Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 1791–1794, March 1997
Biophysics

Direct phase determination in protein electron crystallography: The pseudo-atom approximation

(electron diffraction/crystal structure analysis/direct methods/membrane proteins)

DOUGLAS L. DORSET

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

Communicated by Herbert A. Hauptman, Hauptman–Woodward Medical Research Institute, Buffalo, NY, December 12, 1996 (received for review October 28, 1996)

ABSTRACT The crystal structure of halorhodopsin is determined directly in its centrosymmetric projection using 6.0-Å-resolution electron diffraction intensities, without including any previous phase information from the Fourier transform of electron micrographs. The potential distribution in the projection is assumed *a priori* to be an assembly of globular densities. By an appropriate dimensional re-scaling, these "globs" are then assumed to be pseudo-atoms for normalization of the observed structure factors. After this treatment, the structure is determined directly by conventional direct methods, followed by Fourier refinement, leading to a mean phase deviation of only 20° (from the values originally found from the image transform) for the 45 most intense reflections.

Recently, there has been increasing interest in using direct phasing methods to aid the electron diffraction determination of macromolecular structures at low resolution. Much of this attention has been focused on the problem of phase extension, i.e., starting with a partial phase set and extending it into the whole range of the recorded data set to obtain a clearer view of the molecular envelope. In original x-ray crystallographic studies, such efforts have started with partial information obtained from multiple isomorphous replacement (1, 2). In electron crystallography, the partial phase set would typically be determined from the Fourier transform of experimental electron micrographs after image averaging (3). In either case, successful application of traditional direct methods, such as the Sayre equation (4–6), or newer methods, including maximum entropy and likelihood (7), have shown that the low-resolution data from many macromolecules are accessible to such analyses. Only when the node of averaged intensity occurring near $(5 \text{ Å})^{-1}$ is reached, are significant problems encountered in the phase extension (4).

The prospect of actual *ab initio* phase determinations, assuming that no preliminary information is available, also has been explored. The Sayre equation, followed by phase annealing steps, was quite successful in one case (8), as was the use of maximum entropy and likelihood (9). However, if such multisolution methods are to be effective, there must also be a robust figure of merit that allows identification of the correct structure solution. This requirement may not be easily satisfied, despite the approximate validity of smoothness and flatness criteria (10) for the density distribution when structures are determined at low resolution (6). [The log-likelihood gain criterion in maximum entropy procedures may be a suitable way to solve this problem (9).]

Another approach to such phasing problems, especially in cases where the structures have appropriate distributions of mass, would be to adopt a pseudo-atom approach. The concept of using globular sub-units as quasi-atoms was discussed by David Harker in 1953, when he showed that an appropriate globular scattering factor could be used to normalize the low-resolution diffraction intensities with higher accuracy than the actual atomic scattering factors employed for small molecule structures (11). In a sense, this idea has already been employed (in real space) for phase determination in protein x-ray crystallography, when clusters of globular subunits, randomly arrayed in numerous patterns, have been generated to seek an adequate approximation to the density distribution in the unit cell (12).

There is, possibly, a more straightforward use of this idea in reciprocal space, and that is to test the suitability of intensity data normalized by a "glob" transform for analysis of crystallographic phases by conventional direct methods, as if the structure were that of a small molecule. The successful analysis of a projected membrane protein structure is described in this paper.

ANALYSIS

Data Set. Electron diffraction intensity data, which had been collected at 120 kV to 6-Å resolution from frozen-hydrated two-dimensional (*Halobacterium halobium*) halorhodopsin crystals by Havelka *et al.* (13), were used in this analysis. There were 101 unique reflections in the published list. The centrosymmetric plane group for the square projection (a = 102.0 Å) is p4 gm (# 12). The crystallographic phases, derived from the Fourier transform of averaged electron micrographs, had been published in the original report of this structure (13).

Normalization of Structure Factors. Harker (11) had treated the globular subunits of a protein as hard spheres in his original treatment and, if these were to have a diameter x, then their Fourier transforms would have cross-sectional terms containing the function $\text{sinc}(\pi x) = \sin(\pi x)/\pi x$. Since it is only the first envelope of the two-dimensional "Airy disc" transform that would be used to approximate a globular scattering factor, it might be preferable to replace this function by a Gaussian term (or something close to a Gaussian function), since the Fourier transform of this function (i.e., another Gaussian function) does not produce the "edge diffraction" effect found in the sinc function. It is also well known that several functions (e.g., Gauss and sinc) can be used as good approximations of one another in the most intense part of their envelopes (14).

