Acta Cryst. (1997). A**53**, 445–455

The Pseudo-Atom Approach to Phase Determination in Protein Electron Crystallography – Noncentrosymmetric Projections

Douglas L. Dorset

Electron Diffraction Department, Hauptman–Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, New York 14203-1196, USA. E-mail: dorset@hwi.buffalo.edu

(Received 22 November 1996; accepted 3 March 1997)

Abstract

From an idea proposed by David Harker [*Acta Cryst.* (1953), **6**, 731–736], the assembly of globular subunits in a protein can be treated as pseudo-atoms for normalization of observed electron diffraction intensities. As demonstrated with published data from native or deoxycholate-treated bacteriorhodopsin, a multisolution approach *via* the Sayre–Hughes equation can then generate phase solutions to 6 Å resolution that compare quite favorably with those determined earlier by phase extension. The major problem in such determinations is identification of the best phase set, especially if no lower-resolution images of the protein are available. (However, 15 to 10 Å resolution image-derived phases could be used as a reference set to identify the correct solution.) A viable option may be to compare Patterson maps, calculated from trial map peak positions, to the experimental autocorrelation function. Trial phase determinations for the Omp F porin from *E. coli* outer membrane, on the other hand, are somewhat less successful because the β-sheet secondary structure is less well modeled by an array of 'globs'.

1. Introduction

In recent years, there has been increasing activity, exploring the possibilities of direct determination of crystallographic phases for diffraction data from proteins. Although the large number of atoms in the unit cell can severely limit the application of traditional probabilistic methods (Karle, 1989), there have been alternative approaches (Hauptman, 1993; Miller, DeTitta, Jones, Langs, Weeks & Hauptman, 1993), where a nearly correct phase solution in a multiple set can be identified to lie near a previously identified optimal figure of merit so that these phases can be improved by annealing to determine the correct crystal structure (Weeks, Hauptman, Smith, Blessing, Teeter & Miller, 1995). Thus, given experimental X-ray intensities from a protein (*e.g.* containing up to 800 atoms) measured to near 1.0 Å resolution, there is now good reason to expect a favorable structure determination without the need to prepare heavy-atom derivatives or reliance on anomalous-scattering information.

Currently, lower-resolution data sets cannot be assigned phase values by this procedure. However, there are reasons to believe that, at least within the 5 to 6 Å diffraction resolution limit, the intensities themselves also should be amenable to direct analysis. As pointed out by Fan, Hao & Woolfson (1991), the low-angle region of the diffraction pattern contains the most intense reflections. Although there are fewer phase relationships per reflection than for a small-molecule structure, it has been argued by these authors that there should be no change in their quality for low-resolution protein data. Previous explorations of conventional probabilistic direct methods in protein X-ray crystallography, starting with initial information, *e.g.* from an incomplete MIR set, have demonstrated that these phase relationships can be very effective for completing the phase set (Reeke & Lipscomb, 1969; Podjarny, Schevitz & Digler, 1981). The same result has been noted in electron crystallography, where lower-resolution information from the Fourier transform of an electron micrograph has been extended to the limit of the electron diffraction pattern (Gilmore, Shankland & Fryer, 1993; Dorset, Kopp, Fryer & Tivol, 1995; Dorset, 1996). Within the sampling limits of goniometry, an actual advantage of electron crystallography over X-ray crystallography is that, for macromolecules, all of the diffracted intensities are collected, including those in the very low angle region often occluded by beam stops in X-ray measurements.

In electron crystallography, true *ab initio* phase determinations at low resolution have also been attempted. After building up a trial basis set, the Sayre equation and/or maximum entropy and likelihood have been found to be somewhat effective for electron diffraction intensities, the former when coupled with a phase-annealing process (Dorset, 1995*a*) and the latter when likelihood prediction was used to prune the phasing tree (Gilmore, Nicholson & Dorset, 1996). In X-ray crystallography, known information about the protein structure has been incorporated into the phase determination by constraining a match to a density histogram (Lunin, Urzhumetsev & Skovoroda, 1990).

© 1997 International Union of Crystallography
Printed in Great Britain – all rights reserved

Alternatively, trial structures have been generated using random spherical density generators followed by a calculation of structure-factor amplitudes to match to the observed data set (Lunin, Lunina, Petrova, Vernoslova, Urzhumtsev & Podjarny, 1995; Andersson & Hovmöller, 1996).

As early as 1953, Harker proposed that the Fourier transform of globular density units in proteins would be the optimal means for normalizing the low-resolution intensity data. By inference, it is clear also that the use of pseudo-atom transforms might also affect how a protein structure could be determined by direct methods at low resolution. For example, as is shown in another publication (Dorset, 1997), the 6 Å centrosymmetric projected electron diffraction data set from the integral membrane protein halorhodopsin reduces to a small-molecule problem, easily solved by symbolic addition after these data are adjusted for the fall-off in the Fourier transform of an average Gaussian glob, corresponding approximately to the cross section of an α-helix. The appropriateness of this approximation is further evaluated in this communication, considering the case of non-centrosymmetric projections, as well as density distributions (*e.g.* β-barrels) that may not correspond so closely to the atomistic model.

2. Electron diffraction data

2.1. *Native bacteriorhodopsin*

Electron diffraction amplitudes and image-derived crystallographic phases from glucose-embedded bacteriorhodopsin have been published to 3.5 Å resolution (Henderson, Baldwin, Downing, Lepault & Zemlin, 1986). In the [001] projection, the protein crystallizes in plane group $p3$, where $a = 62.4$ Å. The most reliable phase information from image Fourier transforms, however, was measured out to 6 Å, the resolution cut-off used in this study. [This corresponds approximately to the resolution limit reliably extended from a 10 Å image-derived basis set by the Sayre equation (Dorset, Kopp, Fryer & Tivol, 1995) and is near the minimum of $\langle I_{\rm obs}\rangle$ *vs* $\sin\theta/\lambda$ found for many proteins.] The layer structure at this resolution contains the characteristic assembly of seven α-helices in the asymmetric unit (Fig. 1*a*).

