components of the density,

$$p(\mathbf{r}) = 1/V \sum_{\mathbf{h}} |F_{\mathbf{h}}| \cos(\varphi_{\mathbf{h}} - 2\pi\mathbf{h}\cdot\mathbf{r}), \qquad (20)$$

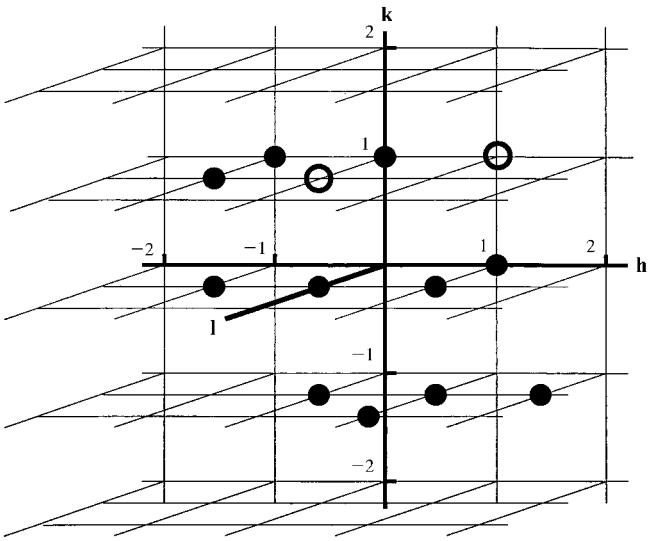

Figure 8
Schematic illustration of the GPD reciprocal lattice for the low-order reflections with $d_{hkl} > 40$ Å. The 110 and 011 reflections denoted by open circles were among the measured data, but the 11 reflections denoted by filled circles are missing from the measured data set.

are generally low-resolution terms, *i.e.* leading terms in the sequence $0 \le |\mathbf{h}| \le |\mathbf{h}|_{\rm max}$. On account of **h**-space and **r**-space reciprocity, the leading low-resolution terms carry most information on longer range structural ordering, solute/solvent molecular envelopes and molecular-packing patterns.

7. Concluding remarks

We have shown that probabilistic direct-methods phasing theory, originally developed from a starting hypothesis of

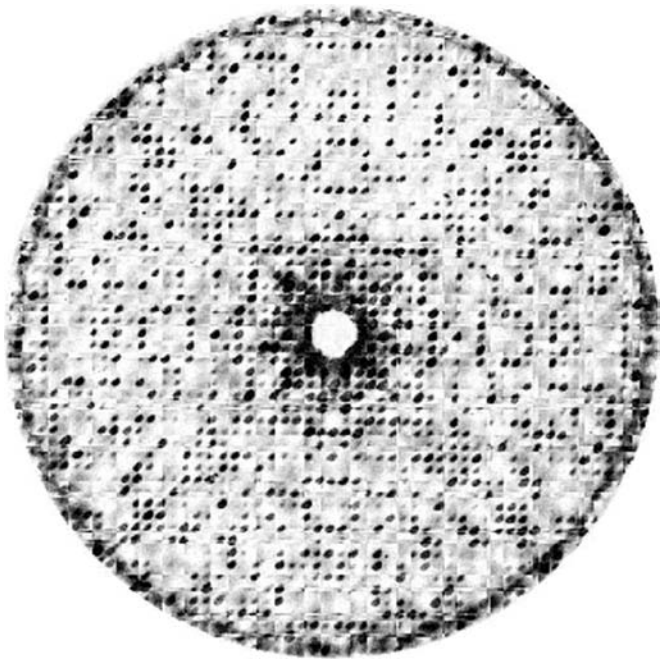

Figure 9
X-ray precession photograph corresponding to the *hk*0 reciprocal-lattice net of tetragonal lysozyme. Note the tendency toward local clustering of strong reflections.

