Acta Cryst. (1998). A**54**, 290–295

Symbolic Addition in Protein Electron Crystallography – a Method for Finding Projected Helices

DOUGLAS L. DORSET

Electron Diffraction Department, Hauptman–Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203-1196, USA

(*Received 9 July 1997; accepted 14 November 1997*)

Abstract

The crystal structure of orthorhombic bacteriorhodopsin was determined in projection by direct methods from electron diffraction amplitudes, assuming that, after re-scaling the problem, the Fourier transform of projected α-helices could be modeled by atomic scattering factors. A basic set comprising two origin-defining phases, two phase values from Σ_1 triple estimates and an algebraic unknown (resolved early in the phase determination) was extended to a total set of 20 terms, with only two errors. Five helix sites were observed in the first potential map and, after three cycles of Fourier refinement, the rest of the asymmetric unit was found. The overall phase accuracy was 47° or 22° for the 25 most intense reflections.

1. Introduction

After the pioneering work of Podjarny *et al.* (1981), direct phasing methods have found a place in protein crystallography for the definition of molecular envelopes. Unlike more recent applications to atomic resolution data sets, the analysis utilizes diffraction information from the low-angle region so that the boundary between the protein domain and solvent can facilitate structural interpretation of electron-density maps. Often, the major drawback to this approach is not the applicability of standard direct methods for phase extension but that a complete low-resolution X-ray intensity set is rarely collected. Electron diffraction studies of thin protein microcrystals, on the other hand, routinely record all of the reflection intensities within the tilt sampling limits of the goniometer (Amos *et al.*, 1982). For this reason, there has been a parallel effort in applying direct methods to the phasing of such data, either starting from a lower-resolution basis set provided by the Fourier transform of an electron micrograph (*i.e.* phase extension) (Gilmore *et al.*, 1993; Dorset *et al.*, 1995), or actual *ab initio* techniques, *e.g.* involving the permutation of phases and screening likely solutions by a suitable figure of merit (Dorset, 1995; Gilmore *et al.*, 1996).

It must be emphasized here that the application of direct methods to electron diffraction data in no way implies a criticism of electron micrographs as a source of crystallographic phases (*via* image processing). The correctness of such image-derived phases is underscored by a recent comparison of X-ray and electron crystallographic determinations of the bacteriorhodopsin structure at high resolution (Pebay-Peyroula *et al.*, 1997). Again, in the current context, projected electron diffraction patterns are exploited only for the completeness of the low-resolution intensity data. However, the low-resolution phasing problem has a much wider significance to protein crystallography, particularly for the elucidation of molecular envelopes (Podjarny *et al.*, 1981).

Based on an idea proposed by Harker (1953), one approach to *ab initio* phasing of protein diffraction data has been to exploit the Fourier transform of a 'glob' density element. The scale of this phasing procedure can vary greatly, from simulation of the overall protein mass distribution (Andersson & Hovmöller, 1996) to generalized scattering factors for amino acid segments (Guo *et al.*, 1995). In electron crystallography, it has been observed that the projection of a helical column might also be treated as a pseudoatomic glob (Dorset, 1997*a*). In the initial evaluation of this hypothesis, a centrosymmetric projection for halorhodopsin was solved by symbolic addition, as if the cluster of helices could be treated as a small-molecule problem (Dorset, 1997*a*). The determination can be somewhat more complicated when the projection is noncentrosymmetric. For example, in a case where the trigonal unit-cell symmetry causes many *hk*0 reflections to be phase invariants, few origin-defining reflections can be specified initially, requiring that multiple solutions be generated (Dorset, 1997*b*). Although there may be some difficulty in specifying the correct solution, the phase determination itself is still quite accurate, as long as the protein itself is largely composed of α-helices.

The generality of this procedure, nevertheless, still needs to be established. In this paper, another representative centrosymmetric projection, *i.e.* from the orthorhombic form of bacteriorhodopsin, is analyzed.