Indeed, the scattering factors of atoms are approximately Gaussian, since they can be expressed as a weighted sum of Gaussian functions (15). [The actual Lorentzian shape of this sum (15) is assumed not to be a significant deviation.] In this study, therefore, it was assumed that all details of the structure could be rescaled dimensionally by a factor of 10. The rationale for this 10-fold rescaling (and the choice of an approximate

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

0027-8424/97/941791-4$2.00/0
PNAS is available online at **http://www.pnas.org**.

scattering factor) is easily found by comparing the center-to-center distance for two touching α-helices (16), i.e., $\approx$15 Å, to the distance, 1.54 Å, of a carbon–carbon single bond (17). Thus, a square unit cell with a = 102.0 Å would become one with a′ = 10.20 Å edges, and the data resolution would then be assumed to be recorded to 0.6-Å resolution instead of just 6 Å. Therefore, for normalization after this rescaling of dimensions, the glob transform was approximated by the electron scattering factor of carbon (15). [This approach is to be distinguished from a "globular" approximation on another scale, i.e., where an atomic scattering factor (7) or a more accurate phenomenological distribution (18) simulates the Fourier transform of an amino acid residue, for example.]

As usual, the intensities were evaluated with a Wilson plot (19), based on $\ln \langle I_{\mathrm{h}}^{\mathrm{obs}}/\Sigma f^2\rangle = \ln \mathrm{C} - 2\mathrm{B}_{\mathrm{iso}}(\sin^2\theta/\lambda^2)$, to determine the overall temperature factor $\mathrm{B}_{\mathrm{iso}} = 2.8$ Å (2), as shown in Fig. 1*a*. After adjusting the carbon scattering factor, $f' = f \exp(-\mathrm{B} \sin^2\theta/\lambda^2)$, i.e., the approximation for the glob transform after rescaling, for "thermal motion," normalized structure-factor magnitudes $|E_{\mathrm{h}}|$ were calculated from the observed intensities in the usual way (20): $|E_{\mathrm{h}}| = I_{\mathrm{h}}^{\mathrm{obs}}/\varepsilon\Sigma(f')^2$. Here the statistical weight ε compensates for certain reflection index classes due to the symmetry properties of the plane group. After these normalized quantities were calculated, 95 were found to have suitably high values to be used for direct phase determination.

Direct Phase Determination. Three-phase structure invariants (20), defined $\phi_{\mathrm{h}} = \phi_{\mathrm{k}} + \phi_{\mathrm{h-k}}$, were employed for the analysis. (Here ϕ_{h} is the phase of a reflection with Miller indices $\mathbf{h} = \mathrm{h}_1\mathrm{k}_1\ell_1$ and $\mathbf{k} = \mathrm{h}_2\mathrm{k}_2\ell_2$. When all indices are different, the equality defines a so-called Σ_2 invariant, but when $\mathbf{h} = 2\mathbf{k}$, it defines a Σ_1 invariant. The former is the most useful.) These generate simultaneous equations in the crystallographic phases that can be ranked according to their decreasing reliability of being correctly predicted according to, e.g., $A = (\sqrt{N}/2)|E_{\mathrm{h}}E_{\mathrm{k}}E_{\mathrm{h-k}}|$ for the Σ_2 invariants. Here N is the number of atoms in the unit cell (assumed to be equally weighted). (Note that the correct number of globs, i.e., seven in the asymmetric unit, was used in this determination. This exact number is not absolutely required since, to a first approximation, it will only affect the relative scaling of the A terms but not their ordering.) Using a convergence procedure (21), all possible phase contributors (from a total set of 507 triples generated to a minimum value of $A = 0.5$) to a given reflection **h** were considered in the sequence most optimal for finding the phases of the most number of reflections after definition of a small basis set.