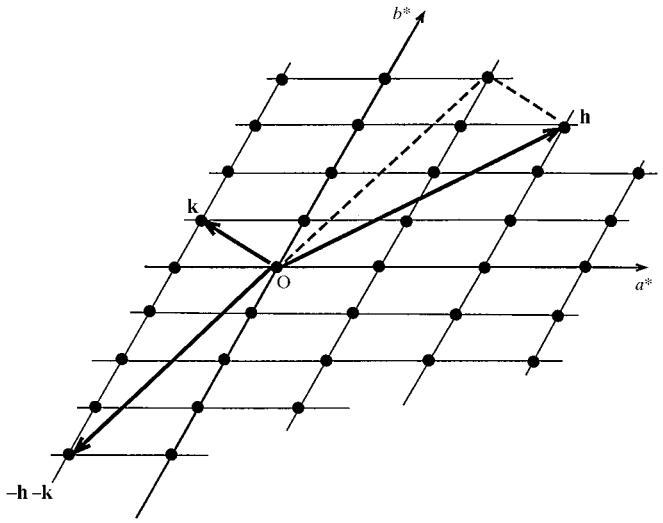

Figure 10
Schematic reciprocal-lattice vector diagram for a three-phase structure invariant $\varphi_{\mathbf{hk}} = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}$ that involves one reflection corresponding to a short reciprocal-lattice vector.

uniform random distributions of atomic positions, can be adapted simply to a non-uniform bulk-solvent-compensated globbic approximation for low-resolution protein crystal structures. The effective number n_g of atoms per polyatomic glob increases with decreasing resolution, so that low-resolution phasing involves positioning only $N_g = N_a/n_g$ globs rather than N_a atoms. Probabilistic estimation of low-resolution phases depends critically on completeness of the low-resolution data; if even a very few of the lowest order data are missing, average phase errors increase dramatically. This reinforces the lesson that it is worth taking extra care in diffraction experiments to ensure completeness of the low-resolution data.

All our conclusions from our calculations on glyceraldehyde phosphate dehydrogenase (GPD) have also been confirmed by parallel calculations on the Se_{30} selenomethionine form of *S*-adenosylhomocysteine hydrolase (SAH; see also Guo *et al.*, 2000*a*).

The work we have reported here is based on ideas on group scattering factors originated by Debye (1915) [as cited by Guinier (1994*a*) and Warren (1990)], Harker (1953), Main (1976) and Heinermann (1977) and ideas on bulk-solvent compensation developed by Moews & Kretsinger (1975), Wang (1985), Kostrewa (1997), Badger (1997) and Tronrud (1997). Our work has also been informed by the low-resolution uniform sphere model of Andersson & Hovmöller (1996), the few-atoms model of Podjarny *et al.* (1998) and the solvent-contrast method of Carter (1998).

APPENDIX *A*

Normalized crystal structure factors are defined by

$$E_{\mathbf{h}} = \frac{F_{\mathbf{h}}}{\langle |F|^2 \rangle_{\mathbf{h}}^{1/2}} = \sum_{j=1}^{N_a} \frac{f_{j\mathbf{h}}}{\alpha_{\mathbf{h}}^{1/2}} \exp(2\pi i \mathbf{h}\cdot\mathbf{r}_j) = |E_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}), \quad (21)$$

where $\langle |E|^2 \rangle_{\mathbf{h}} = 1$ since

$$\alpha_{\mathbf{h}} = \langle |F|^2 \rangle_{\mathbf{h}} = \sum_{j=1}^{N_a} f_{j\mathbf{h}}^2 \quad (22)$$

is the Wilson (1949) expectation value for the intensity of scattering by N_a atoms distributed independently and uniformly at position vectors $\mathbf{r}_j$ throughout the unit cell.

The derivations of Hauptman & Karle (1958) and Karle & Hauptman (1958) give joint and conditional probability distributions for the amplitudes $|E_{\mathbf{h}1}|$, $|E_{\mathbf{h}2}|$, $|E_{\mathbf{h}3}|$ and phases $\varphi_{\mathbf{h}1}, \varphi_{\mathbf{h}2}, \varphi_{\mathbf{h}3}$ for triplets of structure factors with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$. The joint probability density function is

$$\begin{aligned} P_J(R_1,R_2,R_3,\Phi_1,\Phi_2,\Phi_3) &= \frac{R_1R_2R_3}{(2\pi)^6} \int_0^\infty \mathrm{d}\rho_1 \int_0^\infty \mathrm{d}\rho_2 \int_0^\infty \mathrm{d}\rho_3 \int_0^{2\pi} \mathrm{d}\theta_1 \int_0^{2\pi} \mathrm{d}\theta_2 \int_0^{2\pi} \mathrm{d}\theta_3 \rho_1\rho_2\rho_3 \\ &\times \exp\left[-i\sum_{j=1}^{3} R_j\rho_j \cos(\theta_j - \Phi_j)\right] \prod_{k=1}^{N_a} q_k, \end{aligned} \quad (23)$$