2. Materials and methods

2.1. *Preparation and electron diffraction*

The orthorhombic form of bacteriorhodopsin was prepared by Michel *et al.* (1980) by combined action of two detergents at low pH. Electron diffraction patterns

from the glucose-embedded two-dimensional crystalline sheets, observed to 3.5 Å resolution [procedures given by Unwin & Henderson (1975)], revealed that the projected plane-group symmetry was *pgg* with cell constants $a = 57.6$, $b = 73.5$ Å. Crystallographic phases were also obtained to a resolution of 6.5 Å from the Fourier transform of averaged electron micrographs (Michel *et al.*, 1980) and were used to monitor the direct phasing procedure to be described below. Combination of electron diffraction amplitudes and phases at this resolution produces a potential map where the familiar α-helical cluster of the protein, seen earlier (Henderson & Unwin, 1975) for the trigonal form (plane group $p3$), is readily visualized (Fig. 1).

2.2. *Data preparation and direct phase determination*

The premise behind the structure analysis is that the cross section of an α-helix can, after appropriate rescaling, be regarded as a pseudoatom. Such helices can touch one another with a typical center-to-center distance of about 15 Å (Parsons & Martius, 1964), about ten times the length of a carbon–carbon single bond. [A smaller center-to-center distance of about 12.0 Å would be expressed by the α-polymorph of poly-γ-methyl-L-glutamate (Tatarinova & Vainshtein, 1962).] If the density cross section is Gaussian then, after the dimensions of the determination are reduced tenfold, the glob scattering factor can be well approximated by, say, the electron form factor for carbon (Doyle & Turner, 1968). (Here the difference between the actual Lorentz shape of this form factor and the Gaussian transform will not be considered to be very important.) Thus, the actual cell dimensions for the determination are regarded to be $a = 5.76$, $b = 7.35$ Å.

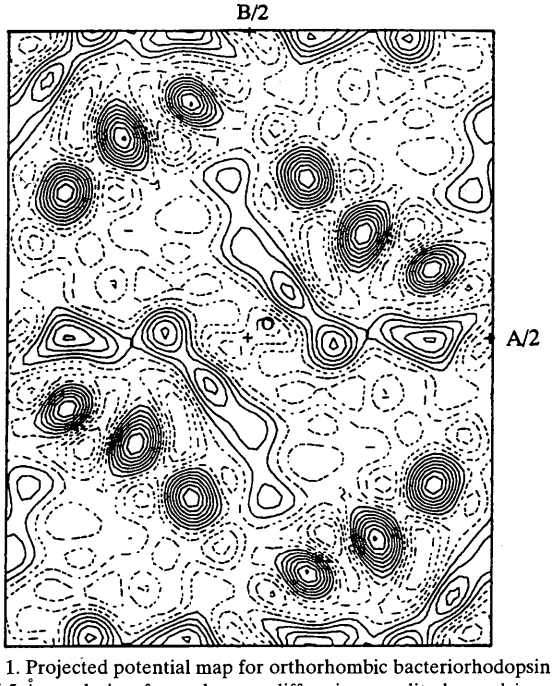

Fig. 1. Projected potential map for orthorhombic bacteriorhodopsin at 6.5 Å resolution from electron diffraction amplitudes and image-derived crystallographic phases. The map origin is marked O and crosses + are marked at $\frac{1}{2}$ of the unit-cell edges (here written as $A/2$, $B/2$).

Based on the carbon scattering factor, a Wilson (1942) plot was made for the intensity data I_h^{obs} reported by Michel *et al.* (1980), again after the dimensional adjustment. From this, normalized structure factors were calculated from $|E_h|^2 = I_h^{obs}/\varepsilon \sum (f_c')^2$, where ε is a statistical weight (in this example, accounting for reciprocal axial reflection classes including systematic absences) and f_c' is the scattering factor corrected for an overall Debye–Waller factor. From the weights $A = (2/N^{1/2})|E_{h_1} E_{h_2} E_{h_3}|$, three-phase Σ_2 invariants (Hauptman, 1972) were generated and sequenced in order of decreasing probability that the value of $\psi = \phi_{h_1} + \phi_{h_2} + \phi_{h_3}$ could be predicted reliably. In addition, a smaller number of highly probable Σ_1 three-phase invariants, where $h_1 = h_2 = -\frac{1}{2}h_3$, were consulted.

