Acta Cryst. (1996). A52, 937–946

Direct Methods in Protein Electron Crystallography: the *Ab Initio* Structure Determination of Two Membrane Protein Structures in Projection using Maximum Entropy and Likelihood

Christopher J. Gilmore,[a]* William V. Nicholson[a] and Douglas L. Dorset[b]

[a]*Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, and* [b]*Electron Diffraction Department, Hauptman–Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, New York 14203-1196, USA. E-mail: chris@chem.gla.ac.uk*

(Received 22 March 1996; accepted 1 July 1996)

Abstract

Using maximum entropy and likelihood, an *ab initio* phase determination was carried out in projection at *ca* 6–10 Å resolution for two dissimilar membrane proteins: the Omp F porin from the outer membrane of *E. coli* (largely β-sheet) and halorhodopsin (largely α-helix). Accurate phase information found for the most likely solutions enabled potential maps to be calculated that contained most of the essential structural details of these macromolecules without the need for any image-derived phases as a starting set for phase extension or the necessity to use envelopes or electron-density histograms. A comparison with earlier calculations using the Sayre–Hughes equation coupled with phase annealing and the Luzzati flatness criterion used as a figure of merit is made.

1. Introduction

Because of their occurrence as two-dimensional arrays in a phospholipid bilayer matrix, integral membrane proteins are most conveniently studied in the electron microscope. The development of electron crystallographic procedures incorporating the processing of low-dose high-resolution electron micrographs from tilted two-dimensional crystals has become a standard technique for determining their structures (Amos, Henderson & Unwin, 1982). Nevertheless, as the resolution of the determination is increased, reliance on experimental images as a sole source of crystallographic phases for the electron diffraction amplitudes becomes more and more of an experimental challenge (Henderson, Baldwin, Downing, Lepault & Zemlin, 1986). Firstly, at the highest resolutions, more sampled pixels are needed to resolve a detail between two points. Even though averaging over the repeat of the two-dimensional space lattice can be used to minimize the actual radiation dose to the specimen when recording the image, the damage induced by inelastic interactions between electron and sample can be problematic (Henderson & Glaeser, 1985). Secondly, many such two-dimensional crystals contain a curvilinear paracrystalline distortion, probably because of the lipid matrix in which the proteins are embedded. Thus, somewhat paradoxically, although the observed electron diffraction pattern might extend to *e.g.* 3 Å, the Fourier transform of a micrograph from a similar area may vanish somewhere in the range from 10 to 6 Å. Lattice 'unbending' has been used to restore the higher-resolution information (Henderson, Baldwin, Downing, Lepault & Zemlin, 1986) but the actual amount of artefact involved is not known. Lastly, there are slight variations of specimen height from the perspective of the microscope objective lens. Although a nearby correction can be made locally for the lens focus before the low-dose image is recorded, the actual transfer function of the micrograph is often unknown. Small deviations can be critical at higher spatial frequencies, affecting rather narrow neighbouring bands of reciprocal space by the rapid change of contrast.

The use of lower-resolution images, which can most easily be recorded in the electron microscope, as the source of crystallographic phases may be a convenient starting point for actual direct phase extensions to higher resolution. This concept was used by Gilmore, Shankland & Fryer (1993), who demonstrated that 15 Å resolution image-derived phases from bacteriorhodopsin can be extended by maximum entropy and likelihood procedures to the diffraction limit to produce potential maps closely resembling those derived solely from high-resolution images. The generality of this approach was demonstrated later when convolutional phase-extension procedures employing the Sayre equation were used successfully for the proteins bacteriorhodopsin, halorhodopsin and the Omp F porin from the outer membrane of *E. coli* (Dorset, Kopp, Fryer & Tivol, 1995; Dorset, 1996).

During the course of these investigations, the possibility of true *ab initio* phase determination, in which it was assumed that no image information was available, was also explored for the first time (Dorset, 1995, 1996) and the Sayre equation in a multisolution environment, followed by phase annealing, was found to produce useful results for the centrosymmetric projection of halorhodopsin (Havelka, Henderson, Heymann & Oesterhelt, 1993) to 6 Å. While the phase accuracy was not so great as found earlier in the phase-extension experiments, the essential features of the protein could be identified, *i.e.* the presence of an α-helix bundle similar to that found for bacteriorhodopsin (Henderson, Baldwin, Downing, Lepault & Zemlin, 1986). Attempts to solve the Omp F porin structure by this method were less successful, however, because the criterion of density 'flatness' (Luzzati, Mariani & Delacroix, 1988), employed as a figure of merit for the annealing step (Dorset, 1995), was not very useful in this case, even though the concept of this density constraint should be quite correct for this resolution range.