where R_1, R_2, R_3 and Φ_1, Φ_2, Φ_3 denote the triplet amplitude and phase random variables corresponding to N_a uniformly randomly distributed atomic positions $r_{\mathbf{k}}$. The factors $q_{\mathbf{k}}$ in the continued product in (23) are statistical averages over the random positions and are given by

$$q_k = \left\langle \exp\left[i\sum_{j=1}^{3} (f_{k\mathbf{h}_j}/\alpha_{\mathbf{h}_j}^{1/2}) \rho_j \cos(2\pi\mathbf{h}_j\cdot\mathbf{r}_k - \theta_j)\right]\right\rangle_{\mathbf{r}_k \subseteq V_{\mathrm{cell}}}. \quad (24)$$

Using the notation

$$c_{kj} = (f_{k\mathbf{h}_j}/\alpha_{\mathbf{h}_j}^{1/2}) \rho_j \cos(2\pi\mathbf{h}_j\cdot\mathbf{r}_k - \theta_j), \quad (25)$$

series expansion gives the average of complex exponentials on the right of (24) as a sum of averages of cosine products,

$$\begin{aligned} q_k = 1 &+ i\sum_{l=1}^{3} \langle c_{kl}\rangle_{\mathbf{r}_k} + (i^2/2)\sum_{l=1}^{3}\sum_{m=1}^{3} \langle c_{kl}c_{km}\rangle_{\mathbf{r}_k} \\ &+ (i^3/3!)\sum_{l=1}^{3}\sum_{m=1}^{3}\sum_{n=1}^{3} \langle c_{kl}c_{km}c_{kn}\rangle_{\mathbf{r}_k} + \ldots, \end{aligned} \quad (26)$$

which, using

$$\cos x_1 \cos x_2 = \tfrac{1}{2}[\cos(x_1 + x_2) + \cos(x_1 - x_2)],$$

can be shown to be expressible in terms of averages of cosine and squared-cosine functions. Then, since by the uniform random atomic distribution hypothesis the function arguments span the whole interval $0 \le x \pmod{2\pi} < 2\pi$, the averages $\langle \cos x\rangle = 0$ and $\langle \cos^2 x\rangle = \frac{1}{2}$ allow simplification of (26), (24) and (23) to practically useful forms.

Specifying given values for the amplitudes $|E_{\mathbf{h}1}|$, $|E_{\mathbf{h}2}|$, $|E_{\mathbf{h}3}|$ with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$ leads from the joint distribution function (23) to the conditional distribution function,

$$\begin{aligned} P_C(\Phi_1+\Phi_2+\Phi_3|R_1,R_2,R_3) &= \frac{1}{2\pi I_0(A)} \exp[A\cos(\Phi_1+\Phi_2+\Phi_3)]. \end{aligned} \quad (27)$$

In customary notation, the conditional distribution function is written as

$$P_C(\Phi_{\mathbf{hk}}|A_{\mathbf{hk}}) = \frac{1}{2\pi I_0(A_{\mathbf{hk}})} \exp(A_{\mathbf{hk}}\cos\Phi_{\mathbf{hk}}), \quad (28)$$

where

$$\Phi_{\mathbf{hk}} = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}} \quad (29)$$

is the *triplet* or *three-phase structure invariant*, $I_n(\mathbf{x})$ denotes an nth-order modified Bessel function of the second kind and

$$A_{\mathbf{hk}} = 2\frac{\sum_{j=1}^{N_a} f_{j,\mathbf{h}} f_{j,\mathbf{k}} f_{j,-\mathbf{h}-\mathbf{k}}}{(\alpha_{\mathbf{h}}\alpha_{\mathbf{k}}\alpha_{-\mathbf{h}-\mathbf{k}})^{1/2}} |E_{\mathbf{h}}||E_{\mathbf{k}}||E_{-\mathbf{h}-\mathbf{k}}|, \quad (30)$$

or, for equal atoms,

$$A_{\mathbf{hk}} = (2/N_a^{1/2})|E_{\mathbf{h}}||E_{\mathbf{k}}||E_{-\mathbf{h}-\mathbf{k}}|. \quad (31)$$