To give the least-biased determination of phases, a logical sequence of new phases determined from a defined basis set was established by the well known convergence procedure (Germain *et al.*, 1970). That is, given the origin-defining phases and others (here assumed to be algebraic unknowns), the sequence of ϕ_h is ranked with all possible contributors in the Σ_2 invariants $\phi_h = \langle \phi_k + \phi_{h-k} \rangle$, these individual contributors weighted according to the value of a parameter α_{est}, defined elsewhere (DeTitta *et al.*, 1975). In this determination, the symbolic addition procedure (Karle & Karle, 1966) was used to evaluate each contributing phase contribution in the triples leading to the sequential phase. It may be found that some of the individual phase determinations are inconclusive (*i.e.* self-contradictory) so that they would not be added to the growing set of new values.

2.3. *Refinement*

After identification of 'atom' (*i.e.* helix) positions in the initial potential map, a Fourier refinement was begun. In the usual structure-factor expression

$$F_h = \sum_i f_i \cos(2\pi \mathbf{h} \cdot \mathbf{r}),$$

the scattering factor for carbon is retained while the dimension of the problem remains reduced tenfold.

3. Results

3.1. *Adequacy of the scattering-factor model*

As in the original direct determination of bacteriorhodopsin with data in the $p3$ cell (Dorset, 1997*b*), the density distribution of the molecule could be modeled by either a seven-atom or an eight-atom cluster, after

re-scaling the dimensions of the problem. That is to say, the positions of at least six helices were clearly defined by the density distribution in the potential map. An additional density site could be assumed while yet another in a 'tail' region could also be added. For the 76 unique reflections in the data set, either model gave the same phase accuracy when compared to those found from the electron micrograph transform, *i.e.* 12 errors in all, or an overall mean deviation of 28°. (This result can be compared to the two trigonal forms of the protein investigated earlier, *e.g.* mean phase errors of 37° for the native protein and 29° for the deoxycholate-treated form when pseudoatom globs were simulated by an atomic scattering factor after re-scaling of dimensions.) Neither model accounted for the intensity distribution of the diffraction pattern very well. When no Debye–Waller factor was given to the carbon scattering factor (*i.e.* $B = 0.0$ Å^2), $R = 0.61$ for the seven-atom model and 0.54 for the eight-atom model. This intensity simulation was therefore somewhat less favorable than found for the two trigonal forms of the same protein (Dorset, 1997*b*) (*i.e.* 0.48 for the native protein and 0.33 for the delipidized form).

3.2. *Phase determination*

A Wilson (1942) plot, assuming the carbon scattering factor could serve as a model for the helix sites in the protein, gave a negative overall temperature factor, $B = -4.2$ Å^2. This was used for the subsequent calculation of $|E_h|$ and, hence, A, for ordering of the Σ_2 triples. In addition, $N = 8$ was assumed (but $N = 7$ or a near estimate would only change the absolute magnitude of A but not the sequence of the triple invariants).

From the list of reflections, ordered on $|E_h|$, the 630 reflection (third in the list and the first with allowable index parity for origin definition) was tested *via* the convergence procedure (Germain *et al.*, 1970) to ascertain what other reflections would be needed to define the greatest number of new phase terms. It was found that 650, 340 and 250 were also required. Since the first of these has the same parity as the 630 reflection, the 340 reflection was chosen as the second origin-defining reflection and algebraic values were assigned to both 650 and 250. However, it was also found from highly probable Σ_1 triples that $\phi_{040} = \phi_{080} = 0$. With the assumption $\phi_{630} = \phi_{340} = \pi$ rad [but this combination was chosen only to preserve the origin found from image averaging in the original study of Michel *et al.* (1980)], the value of $\phi_{650} = 0$ was predicted from the first Σ_2 triple, knowing the value of ϕ_{080} from the Σ_1 estimate. The value of $\phi_{250} = \pi$ was soon established. Working through the sequence of invariants, again according to the convergence procedure, it was interesting to note that the phase of the 420 reflection was incorrectly predicted to be 0 in the second triple of the sequence. However, using this false value allowed many other reflections to be determined correctly. Further, contradictory values were indicated for the phase of the 510 reflection so this was rejected from the final list of 20 phase terms (Table 1). There were only two errors in this initial set.