2.2. *Deoxycholate-treated bacteriorhodopsin*

If the purple membranes from *Halobacterium halobium* are extracted with sodium deoxycholate, the unit-cell axis of the bacteriorhodopsin two-dimensional crystals is found to shrink to $a = 57.3$ Å, while retaining the $p3$ symmetry. Electron diffraction amplitudes and image-derived phases from glucose-embedded preparations were used by Glaeser, Jubb & Henderson (1985) to determine the delipidized crystal structure to 6 Å resolution and these data were used as a basis for the direct analysis described below. As shown in Fig. 1(*b*), there is a close resemblance of the protein structure to that of the native membrane form (Fig. 1*a*).

2.3. *Omp F porin*

Two crystalline forms of Omp F porin, reconstituted in phospholipid bilayers from the outer membrane of *Escherichia coli* and then glucose embedded, had been investigated by cryoelectron microscopy and Fourier transforms of experimental micrographs yielded phase information to 3.2 Å resolution (Sass, Büldt, Beckmann, Zemlin, Van Heel, Zeitler, Rosenbusch, Dorset & Massalski, 1989). Diffraction amplitude data from the double-layer form, involving stacking of two membrane layers in plane group $p31m$, $a = 72.0$ Å, were shown to become weak beyond a resolution limit of 6 Å (Dorset, 1996), so that only this region was used for direct phase determination. Unlike the other two membrane protein projections used in this study, the porin is primarily composed of a β-barrel (Fig. 1*c*).

3. Structure analysis

3.1. *Normalization of data*

The premise of the analytical method described in this paper is that local density distributions in the projected structure, *i.e.* globular cross sections, can be simulated as pseudo-atoms. Unlike the original approach taken by Harker (1953), where these globs were assumed to be spherical, a Gaussian density profile was postulated instead, merely because its Fourier transform (FT), also Gaussian, is well behaved (Gaskill, 1978) [*i.e.* no 'ringing' effect found in the sinc(u) function in the FT of a sphere cross section]. Furthermore, it was also assumed that the cell edges could be re-scaled to 1/10 their original size so that the electron scattering-factor curve for carbon could serve as an approximation for the Gaussian glob transform. [As pointed out earlier (Dorset, 1997), the tenfold factor is justified by comparing the 15 Å center-to-center distance for two touching α-helices to the 1.54 Å C—C single-bond distance.] Thus, with re-scaling, for a unit-cell edge from *e.g.* 62 to 6.2 Å, the resolution of the data set would be transformed from 6 to 0.6 Å for purposes of the simulation. Approximation of a Gaussian function by $f_{\rm C}$ is partially justified by its fit by a sum of weighted Gaussians (Doyle & Turner, 1968) but, strictly speaking, its shape is more Lorentzian in character [so that its Fourier transform is actually an exponential function rather than a Gaussian (Champeney, 1963)].

The scattering factors were then used to calculate Wilson (1949) plots from the observed intensity data from all three proteins. In all cases $B \simeq 0.0$ Å^2, indicating that the fall-off of diffracted intensity is matched reasonably well by the scattering-factor approximation without need for further shape adjustments.

From the normalized intensity values, $|E_h|$ magnitudes were calculated from $|E_h|^2 = I_h^{obs}/\varepsilon f_C$, again where f_C is the carbon scattering factor and the statistical weight ε compensates for special classes of reflections (for the examples considered, important only for plane group $p31m$). As usual (Karle & Hauptman, 1956), the $|E_h|$ were scaled such that $\langle|E_h|^2\rangle = 1.000$. An E_{000} value was also estimated from the number of glob sites in the unit cell.