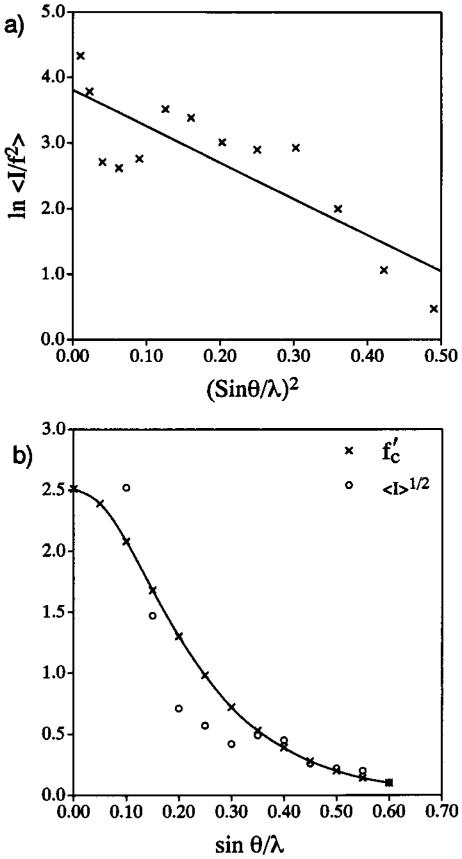

FIG. 1. (*a*) Wilson plot made after rescaling of halorhodopsin electron diffraction intensities. (*b*) Fit of rescaled plot of average-observed structure-factor magnitudes vs. resolution to the carbon scattering factor curve (with temperature factor included).

Crystallographic phases were determined by a procedure termed "symbolic addition" by some researchers (22). Only one reflection could be used for origin definition in this projection, since both gg0 and uu0 combinations are phase seminvariants for this plane group (23). (Here, g = "gerade" = even and u = "ungerade" = odd index values.) From the sequential rank of $|E_{\mathrm{h}}|$ terms, a ug0 reflection, the phase $\phi_{560} = \pi$, was assigned. [The other possible value, 0, also could have been used legitimately, but this term was chosen to preserve the origin used in earlier determinations (5, 9, 13), just to facilitate phase comparison.] From the most probable Σ_1 relationships, the phases: $\phi_{12,12,0} = \phi_{880} = \phi_{0,16,0} = \phi_{0,10,0} = \phi_{2,10,0} = 0$, were also accepted. Finally, two other reflections were initially given algebraic phase values, $\phi_{080} = \mathrm{a}$, $\phi_{590} = \mathrm{b}$. Their actual values (respectively, 0 and π) were determined unequivocally in the course of the phase solution.

Fourier Refinement. As will be shown below, the initial phase set yielded a potential map [via $\rho(r) = \frac{1}{V}\int|F_{\mathrm{h}}^{\mathrm{obs}}| \exp(i\phi_{\mathrm{h}}) \exp(-2\pi i\mathbf{h}\cdot\mathbf{r})dr$] with a partial structure solution. Glob positions were picked from the map as atoms would have been for a small molecule structure and structure factors were then calculated with these identified positions to find new phase terms, again using the scaling approximation above and the carbon scattering factor. In successive cycles, new atom positions are identified and their positions used for new structure-factor calculations, etc.

In this refinement procedure, it was assumed that the normal indicators for small molecule structure determinations were valid. That is to say, density profiles of maps could suggest new atom sites and these were accepted if the new map, based on the revised phase set, retained the peak as reinforced density. However, the crystallographic R-factor was not used as a figure of merit (see below), nor were difference maps calculated during this refinement.

RESULTS

After definition of the basis set, 32 phase values were determined, the derived set containing 7 errors (mostly associated with weaker reflections). Combining these values with $|F_{\mathrm{h}}^{\mathrm{obs}}|$ to calculate the first potential map, four globular sites were observed strongly (Fig. 2*a*), with the weak indication of a fifth. After a structure-factor calculation based on the five sites to find phases for all unique reflections, a second map (Fig. 2*b*) was calculated, reinforcing the position of the fifth site and indicating a sixth position. The next structure-factor calculation produced a revised phase list. The resultant potential map reinforced the sixth position. It should also be noted that one glob was elongated (Fig. 2*c*), and so an estimated position was given near the end of the elongation. After the next structure-factor calculation, the revised phase list resulted in a potential map (Fig. 2*d*), which is very close in appearance to the one found originally (13) from the Fourier transform of electron micrographs (Fig. 2*e*).