Table 1. *Initial phase determination for orthorhombic bacteriorhodopsin*

hkl	ϕ	ϕ_{im}	*hkl*	ϕ	ϕ_{im}
040	0	0	360	0	0
080	0	0	420	0	π
120	π	π	440	π	π
190	π	π	460	0	0
210	π	π	480	π	π
230	0	0	530	π	π
250	π	π	540	0	0
260	π	π	630	π	π
290	π	0	650	0	0
340	π	π	720	0	0

Table 2. *Comparison of density sites for orthorhombic bacteriorhodopsin*

Direct methods		Earlier study		
x	*y*	*x*	*y*	Helix site†
0.122	0.256	0.122	0.256	7(*B*)
0.234	0.173	0.234	0.173	6(*C*)
0.378	0.112	0.378	0.112	5(*D*)
0.319	0.000	0.372	−0.005	4(*E*)
0.149	0.015	0.181	−0.010	3(*F*)
0.062	0.096	0.089	0.073	2(*G*)
−0.011	0.183	0.004	0.143	1(*A*)
−0.021	0.269	−0.037	0.256	1(*A*)

† See Engelman *et al.* (1980).

Phase determination by symbolic addition was carried out in another less-structured way. Here, the top 50 Σ_2 triples were listed in decreasing order of A. After using the same Σ_1 estimates and origin definers given above, 23 reflections in all were assigned phase values. Again, the conflicting estimates were indicated for the 510 reflection and the correct phase choice for the 440 reflection was made statistically (the greatest number of contributors indicating the value should be π). There were four errors in the list but the starting map was nearly equivalent to the one calculated from the set obtained from the convergence list.

3.3. *Refinement*

From the phase values in Table 1, the potential map in Fig. 2(*a*) could be produced. From the five most intense peaks, trial pseudoatomic positions were chosen for a structure-factor calculation, this giving phase estimates for the complete list of 76 reflections, containing 18 errors. With the associated amplitudes, this gave the map in Fig. 2(*b*) from which two more peak positions were chosen. After a second structure-factor calculation, the map in Fig. 2(*c*) was obtained, suggesting the final peak position. The final structure-factor calculation produced a map very similar in

appearance to Fig. 1. The final phase error was actually worse in terms of numbers (20 out of 76 reflections) corresponding to a mean deviation of 47° from those values obtained from the image transform, but only 22° for the top 25 reflections. Another cycle decreased the total number of false phases to 18 again but the deviation for the top 25 reflections was then 29°. For the eight-atom model used to sample Fig. 1, the recovered peaks from this direct analysis differed on average by a mean value of 1.5 Å. There was essentially no differ-