In X-ray crystallography, there is an extensive literature concerning *ab initio* phasing at low resolution. For example, Subbiah (1993) has developed a methodology based on the packing and diffraction of hard-sphere point scatterers to generate protein envelopes. Carter, Crumley, Coleman, Hage & Bricogne (1990) have used X-ray contrast variation to define the envelope of *Bacillus stearothermophilus* at 18 Å by phasing the envelope structure factors using the *MITHRIL* computer program (Gilmore, 1984; Gilmore & Brown, 1988). Lunin, Lunina, Petrova, Vernoslova, Urzhumtsev & Podjarny (1995) have determined envelopes at very low resolution (*ca* 50 Å) using pseudo atoms, not unlike those of Subbiah, but in addition employing the use of electron-density histograms.

The maximum entropy (ME) and likelihood approach to phase extension in electron crystallography has proven to be somewhat more robust for phase prediction than the Sayre equation, especially since the starting point (*e.g.* for bacteriorhodopsin) could be constrained to a lower resolution (15 *vs* 10 Å) (Gilmore, Shankland & Fryer, 1993; Dorset, Kopp, Fryer & Tivol, 1995). Additionally, Roth (1991) has used the *MICE* maximum entropy program to study the *Rhodobacter sphaeroides* reaction center and Schluenzen, Volkmann, Thygesen, Hansen, Harms, Bennett & Yonath (1994) have used the same program to study ribosome data at 20 Å resolution. From these results, it seems logical to attempt *ab initio* phasing on membrane data using ME methods and, in this paper, we describe how the procedure successfully overcomes some of the difficulties experienced with the Sayre equation when the true *ab initio* phase determinations are attempted even when working in two dimensions with projection data.

2. Data sets and their processing

Electron crystallographic data from two membrane proteins were used for these calculations, testing two extremes of inherent secondary structure: the Omp F porin from the outer membrane of *E. coli*, which is largely β-sheet, and halorhodopsin, which is predominantly α-helix.

Table 1. *Reflection number, h, k, the unitary structure factor* $|U_h|^{obs}$ *and d in Å for Omp F porin*

An asterisk (*) signifies a centric reflection with a phase constrained to be 0 or π; all the unflagged reflections are acentric with unrestricted phases. All reflections are structure seminvariants.

No.	h	k	$\lvert U_{\mathbf{h}}\rvert^{\rm obs}$	d (Å)
1	5	0	0.10330	12.5*
2	4	−2	0.06775	18.0
3	6	−3	0.05809	12.0
4	4	0	0.04797	15.6*
5	5	−2	0.04424	14.3
6	8	−4	0.04341	9.0
7	5	−1	0.04154	13.6
8	3	0	0.04115	20.8*
9	6	−2	0.03350	11.8
10	2	0	0.03275	31.2*
11	7	−3	0.03104	10.2*
12	7	−2	0.02619	10.0
13	3	−1	0.02552	23.6
14	7	0	0.02456	8.9*
15	9	−3	0.02383	7.9
16	2	−1	0.02363	36.0
17	8	−3	0.02313	8.9
18	11	−4	0.02217	6.5
19	8	−2	0.02135	8.6
20	6	−1	0.02075	11.2
21	8	−1	0.01857	8.3
22	7	−1	0.01812	9.5
23	10	−5	0.01730	7.2
24	9	−4	0.01650	8.0
25	10	−2	0.01465	6.8
26	10	−3	0.01438	7.0
27	9	−2	0.01411	7.6
28	4	−1	0.01262	17.3
29	11	−5	0.01185	6.5
30	9	0	0.01049	6.9*
31	10	0	0.00861	6.2*
32	8	0	0.00722	7.8*
33	9	−1	0.00719	7.3
34	10	−4	0.00702	7.1
35	12	−6	0.00617	6.0
36	11	−1	0.00616	5.9
37	10	−1	0.00612	6.5
38	6	0	0.00467	10.4*
39	12	−4	0.00433	5.9
40	11	−2	0.00400	6.1
41	11	−3	0.00349	6.3
42	1	0	0.00342	62.3*

2.1. *Omp F porin*

The structure of Omp F porin from the outer membrane of *Escherichia coli* (MW 36 500 Daltons) was determined by image analysis to 3.2 Å resolution by Sass, Büldt, Beckmann, Zemlin, van Heel, Zeitler, Rosenbusch, Dorset & Massalski (1989). Originally, data had been obtained at 100 kV from glucose-embedded samples on a liquid-helium-cooled superconducting cryomicroscope with the specimen held at 5 K. Amplitudes and phases from an image of a bimembrane stack were obtained from Dr J. Sass and used as the parent data set for this protein. Since most of the diffracted power was contained within a 6 Å limit (Dorset, 1996), the data resolution was not explored beyond this boundary, so that there were 42 unique reflections in the data set considered. The plane group of the projection is $p31m$ and the hexagonal unit-cell constant is $a = 72$ Å.