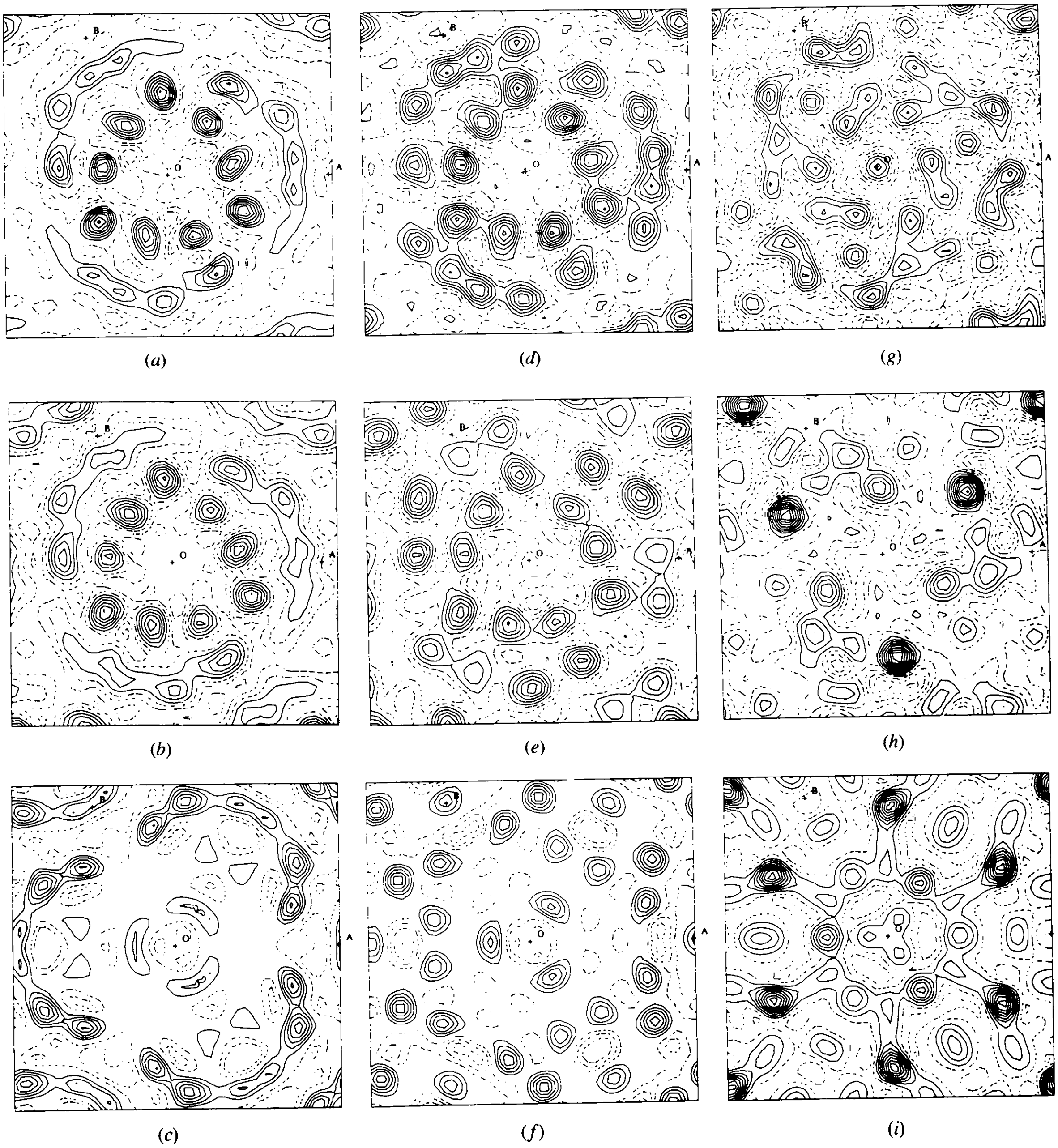


Fig. 1. Potential maps (6 Å resolution) for membrane proteins: situation 1, crystallographic phases from original images: (*a*) native bacteriorhodopsin, (*b*) deoxycholate-treated bacteriorhodopsin, (*c*) Omp F porin; situation 2, phases from atomistic approximations of structure *via* structure-factor calculation: (*d*) native bacteriorhodopsin, (*e*) deoxycholate-treated bacteriorhodopsin, (*f*) Omp F porin; situation 3, phases from Sayre extension of 15 Å resolution phases: (*g*) native bacteriorhodopsin, (*h*) deoxycholate-treated bacteriorhodopsin, (*i*) Omp F porin.

Table 1. *Success of atomistic approximation for simulating low-resolution structure factors from membrane proteins*

Protein (no. of phases)	Mean phase deviation (°)	*R* factor
Bacteriorhodopsin (native) (50)	37	0.48
Bacteriorhodopsin (deoxycholate-treated) (35)	29	0.33
Omp F porin (42)	46	0.41

3.2. *Direct phase determination*

Three-phase structure invariants (Hauptman, 1972) of the Σ_2 type were then generated for the *hk*0 data sets, *i.e.* phase relationships where $\varphi_h = \varphi_k + \varphi_{h-k}$, where $h = h_1k_1l_1$ and $k = h_2k_2l_2$. Next, a convergence procedure (Germain, Main & Woolfson, 1970) determined how each set of reflections could be assigned phases most efficiently from the smallest number of reflections having large $|E_h|$ values. For the two structures in *p*3, three reflections were required. Since every *hk*0 reflection is a structure invariant (Rogers, 1980), these were assigned algebraic values $a = 45, 135°$; $b, c = \pm 45, \pm 135°$ to generate 32 trial permutations. (Inclusion of the $a = -45, -135°$ permutations generates 32 additional enantiomorph sets.) For plane group *p*31*m*, two large $|E_h|$ *h*00 reflections had centrosymmetric values so that the tests, involving four unknowns, were made on $a, b = 0, 180°$; $c = 45, 135°$; $d = \pm 45, \pm 135°$, since each *hk*0 reflection is again invariant (Rogers, 1980).

The permuted algebraic phases were then used as input sets for the Sayre–Hughes equation (Sayre, 1980): $E_h = (1/N)\langle E_k E_{h-k}\rangle$, with the E_{000} term as defined above. After expanding the basis to a new phase set, some method was needed to determine the most probable solution from the calculated test maps based on the phased values of $|E_h|$. This was, in fact, not a straightforward process. In principle, using arguments of Hoppe, Gassmann & Zechmeister (1970), there must be some criterion for finding a 'best' density distribution. Previously, it has been shown (Dorset, 1996) that the Luzzati criterion (Luzzati, Tardieau & Taupin, 1972; Luzzati, Mariani & Delacroix, 1988) of density flatness $\langle \Delta\rho^4\rangle$ is not totally satisfactory for low-resolution macromolecular determinations, particularly for projections. Since a pseudo-atom model is being evaluated, the Stanley (1986) criterion of $\sum_i \rho_i^n$ was also tested, *i.e.* the values $n = 4, 5$ were monitored for each map. As a default criterion, it was assumed that a relatively low resolution image of the two-dimensional crystals could be recorded in the electron microscope and that its transform, yielding phase values to 15 Å, could serve as an independent test for phases also determined in this region by direct methods. [Although, in principle, predicted values of $|E_h|$ could be compared to observed magnitudes, previous studies (Dorset, 1995*a*) have shown that the *R* factor based on these normalized magnitudes is not very reliable at low resolution.]

Trial phase sets from a smaller number of solutions were then generated by another cycle of the Sayre–Hughes equation. After identification of pseudo-atomic positions, a structure-factor calculation was carried out, again *via* re-scaling and employment of the carbon scattering factor as the approximation to the glob Fourier transform. This initiated a Fourier refinement for improvement of the phase set.

In no case were any other techniques, *e.g.* density modification (Wang, 1985), used to improve the initial phase set. The sole object of this study was to test how well a pseudo-atom approach would serve to find a useful starting point for a low-resolution structure analysis.