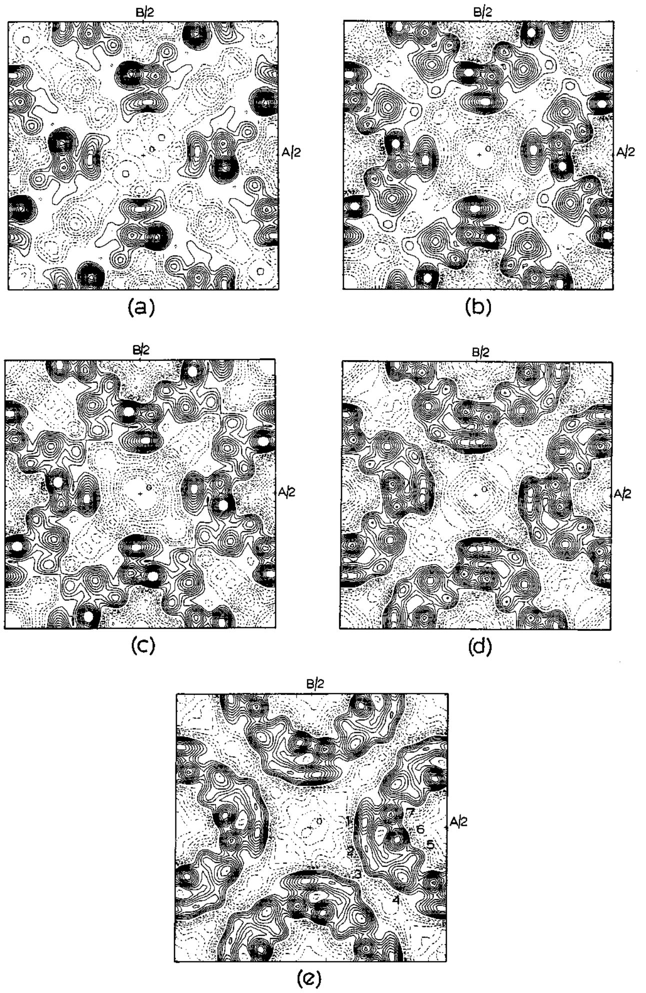

FIG. 2. Progress of structure determination. (*a*) Initial potential map after symbolic addition. (*b*) Map after first structure-factor calculation. (*c*) Map after second structure-factor calculation. (*d*) Map after third structure-factor calculation. (*e*) Structure determined by electron microscopy (13). Helix positions are numbered (to match to those given in Table 2).

The positions of identified α-helix centers (the globs) found in this determination were very close to the ones located in the original determination (Table 1). If all reflections were considered, the mean phase difference to those found from the image transform was $\langle|\Delta\phi|\rangle = 56.8°$. For the 45 most intense reflections this difference was only 20.0°.

DISCUSSION

The results of this straightforward phasing procedure, based on the symbolic addition procedure often employed in small molecule crystallography, are found to be much more accurate for this example than the results obtained earlier by another direct phasing technique, involving an annealing step after expansion in shells of reciprocal space (9). No figure of merit was needed for identification of the structure, and the usual rules for accepting new peaks during refinement were also accepted, again without the reinforcement of the crystallographic *R*-factor.

Table 1. Mean phase errors for halorhodopsin compared with previous determinations

	This determination	Previous direct phasing	Phase extension
To 6 Å, all data	56.8°	74.8°	55.2°
To 6 Å, $|F_h| \geq 1.0$	20.0°	44.0°	24.0°

From this determination, it is apparent that the globular scattering factor suggested by Harker (11) is quite appropriate for determination of optimal normalized structure factors. With this normalization, and assuming that pseudo-atom positions will be sought later in the potential maps, it is clear also that standard direct phasing procedures will be effective for determining what is essentially a small molecule structure problem. Obviously, if the globular pseudo-atoms account well for the unit cell density, then their transforms should faithfully simulate the unit cell Fourier transform (diffraction pattern).

However, it is obvious too that the rescaling device used for this determination is somewhat of an artifice, even if the tabulated atomic scattering factor (15) was a convenient approximation for the globular transform. Just how well did the carbon atom model serve in this case? If a cluster of seven carbon atoms were used to simulate the globs (but with somewhat higher thermal values than found in the Wilson plot above), after rescaling the dimensions of the problem, the crystallographic *R*-factor was calculated to be 0.41, not a highly accurate portrayal of the intensity transform. However, for the complete phase list, there were only 28 errors (for 101 reflections) and these were mostly associated with weaker reflections. The determined phases were, of course, combined with observed structure factors to calculate the potential maps in Fig. 2, so the density distribution would be expected to be quite good. This is shown by the comparison of helix sites in Table 2.

It is clear, therefore, that further work could be devoted to the formulation of more accurate scattering factors for protein subunit globs (including anisotropic distributions of density to simulate tilted helices, for example). The fit of the carbon atom scattering factor curve, adjusted for thermal motion, to the fall-off of $\langle I_{\mathrm{h}}^{\mathrm{obs}}\rangle^{1/2}$ with sin θ/λ is found to be only a fair approximation (Fig. 1*b*), e.g., the curves fall to a near zero value at about the same resolution, a criterion that should be used as an *a priori* test of any model for the glob transform. (It is probably not worthwhile to overcomplicate the simulation by allowing globs of various sizes, since thermal motion can be used as another variable when the map indicates that it might be needed for improving the fit to density.) Thus, the fit to the scattering envelope could be much better, although the present approximation does not have too serious an effect on the results observed in this study.