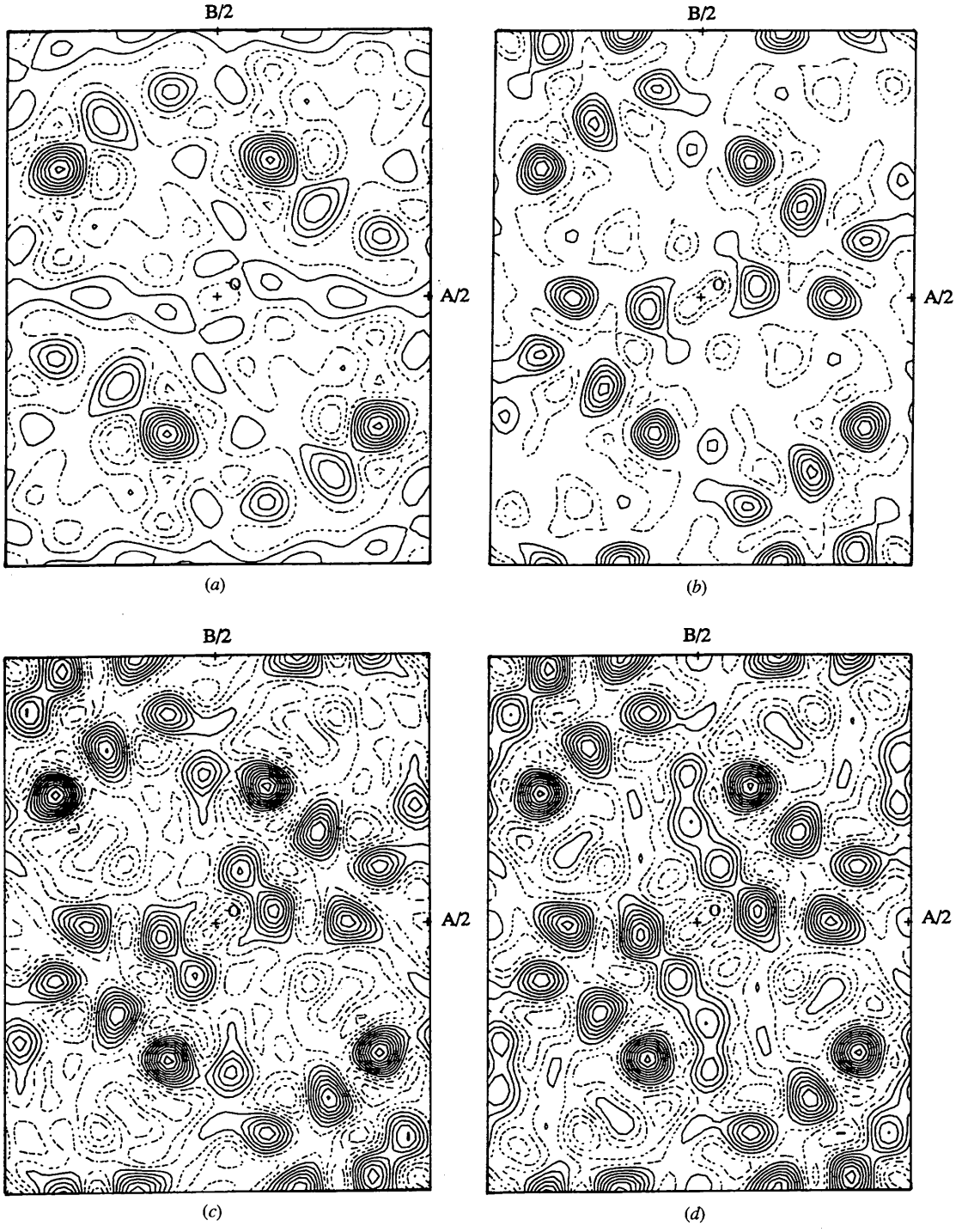

Fig. 2. Direct phase determination of orthorhombic bacteriorhodopsin – potential maps. (*a*) Initial phases from symbolic addition (Table 1); (*b*) Fourier cycle 1; (*c*) Fourier cycle 2; (*d*) Fourier cycle 3.

ence in position for the three most intense peaks, meaning that the others differed on average by 2.5 Å. A comparison of positions is given in Table 2.

4. Discussion

Although the simulation of an average glob scattering factor by an atom scattering factor, after rescaling of dimensions, does not account very well for the intensity transform of the protein crystal structure, the assumption gives a close enough match to the phase set that the structure can be determined by direct methods in a straightforward fashion. The pseudoatomicity of the density distribution, therefore, seems to be the major criterion for the success of such phase determinations although its exact simulation by the model scattering factor need not be exact. A major problem in such determinations is the deviation from a circular cross section of some helices – an anisotropy corresponding to their tilt. The most accurate part of the structural analysis is the location of the inner cluster of three helices that are the least tilted to the projection direction (Henderson & Unwin, 1975; Henderson *et al.*, 1990). In the outer cluster, however, the density centers for the helical columns are less clearly defined. In other words, there were eight pseudoatom positions in this analysis used to simulate seven sites. Two of these pseudoatoms sampled the density area 1 defined in the paper of Engelman *et al.* (1980), a region found later to correspond to the most greatly tilted column axis (Henderson *et al.*, 1990). As shown above, a seven-atom model did not improve the phase agreement. Thus, the outcome of this *ab initio* determination was most accurate for six sites. Nevertheless, a close match was found to the density distribution of the initial determination based on image-derived phases where the agreement with the projected trigonal density profile is also confused in the region sampled by two atoms in this determination (Michel *et al.*, 1980).