4. Results

4.1. *Success of pseudo-atom approximation*

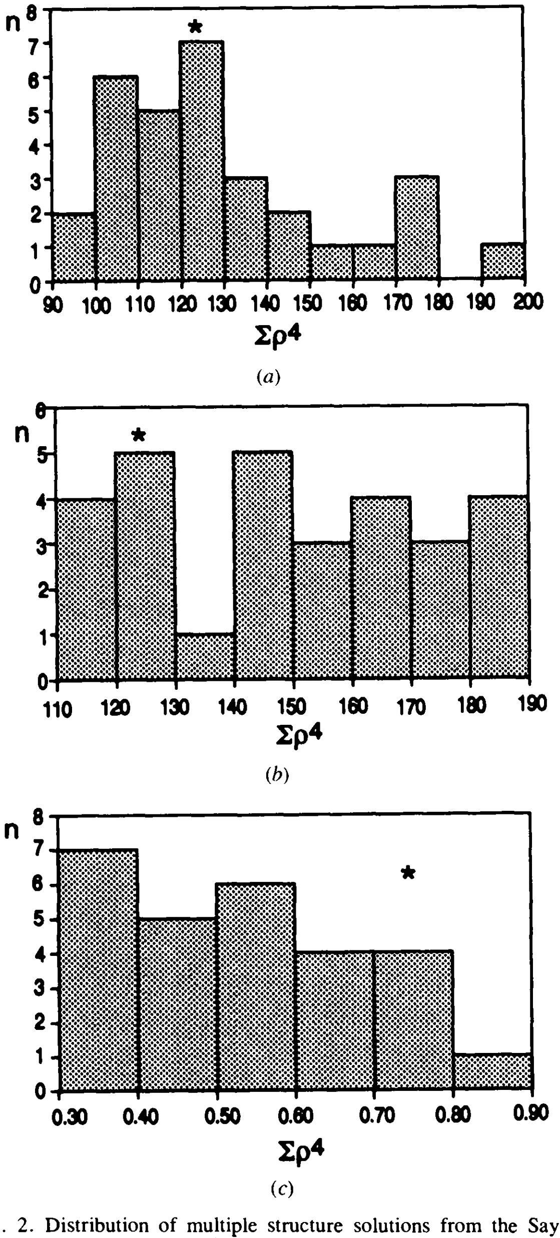
The simulation of the glob Fourier transform by a carbon scattering factor, after a dimensional re-scaling, was found to give an acceptable match to the phase set in all cases, even if the diffraction amplitudes were not accurately simulated by the structure-factor calculation (Table 1). In the case of native bacteriorhodopsin, eight pseudo-atom positions were chosen initially for calculation of structure factors (*i.e.* one tilted helix was initially sampled by two globs) and the generated phases produce a map (Fig. 1*d*) that could be compared favorably to the density distribution found experimentally (Fig. 1*a*). A similar comparison could be made (Figs. 1*e* and *b*) for the deoxycholate-treated protein, where seven peak positions were used to calculate structure factors. The simulation was not so favorable for the Omp F porin, however. Five density maxima were chosen for the structure-factor calculation and these were returned in the ensuing potential map generated from these phases (Figs. 1*f* and *c*). However, there was no continuity of the generated map that would suggest a β-sheet substructure, despite the rather good agreement between calculated and experimental phases.

4.2. *Native bacteriorhodopsin*

After generating 360 Σ_2 triples from 50 unique $|E_h|$ terms ($A_{\rm min} = 0.2$), algebraic values were assigned to the phases, $\varphi_{430} = a$, $\varphi_{350} = b$, $\varphi_{170} = c$, as described above, to find 32 trial solutions. One cycle of the Sayre–Hughes equation produced a total of 11 phases from which *E* maps were generated using $E_{000} = (21)^{1/2}$, assuming seven 'atoms' to be present in the asymmetric unit. (Note the two helix areas originally considered to be independent were combined.) The distribution of maps in terms of the values for $\sum_i \rho_i^4$ is given in Fig. 2(*a*). The best solution ($a = 135$, $b = -45$, $c = 135°$) was contained in a cluster near the lower end of the scale, narrowing the choices to seven. Tests of overlapping phases (three reflections) in the generated sets against a 15 Å image restricted this choice to three. Expansion of these phase sets by another convolution cycle allowed one set to be chosen by the match to the low-resolution image set (nine reflections).

Table 2. *Phase determination (mean deviation) for native bacteriorhodopsin (°)*

	All 50 data	Most intense data (18 reflections)
Structure factors from first Sayre model	63.3	40.9
Second Sayre extension	64.2	45.9
Fourier refinement 1	61.5	41.6
Fourier refinement 2	65.2	38.4
Sayre extension of 15 Å image phases	63.2	52.9



(*a*)

(*b*)

(*c*)

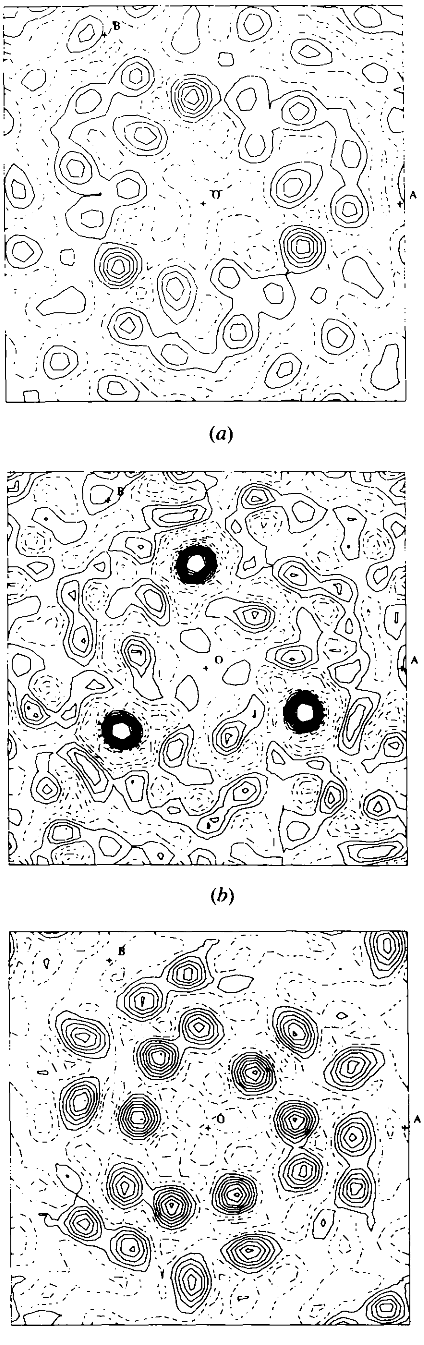
Fig. 2. Distribution of multiple structure solutions from the Sayre equation in terms of $\sum_i \rho_i^4$ (maps calculated with E_h including E_{000}): (*a*) native bacteriorhodopsin, (*b*) deoxycholate-treated bacteriorhodopsin, (*c*) Omp F porin.