The data were normalized using *MITHRIL* (Gilmore, 1984; Gilmore & Brown, 1988) with an imposed overall isotropic temperature factor of zero and using electron scattering factors. Any attempt to derive a temperature factor using Wilson methods (Wilson, 1949) produced overall temperature factors of about 300 Å^2 because of the dominance of very low resolution intensities, the paucity of data and its overall resolution. Table 1 lists the normalized data along with the reflection resolution. The maximum data resolution is 5.9 Å but the effective resolution, defined by the limit for which $|E| > 1.0$, is about 9 Å. Those reflections with a higher resolution that this all have weak amplitudes and will be difficult to phase by any direct method.

2.2. *Halorhodopsin*

Electron diffraction amplitudes and electron-micrograph-derived crystallographic phases from halorhodopsin to 6 Å resolution were reported by Havelka, Henderson, Heymann & Oesterhelt (1993). Original experiments were carried out at 120 kV on frozen-hydrated samples. The centrosymmetric tetragonal plane group is $p4gm$ with lattice constant $a = 102$ Å. Within the 6 Å resolution limit, this corresponds to 76 unique reflections. The data were normalized as for the porin, again with an imposed overall temperature factor of zero, and are listed in Table 2. Although the data have a nominal 6 Å resolution, there are only 16 reflections with a resolution better than 8 Å and these are all weak with a maximum U magnitude of 0.012, which corresponds to an E magnitude of < 0.5. The effective resolution of the data from the viewpoint of direct methods is around 10 Å: those reflections with $|E| > 1.0$ have a maximum resolution of 9.8 Å. This will also have consequences for direct phasing, which will be discussed in the next section.

Table 2. *Reflection number, h, k, the unitary structure factor* $|U_h|^{obs}$ *and d in Å for halorhodopsin*

All reflections are centric with phases constrained to be 0 or π.

No.	h	k	$\|U_{\mathbf{h}}\|^{\rm obs}$	d (Å)	No.	h	k	$\|U_{\mathbf{h}}\|^{\rm obs}$	d (Å)
1	2	2	0.17185	36.1	39	11	5	0.01247	8.4
2	2	0	0.14390	51.0	40	11	8	0.01242	7.5
3	1	1	0.11227	72.1	41	10	0	0.01222	10.2
4	3	3	0.10488	24.0	42	9	1	0.01198	11.3
5	3	1	0.07983	32.3	43	8	5	0.01140	10.8
6	6	5	0.07672	13.1	44	6	2	0.01104	16.1
7	6	3	0.06466	15.2	45	7	5	0.01094	11.9
8	8	0	0.05879	12.8	46	8	7	0.01057	9.6
9	8	1	0.04851	12.7	47	14	1	0.01033	7.3
10	2	1	0.04610	45.6	48	4	4	0.01030	18.0
11	5	5	0.04589	14.4	49	6	6	0.01028	12.0
12	7	4	0.04536	12.6	50	10	5	0.00986	9.1
13	10	3	0.04465	9.8	51	9	4	0.00975	10.4
14	6	0	0.04013	17.0	52	8	8	0.00953	9.0
15	3	2	0.03410	28.3	53	8	6	0.00808	10.2
16	4	3	0.03335	20.4	54	7	2	0.00792	14.0
17	4	1	0.03215	24.7	55	13	3	0.00760	7.6
18	6	4	0.03006	14.1	56	13	4	0.00718	7.5
19	7	6	0.03005	11.1	57	10	4	0.00642	9.5
20	10	2	0.02928	10.0	58	12	7	0.00642	7.3
21	9	5	0.02862	9.9	59	12	4	0.00604	8.1
22	5	1	0.02854	20.0	60	11	3	0.00587	8.9
23	5	2	0.02774	18.9	61	9	7	0.00569	8.9
24	6	1	0.02692	16.8	62	10	9	0.00562	7.6
25	4	2	0.02474	22.8	63	10	6	0.00555	8.7
26	8	3	0.02155	11.9	64	4	0	0.00494	25.5
27	5	4	0.02146	15.9	65	11	7	0.00445	7.8
28	8	2	0.02023	12.4	66	15	8	0.00444	6.0
29	10	7	0.01990	8.4	67	11	1	0.00353	9.2
30	7	1	0.01888	14.4	68	13	5	0.00306	7.3
31	11	2	0.01877	9.1	69	15	2	0.00276	6.7
32	7	7	0.01869	10.3	70	9	8	0.00267	8.5
33	5	3	0.01795	17.5	71	13	8	0.00257	6.7
34	8	4	0.01556	11.4	72	13	1	0.00256	7.8
35	11	6	0.01535	8.1	73	14	7	0.00218	6.5
36	12	0	0.01453	8.5	74	14	5	0.00216	6.9
37	9	3	0.01404	10.7	75	13	9	0.00215	6.4
38	9	2	0.01344	11.1	76	14	0	0.00208	7.3