One serious limitation of this rescaling model is that it restricts the determination to the spatial frequency limit of the phenomenological scattering factor, or else there would be problems with hyper-resolved data sets. In other words, the technique is only suited to a modest diffraction resolution (e.g., the 6 Å limit of this study), because the success of the structure determination probably also relies on the favorable analytical properties of the intensity transform for direct phase determinations (4–8) up to the nodal zero of average scattered intensity observed near 5 Å.

Table 2. Helix positions for halorhodopsin found by direct methods.

	This determination		Image analysis	
Position	x/a	y/b	x/a	y/b
1	0.199	0.018	0.201	0.016
2	0.199	−0.069	0.211	−0.078
3	0.237	−0.164	0.239	−0.144
4	0.330	−0.180	0.339	−0.188
5	0.404	−0.114	0.378	−0.115
6	0.306	−0.045	0.319	−0.045
7	0.294	0.042	0.294	0.045

As will be shown in a future communication, there are other protein structures with pseudo-atomic subunits, such as the lipid-containing (3) or delipidized (24) forms of bacteriorhodopsin, that can also be determined in projection by this technique, even when the projected density distribution is noncentrosymmetric. On the other hand, other nonglobular secondary structure, e.g., β-sheets, may not be so conveniently visualized by this approach.

I acknowledge valuable early discussions of this problem with the late Dr. David Harker, nearly 15 years ago. This research was supported by Grant GM-46733 from the National Institute of General Medical Sciences, which is gratefully acknowledged.

1. Podjarny, A. D., Yonath, A. & Traub, W. (1976) *Acta Crystallogr. A* **32,** 281–292.
2. Podjarny, A. D., Schevitz, R. D. & Sigler, P. (1981) *Acta Crystallogr. A* **37,** 662–668.
3. Henderson, R., Baldwin, J., Downing, K. H., Lepault, J. & Zemlin, F. (1986) *Ultramicroscopy* **19,** 147–178.
4. Dorset, D. L., Kopp, S., Fryer, J. R. & Tivol, W. F. (1995) *Ultramicroscopy* **57,** 59–89.
5. Dorset, D. L. (1995) *Micron* **26,** 511–520.
6. Dorset, D. L. (1996) *Acta Crystallogr. A* **52,** 480–489.
7. Gilmore, C. J., Shankland, K. & Fryer, J. R. (1993) *Ultramicroscopy* **49,** 132–146.
8. Dorset, D. L. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10074–10078.
9. Gilmore, C. J., Nicholson, W. V. & Dorset, D. L. (1996) *Acta Crystallogr. A* **52,** 937–946.
10. Luzzati, V., Mariani, P. & Delacroix, H. (1986) *Makromol. Chem. Macromol. Symp.* **15,** 1–17.
11. Harker, D. (1953) *Acta Crystallogr.* **6,** 731–736.
12. Lunin, V. Yu., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumetsev, A. G. & Podjarny, A. D. (1995) *Acta Crystallogr. D* **51,** 896–903.
13. Havelka, W. A., Henderson, R., Heymann, J. A. W. & Oesterhelt, D. (1993) *J. Mol. Biol.* **234,** 837–846.
14. Gaskill, J. D. (1978) *Linear Systems, Fourier Transforms and Optics* (Wiley, New York).
15. Doyle, P. A. & Turner, P. S. (1968) *Acta Crystallogr. A* **24,** 390–397.
16. Parsons, D. F. & Martius, U. (1964) *J. Mol. Biol.* **10,** 530–533.
17. Ladd, M. F. C. & Palmer, R. A. (1993) *Structure Determination by X-Ray Crystallography* (Plenum, New York), 3rd Ed., p. 434.
18. Guo, D. Y., Smith, G. D., Griffin, J. F. & Langs, D. A. (1995) *Acta Crystallogr. A* **51,** 945–947.
19. Wilson, A. J. C. (1949) *Acta Crystallogr.* **2,** 318–321.
20. Hauptman, H. A. (1972) *Crystal Structure Determination: The Role of the Cosine Seminvariants* (Plenum, New York).
21. Germain, G., Main, P. & Woolfson, M. M. (1970) *Acta Crystallogr. B* **26,** 274–285.
22. Karle, J. & Karle, I. L. (1966) *Acta Crystallogr.* **21,** 849–859.
23. Hahn, T., ed. (1992) *International Tables for Crystallography: Vol. A., Space Group Symmetry* (Kluwer, Dordrecht, The Netherlands), 3rd Ed.
24. Glaeser, R. M., Jubb, J. & Henderson, R. (1985) *Biophys. J.* **48,** 775–780.