The complete phase set from two possible starting points was then monitored. First, the best solution found from the first, incomplete, Sayre expansion of a basis set could be interpreted in terms of possible helical positions (Fig. 3*a*). [Although the exact features of this structure would not be known in a true *ab initio* determination, most of the peaks in this figure correspond to true helix sites depicted in Fig. 1(*a*)]. When these were used to calculate structure factors (*via* the re-scaling process described above), the mean phase agreement to the previous image-derived phases could be calculated (Table 2). A nearly equivalent solution was found when the Sayre-derived phases were used for another convolution cycle. Several attempts were made to improve this structure from this point. The best procedure started with helix coordinates from the map calculated with phased $|E_h|$ values (Fig. 3*b*). After two cycles of Fourier refinement, there was a significant improvement for phases of the most intense reflections, yielding the map in Fig. 3(*c*). A comparison of helix positions from this map to the ones of the ideal structure is given in Table 3. The mean difference in positions is 1.9 Å.

4.3. *Deoxycholate-treated bacteriorhodopsin*

A total of 271 Σ_2 triples ($A_{\min} = 0.2$) were generated from 35 unique reflections in order to find the most optimal assignment of phases from the least number of reflections with large $|E_h|$ magnitudes. Algebraic values were then assigned to three reflections, $\varphi_{420} = a$, $\varphi_{430} = b$, $\varphi_{350} = c$, following the scheme given above, and these were used to generate 32 trial solutions by one cycle of the Sayre–Hughes equation (using the same value for E_{000} as before). After generating trial E maps from the resultant set of ten unique reflections, the distributions of $\sum_i \rho_i^n$ were then plotted, the grouping of the $n = 4$ maps is given in Fig. 2(*b*). Five possible solutions were chosen at the lower end of the scale, corresponding to the group with the largest number of examples. Within this subset, there was only one solution that also corresponded to the best agreement with 15 Å resolution image-derived phases (three overlapping reflections) and that occurred when $a = 135$,

$b = 135$, $c = -45°$. Although some of the helix positions were evident in the potential map calculated from this small phase set [compare Fig. 4(*a*) to Fig. 1(*b*)], a second cycle of the Sayre convolution was carried out, giving the phase agreement shown in Table 4 and the E map shown in Fig. 4(*b*). Helix positions obtained from this map were then used to start two cycles of Fourier refinement, with the increase in phase accuracy (especially for the most intense reflections) reviewed in Table 4. The final potential map is shown in Fig. 4(*c*). Peak positions from it are compared to the helix locations of the ideal structure in Table 3. The mean difference is 1.6 Å.



(*a*)

(*b*)

(*c*)

Fig. 3. Phase determination of native bacteriorhodopsin, potential maps: (*a*) initial phase set from first Sayre expansion, (*b*) E_h map (phases from second Sayre expansion), (*c*) F_h map after Fourier refinement.

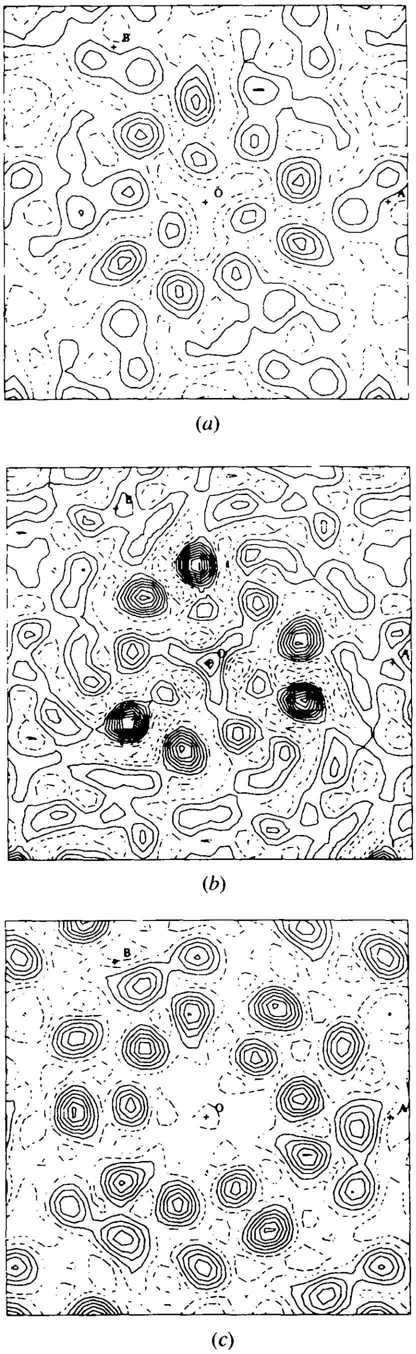
Table 3. *Major pseudo-atom density centers in structure determinations*

	Actual structure (image phases)		After Fourier refinement	
	x	y	x	y
Native bacteriorhodopsin				
	0.359	0.334	0.372	0.299
	0.458	0.213	0.460	0.170
	0.393	0.000	0.359	−0.027
	0.255	−0.200	0.292	−0.171
	0.164	−0.136	0.170	−0.126
	0.218	0.037	0.228	0.017
	0.218	0.190	0.200	0.158
Deoxycholate-treated bacteriorhodopsin				
	0.389	0.361	0.365	0.353
	0.507	0.245	0.493	0.223
	0.434	0.011	0.400	−0.026
	0.349	−0.159	0.287	−0.230
	0.199	−0.137	0.185	−0.128
	0.245	0.047	0.280	0.057
	0.237	0.207	0.239	0.201
Omp F porin				
	0.122	0.122	0.117	0.117
	0.428	0.131	0.467	0.144
	0.337	0.251	0.295	0.189
	0.521	0.292	0.538	0.281
	0.491	0.491	0.482	0.482

4.4. *Omp F porin*

From 42 unique data, 291 Σ_2 triples were generated to find a phasing sequence *via* the convergence method. Algebraic terms were assigned, therefore, to four strong reflections, $\varphi_{400} = a$, $\varphi_{500} = b$, $\varphi_{630} = c$, $\varphi_{330} = d$, the former two axial maxima having centrosymmetric values. These were then expanded by the Sayre–Hughes equation into 27 unique phases. [The value assumed for E_{000} was 3.56, *i.e.* $(12)^{1/2}$.] Using the values of $\sum_i \rho_i^n$, $n = 4, 5$, to evaluate the 32 generated E maps, it was not possible (Fig 2*c*) to find a cluster of solutions that satisfied criteria similar to the other proteins investigated in this study. On the other hand, comparison of trial phase sets with values from an assumed 15 Å resolution image (five overlapping values) readily identified the best phase set (generated when $a = 0, b = 180, c = 45, d = -135°$). A second convolution was carried out with this basis to give the phase agreement indicated in Table 5.