Why is this *ab initio* phase determination so accurate? In an earlier work at similar resolution on the delipidized $p3$ structure of the same protein (Dorset, 1997*b*), it was found that the 19 most probable Σ_2 triple invariants predicted a mean average value of $\psi = 51°$. In this determination, the average value of the invariant sum was 61.2° for the top 50 triples, again where a value of 0° was expected. After Fourier refinement of a best model, the best overall phase error was 53.8° for 35 reflections or 22.9° for the 14 most intense reflections. Helix sites were located within 1.6 Å of their actual positions. For the native protein in its trigonal form, the mean phase error was 65.2° for all 50 data but 38.4° for the 18 most intense reflections. Helix sites were found within 1.9 Å of their actual positions (Dorset, 1997*b*).

These findings support the statement made by Fan *et al.* (1991) that the phase invariants for low-resolution protein diffraction data might not be less valid than similar values for small molecules, even though there are fewer of them. Because most of the scattering intensity from the macromolecule is also concentrated in the low-resolution range, these invariants should also have the highest probability. Obviously, if some other feature of the density distribution can be assumed *a priori*, then the determination can be adequently constrained to a successful conclusion. If it cannot, *e.g.* in the case of a structure consisting of mostly β-sheets (Dorset, 1997*b*), a case difficult to model with pseudoatom scattering factors, the outcome will be much less favorable.

Finally, it should be pointed out that the four favorable *ab initio* determinations described so far have essentially the same density motif arranged in different packing arrangements. Does this mean that a protein with high helix content but also with a different clustering of these projected density features might be more difficult to analyze by direct methods and the pseudoatomic glob model? Efforts will be made to answer this question in future investigations of other protein structures. Also, attempts will be made to extend these techniques to three-dimensional data.

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

References

Amos, L. A., Henderson, R. & Unwin, P. N. T. (1982). *Prog. Biophys. Mol. Biol.* **39**, 183–231.

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

DeTitta, G. T., Edmonds, J. W., Langs, D. A. & Hauptman, H. (1975). *Acta Cryst.* A**31**, 472–479.

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

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

Dorset, D. L. (1997*b*). *Acta Cryst.* A**53**, 445–455.

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.

Engelman, D. M., Henderson, R., McLachlan, A. D. & Wallace, B. A. (1980). *Proc. Natl Acad. Sci. USA*, **77**, 2023–2027.

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

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.

Guo, D. Y., Smith, G. D., Griffin, J. F. & Langs, D. A. (1995). *Acta Cryst.* A**51**, 945–947.

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

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

Henderson, R., Baldwin, J. M., Ceska, T., Zemlin, F., Beckmann, E. & Downing, K. H. (1990). *J. Mol. Biol.* **213**, 899–929.
Henderson, R. & Unwin, P. N. T. (1975). *Nature (London)*, **257**, 28–32.
Karle, J. & Karle, I. L. (1966). *Acta Cryst.* A**24**, 390–397.
Michel, H., Oesterhelt, D. & Henderson, R. (1980). *Proc. Natl Acad. Sci. USA*, **77**, 338–342.
Parsons, D. F. & Martius, U. (1964). *J. Mol. Biol.* **10**, 530–533.
Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. (1997). *Science*, **277**, 1676–1681.
Podjarny, A. D., Schevitz, R. W. & Sigler, P. B. (1981). *Acta Cryst.* A**37**, 662–668.
Tatarinova, L. I. & Vainshtein, B. K. (1962). *Vysokomolek. Soed.* **4**, 261–269.
Unwin, P. N. T. & Henderson, R. (1975). *J. Mol. Biol.* **94**, 425–440.
Wilson, A. J. C. (1942). *Nature (London)*, **150**, 151–152.