The conditional distribution of the three-phase invariant gives expectation values

$$\begin{aligned} \langle y\rangle &= \int_{-\infty}^{+\infty} y(x)P(x)\,\mathrm{d}x \\ \langle \Phi_{\mathbf{hk}}\rangle &= 0 \end{aligned} \quad (32)$$

and

$$\langle \cos \Phi_{\mathbf{hk}} \rangle = I_1(A_{\mathbf{hk}})/I_0(A_{\mathbf{hk}}) \quad (33)$$

and leads to the tangent formula

$$\tan \varphi_{\mathbf{h}} \simeq \frac{\sum_{\mathbf{k}} |E_{\mathbf{k}}||E_{-\mathbf{h}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |E_{\mathbf{k}}||E_{-\mathbf{h}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}})}. \quad (34)$$

These now classical results have been cornerstones of direct-methods phasing procedures.

We thank Herbert Hauptman and Douglas Dorset for helpful discussions of the low-resolution phase problem; we are grateful for research support from USDHHS PHS NIH grant No. GM46733.

References

Andersson, K. M. & Hovmöller, S. (1996). *Acta Cryst.* D**52**, 1174–1180.

Badger, J. (1997). *Methods Enzymol.* **277**, 344–352.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1998). *Direct Methods for Solving Macromolecular Structures*, NATO ASI Series Volume, Series C: *Mathematical and Physical Sciences*, Vol. 507, edited by S. Fortier, pp. 47–71. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Carter, C. W. Jr (1998). *Direct Methods for Solving Macromolecular Structures*, NATO ASI Series Volume, Series C: *Mathematical and Physical Sciences*, Vol. 507, edited by S. Fortier, pp. 227–237. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Debye, P. (1915). *Ann. Phys. (Leipzig)*, **46**, 809–823.

Guinier, A. (1994*a*). *X-ray Diffraction in Crystals, Imperfect Crystals and Amorphous Bodies*, pp. 49–53. New York: Dover Publications, Inc.

Guinier, A. (1994*b*). *X-ray Diffraction in Crystals, Imperfect Crystals and Amorphous Bodies*, p. 359. New York: Dover Publications, Inc.

Guo, D. Y., Blessing, R. H. & Langs, D. A. (2000*a*). *Acta Cryst.* D**56**, 451–457.

Guo, D. Y., Blessing, R. H. & Langs, D. A. (2000*b*). *Acta Cryst.* D**56**, 936.

Guo, D. Y., Blessing, R. H., Langs, D. A. & Smith, G. D. (1999). *Acta Cryst.* D**55**, 230–237.

Harker, D. (1953). *Acta Cryst.* **6**, 731–736.

Hauptman, H. A. & Karle, J. (1958). *Acta Cryst.* **11**, 149–157.

Heinermann, J. J. L. (1977). *Acta Cryst.* A**33**, 100–106.

Karle, J. & Hauptman, H. A. (1958). *Acta Cryst.* **11**, 264–269.

Kostrewa, D. (1997). *CCP4 Newslett.* **34**, 9–22.

Main, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, K. Huml & B. Sedlácek, pp. 97–105. Copenhagen: Munksgaard.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

Murthy, M. R. N., Garavito, R. M., Johnson, J. E., Rossmann, M. G. (1980). *J. Mol. Biol.* **138**, 859.

Patterson, A. L. (1967). *International Tables for X-ray Crystallography*, Vol. II, p. 72. Birmingham, England: The Kynoch Press.

Podjarny, A., Urzhumtsev, A. & Lunin, V. (1998). *Direct Methods for Solving Macromolecular Structures*, NATO ASI Series Volume, Series C: *Mathematical and Physical Sciences*, Vol. 507, edited by S. Fortier, pp. 421–431. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes. The Art of Scientific Computing*, pp. 289–293. Cambridge University Press.

Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.

Warren, B. E. (1990). *X-ray Diffraction*, pp. 116–117. New York: Dover Publications, Inc.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–320.