The potential map found from after the second cycle of the Sayre–Hughes equation, shown in Fig. 5(*a*), was not appreciably different from the one calculated from phased E values. If five unique density sites were identified and used for Fourier refinement, the phase accuracy did not improve (Table 5). Nevertheless, the potential maps (Fig. 5*b*) resemble the one shown in Fig. 1(*f*). A comparison of major peak positions from this map with major density centers in Fig. 1(*c*) is given in Table 3. The mean difference is 1.4 Å.



(*a*)

(*b*)

(*c*)

Fig. 4. Phase determination of deoxycholate-treated bacteriorhodopsin, potential maps: (*a*) initial phase set from first Sayre expansion, (*b*) E_h map after second Sayre expansion, (*c*) F_h map after Fourier refinement.

Table 4. *Phase determination (mean deviation) for deoxycholate-treated bacteriorhodopsin (°)*

	All 35 data	Most intense data (14 reflections)
From second Sayre expansion	61.1	38.6
Fourier refinement 1	50.3	24.0
Fourier refinement 2	53.8	22.9
Sayre expansion of 15 Å image phases	78.7	92.6

Table 5. *Phase determination (mean deviation) for Omp F porin (°)*

	All 42 data	Most intense data (15 reflections)
From second Sayre expansion	62.0	53.5
Fourier refinement 1	69.2	61.9
Fourier refinement 2	76.4	65.9
Sayre expansion of 15 Å image phases	72.2	72.3

4.5. *Improvement of scattering factors*

Attempts were made to improve the fit of the amplitude transform of the glob model to the observed structure-factor amplitudes for the two bacteriorhodopsin structures. While the f_C approximation was meant to approximate a Gaussian function, it is really closer to Lorentzian shape (Doyle & Turner, 1968). Thus, various Gaussian and Lorentzian functions were evaluated to improve the phenomenological scattering factor with the intent to match the fall-off of $\langle|F_{\rm obs}|\rangle$ *vs* $\sin\varphi/\lambda$. This was, in fact, a difficult task, since the average of $|F_{\rm obs}|$ within overlapping resolution shells did not produce a smooth sampling of the glob scattering envelope. This was because of a strong influence of a 'molecular transform' of glob clusters (*i.e.* interference between scattering centers).

For native bacteriorhodopsin, a trial Gaussian function could lower the R factor, either with an eight-atom (0.38) or a seven-atom (0.44) model (compare to Table 1). Similar improvements could be found with various Lorentzian functions. Equivalent R values to the value in Table 1 could be found for the deoxycholate-treated protein when a Lorentzian scattering factor was used. Because of the inadequacy of the glob approximation, such fits were not attempted for the Omp F porin.

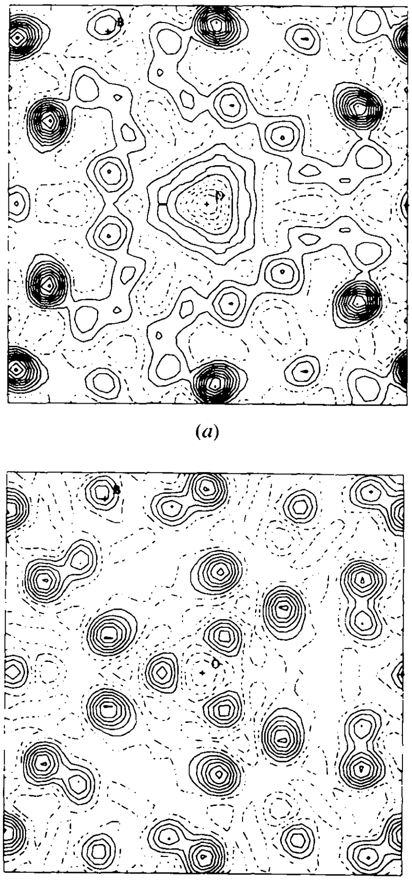
4.6. *Evaluation of the phase determination and figures of merit*

Table 6. *Deoxycholate-treated bacteriorhodopsin phases (°) after symbolic addition, assuming that $\varphi_{420} = 135$, $\varphi_{430} = 135$, $\varphi_{350} = -45°$ are known*

$hk0$	$\|F_{\rm obs}\|$	φ (symbolic addition)	φ (image transform)
400	52.0	0	−52
700	27.0	−225	−154
410	45.0	−90	−41
610	27.0	−45	−75
320	46.0	−45	11
420	50.0	135	122
330	42.0	90	132
430	45.0	135	143
630	17.0	−90	3
140	59.0	−90	−87
240	45.0	−90	−157
250	26.0	0	−75
350	31.0	−45	−44
450	15.0	−45	90
160	24.0	135	162
260	19.0	−45	169
170	19.0	−135	98

Since the best results were obtained from the deoxycholate-treated bacteriorhodopsin data, the accuracy of the phase determination itself was evaluated further in terms of the pseudo-atomistic model. After the structure-factor magnitudes had been normalized with f_C and the assembly of Σ_2 invariants, as described above, the triples were then listed in descending order of $A = (2/N^{1/2})|E_hE_kE_{h+k}|$. In the list of the 19 largest A values (down to $A = 0.84$), the mean average value of $\psi = \varphi_h + \varphi_k + \varphi_{-h-k}$ was found to be $51 \pm 32°$ (expected value 0°), assuming that the true (*i.e.* image-transform) values for the algebraic terms defined above known *a priori*. This appraisal, as well as the earlier analysis of halorhodopsin (Dorset, 1997), support the statement made by Fan, Hao & Woolfson (1991) that the phase relationships themselves should be reliable in this low-angle region. If a phase solution *via* the algebraic unknowns was then attempted by 'symbolic addition', $\langle|\Delta\varphi|\rangle = 54°$ for 17 reflections when the nearly correct test values ($a = 135$, $b = 135$, $c = -45°$) were chosen. Most of the error was found in the weaker reflections (Table 6), as demonstrated by the appearance of the potential map calculated from this partial phase set (Fig. 6). To justify this solution experimentally with phases from reflections present in both sets, an image of the protein crystals would have to be recorded to at least 10 Å resolution.

However, there may be other options to be used as figures of merit (FOM) so that the correct structure can be identified in a multiple set just from the diffraction data. If the ten reflections phased from the first Sayre convolution were used to calculate 32 trial potential maps, the peaks from these could then be used for a subsequent structure-factor calculation. There were two possible FOM's, therefore, for structure identification. If the approximate scattering factor were good enough,



(*a*)

(*b*)

Fig. 5. Phase determination of Omp F porin, potential maps: (*a*) F_h map after second Sayre expansion, (*b*) F_h map after Fourier refinement.

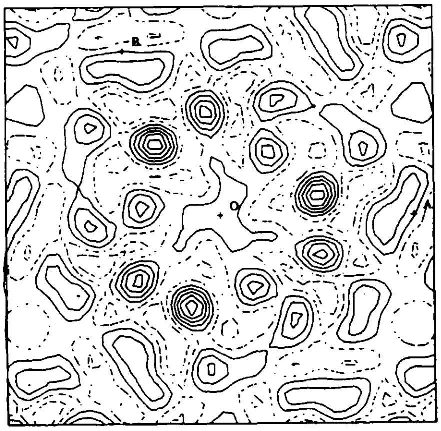


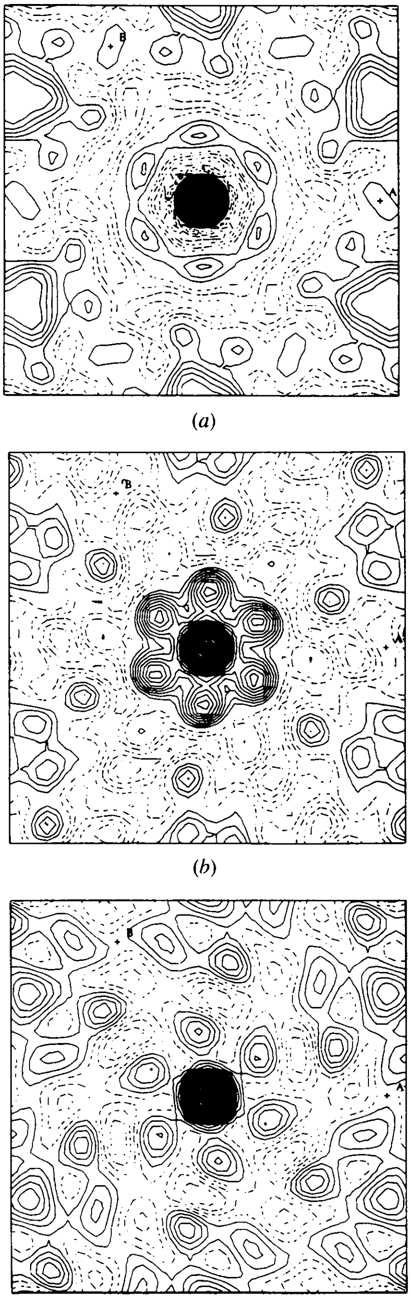
Fig. 6. Potential map for deoxycholate-treated bacteriorhodopsin from partial phase set found by symbolic addition (Table 6).

perhaps the R factor would suffice. If not, the complete I_{calc} list could then be transformed to a calculated Patterson function to compare to the one (Fig. 7*a*) determined from I_{obs}. Use of just the R factor to distinguish the correct structure proved to be completely worthless (even though it was of some use for Fourier refinement). On the other hand, visual match of Patterson functions narrowed the solution choices down to two (or possibly one if the flatness of the map was also considered) – one of which corresponded to the most correct phase set (Fig. 7*b*). [This comparison could be quantitated with Patterson correlation coefficients (*e.g.* see Beurskens & Smykalla (1991).] The identified solution could then be expanded as described above.

5. Discussion

From the results presented above, it is clear that an atomistic approach to solving protein crystal structures in projection is best suited to cases where the density distribution of subunits is well modeled by an assembly of globs. Halorhodopsin [considered in a previous study (Dorset, 1997)] and two forms of bacteriorhodopsin provide views down the axes of α-helices, which, on rescaling, might just as well be treated as 'atoms' to a good first approximation. The Omp F porin structure, on the other hand, because it is mostly composed of a β-sheet, can only be visualized as the loci of maximal density but their continuity in the projected barrel structure is not apparent after the analysis. (However, another difficulty with this analysis may be that only the amplitudes of the image transform were used rather than structure-factor magnitudes measured from electron diffraction patterns.)

The enticement of using multisolution methods in protein electron crystallography is that information from an electron microscope image recorded at modest resolution might also be obtained with very little difficulty. It is instructive in this context to compare the direct phasing results given above to those found if just the 15 Å image phases are extended to the 6 Å diffraction limit by the Sayre equation [using E values, which have been found to provide the most accurate phase prediction in earlier work (Dorset, 1995*a*,*b*)]. The phase accuracy for native bacteriorhodopsin is given in Table 2. While the overall mean error for the complete phase set, after extension, does not appear to be very different from the one found in this *ab initio* study (*i.e.* where 15 Å phases were used only to identify the correct solution), it is apparent from the projected potential map (Fig. 1*g*) that the correct structure could not be generated by phase extension from this lower resolution. Therefore, the direct determination gives much more accurate phase estimates for the most intense reflections than does the phase extension from a 15 Å basis. [However, note that another phase extension, based on maximum-entropy and likelihood methods (Gilmore, Shankland & Fryer, 1993) had been much more successful from this lower-resolution starting point. Also, it is clear that there are optimal starting resolutions for phase extension *via* the Sayre equation (Dorset, 1996).] Comparison of phase devia-


(*a*)

(*b*)

(*c*)

Fig. 7. Patterson functions for deoxycholate-treated bacteriorhodopsin: (*a*) calculated from I_{obs}; (*b*) calculated from I_{calc} when 'atoms' from $a = 135, b = 135, c = -45°$ map are used for a structure-factor calculation; (*c*) as in (*b*), but for $a = 45, b = 45, c = -45°$.

tions after extension of the 15 Å resolution image phases to those from the direct determination also is far worse for deoxycholate-treated bacteriorhodopsin (Table 4) or Omp F porin (Table 5), as revealed by the projected potential maps [Figs. 1(*h*), (*i*), respectively]. A similar comparison has also been made for halorhodopsin in favor of the pseudo-atom approach to phase determination (D. L. Dorset, unpublished data).

Thus, when the atomistic model is a suitable approximation, *ab initio* phase determination itself seems to be sufficiently accurate for generating a nearly correct structure, particularly when the initial solution is optimized by Fourier refinement. There is, of course, much room for further improvement of the phase set from a variety of approaches including solvent flattening (Wang, 1985), histogram matching (Zhang & Main, 1990), as well as maximum-likelihood tests (Gilmore, Nicholson & Dorset, 1996).

The weakness of the current methodology seems to be the lack of a robust figure of merit to identify the best solution after generation of numerous trials, especially if only diffraction data are used. It is already known that standard FOM's employed in small-molecule direct phasing are not reliable, as discussed earlier (Fan, Hao & Woolfson, 1991). Obviously, there have to be optimal constraints imposed on accepting trial potential maps calculated from incomplete phase sets. While the concepts of map smoothness and flatness have been shown to be insufficient in themselves for structure identification (Dorset, 1996), it is clear from Fig. 2 that correct solutions also cannot have a density distribution that is too 'peaky'. Part of the problem may be the breakdown of these criteria as absolute indicators for structure projections, a criticism that can be made of the Cochran (1952) condition at atomic resolution. On the other hand, at atomic resolution, prior knowledge of chemical composition and geometry is an additional criterion that can be advantageous for identification of a correct structure - but this foreknowledge is not available for an unknown protein structure at low resolution. The demands placed on protein structure determination at low resolution, therefore, are actually more stringent than for the small-molecule case. For future work, other approaches, such as maximum-likelihood predictions (Gilmore, Nicholson & Dorset, 1996) may be more successful than traditionally employed FOM's. Alternatively, as demonstrated in the above trial, the comparison of calculated Patterson functions to the observed autocorrelation function seems to be an option worthy of further exploration.

In summary, the pseudo-atom approach to determining projected protein structures at low resolution seems to be suitably valid as long the protein itself contains a substructure that can be well modeled by globs. The most probable phase-invariant relationships are reasonably accurate to lead to a useful solution. Although better approximates to the glob scattering factors than a re-scaled f_C curve can be found, they still do not permit the determination to be completed with the same accuracy found in small-molecule crystallography, partially because the exact nature of an unidentified globular subunit is unknown *a priori* - in terms of both its breadth and its ellipticity.

Research was supported by a grant from the National Institute of General Medical Science, GM-46733, which is gratefully acknowledged.

References

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

Beurskens, P. T. & Smykalla, C. (1991). *Direct Methods of Solving Crystal Structures*, edited by H. Schenk, pp. 281-290. New York: Plenum.

Champeney, D. C. (1963). *Fourier transforms and their Physical Applications*, p. 22. London: Academic.

Cochran, W. (1952). *Acta Cryst.* **5**, 65-67.

Dorset, D. L. (1995*a*). *Proc. Natl Acad. Sci. USA*, **92**, 10074-10078.

Dorset, D. L. (1995*b*). *Micron*, **26**, 511-520.

Dorset, D. L. (1996). *Acta Cryst.* A**52**, 480-489.

Dorset, D. L. (1997). *Proc. Natl Acad Sci. USA*, **94**, 1791-1794.

Dorset, D. L., Kopp, S., Fryer, J. R. & Tivol, W. F. (1995). *Ultramicroscopy*, **57**, 59-89.

Doyle, P. A. & Turner, P. S. (1968). *Acta Cryst.* A**24**, 390-397.

Fan, H. F., Hao, Q. & Woolfson, M. M. (1991). *Z. Kristallogr.* **197**, 197-208.

Gaskill, J. D. (1978). *Linear Systems, Fourier Transforms, and Optics*, pp. 179-208. New York: Wiley.

Germain, G., Main, P. & Woolfson, M. M. (1970). *Acta Cryst.* B**26**, 274-285.

Gilmore, C. J., Nicholson, W. V. & Dorset, D. L. (1996). *Acta Cryst.* A**52**, 937-946.

Gilmore, C. J., Shankland, K. & Fryer, J. R. (1993). *Ultramicroscopy*, **49**, 132-146.

Glaeser, R. M., Jubb, J. S. & Henderson, R. (1985). *Biophys. J.* **48**, 775-780.

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

Hauptman, H. A. (1972). *Crystal Structure Determination. The Role of the Cosine Seminvariants*. New York: Plenum.

Hauptman, H. A. (1993). *Proc. R. Soc. London Ser. A*, **442**, 3-12.

Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. (1986). *Ultramicroscopy*, **191**, 147-178.

Hoppe, W., Gassmann, J. & Zechmeister, K. (1970). *Crystallographic Computing*, edited by F. R. Ahmed, pp. 26-36. Copenhagen: Munksgaard.

Karle, J. (1989). *Acta Cryst.* A**45**, 765-781.

Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635-651.

Lunin, V. Yu., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 896-903.

Lunin, V. Yu., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540-544.

Luzzati, V., Mariani, P. & Delacroix, H. (1988). *Makromol. Chem. Macromol. Symp.* **15**, 1–17.

Luzzati, V., Tardieu, A. & Taupin, D. (1972). *J. Mol. Biol.* **64**, 269–286.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

Podjarny, A. D., Schevitz, R. W. & Sigler, P. B. (1981). *Acta Cryst.* A**37**, 662–668.

Reeke, G. N. & Lipscomb, W. N. (1969). *Acta Cryst.* B**25**, 2614–2623.

Rogers, D. (1980). *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. Ladd & R. A. Palmer, pp. 23–92. New York: Plenum.

Sass, H. J., Büldt, G., Beckman, E., Zemlin, F., Van Heel, M., Zeitler, E., Rosenbusch, J. P., Dorset, D. L. & Massalski, A. (1989). *J. Mol. Biol.* **209**, 171–175.

Sayre, D. (1980). *Theory and Practice of Direct Methods in Crystallogrphy*, edited by M. F. C. Ladd & R. A. Palmer, pp. 271–286. New York: Plenum.

Stanley, E. (1986). *Acta Cryst.* A**42**, 297–299.

Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–113.

Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Zhang, K. Y. J. & Main, P. (1980). *Acta Cryst.* A**46**, 41–46.