3. Phase determination

3.1. *The ME method*

The technique used here is that of multisolution constrained entropy maximization combined with likelihood evaluation as a source of selecting the most probable phase choices. The theory comes from Bricogne (1984, 1988*a*,*b*, 1993) as implemented in the *MICE* computer program (Gilmore, Bricogne & Bannister, 1990; Bricogne & Gilmore, 1990; Shankland, Gilmore, Bricogne & Hashizume, 1993; Gilmore, Shankland & Bricogne, 1993; Gilmore, 1996). For recent applications to small-molecule electron diffraction data sets, see Voigt-Martin, Yan, Gilmore, Shankland & Bricogne (1994) and Voigt-Martin, Yan, Yakimansky, Schollmeyer, Gilmore & Bricogne (1995), and for phase extension applied to membrane proteins, see Gilmore, Shankland & Fryer (1993). The methods described in these references closely match the procedures used here with only small resolution-based differences; however, for completeness, a brief overview of the formalism is presented here:

(i) The diffraction intensities are normalized to give unitary structure factors, $|U_{\mathbf{h}}|^{\rm obs}$ and their associated standard deviations $\sigma_{\mathbf{h}}$.

(ii) The observed U magnitudes are partitioned into two sets: the basis set $H = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$ comprises the reflections for which reliable phase information $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$ for n reflections is available. In this case, $\{H\}$ comprises only the origin-defining reflections (where any exist) with their associated phases. The set $\{H\}$ defines the root node of a phasing tree. The disjoint set $\{K\}$ comprises the remaining unphased reflections.

(iii) The basis set reflections (both phase and amplitude) are used as constraints in an entropy maximization to give a maximum entropy distribution $q^{\rm ME}(\mathbf{x})$, which reproduces the intensities and phases of the basis set but also has Fourier coefficients $U_{\mathbf{k}}^{\rm ME}$ with non-negligible amplitude for many non-basis set reflections. This is the process of maximum entropy extrapolation.

(iv) Initially, the extrapolation is usually too weak to be of any value, so new phase information is incorporated into the basis set by adding strong reflections, which are hitherto inconclusively extrapolated and which optimally enlarge the second neighbourhood of the current basis set. (The second neighbourhood is defined by reflections $\mathbf{h}_1 \pm {}^t\mathbf{R}_g \cdot \mathbf{h}_2$ for $\mathbf{h}_1, \mathbf{h}_2 \in \mathbf{H}$, where ${}^t\mathbf{R}_g$ is the transpose of a rotation matrix obtained from the crystal space group.) Since the phases of these new reflections are unknown, this gives rise to series of phase choices in which each centric reflection is given both of its possible values (*e.g.* 0, π or $\pm\pi/2$) and each acentric reflection is quadrant fixed using the choices $\pm\pi/4$, $\pm3\pi/4$. Each phase choice is represented as a node on the second level of the phasing tree. If there are n_c centric phases and n_a acentric phases to be added to the basis set, this will generate $2^{n_c}4^{n_a}$ nodes.

(v) Each node on the tree is now subjected to constrained entropy maximization just as before. To rank the nodes, hopefully in order of phase error, a Rice-type likelihood function is used, which evaluates the agreement between the extrapolated structure-factor magnitudes from the relevant maximum entropy distribution and the experimentally measured ones. The log-likelihood gain (LLG) will be largest when the phase assumptions for the basis set lead to predictions of deviations from the Wilson distribution in the unphased reflections, which in turn mirror the measured intensities, and in this context the LLG is used as a powerful figure of merit (Bricogne, 1984, 1993; Bricogne & Gilmore, 1990; Gilmore, Bricogne & Bannister, 1990).

(vi) The LLGs are analysed for phase relationships using the Student t test, which defines the level of significance in the contrast between two means (Shankland, Gilmore, Bricogne & Hashizume, 1993). The simplest t test involves the detection of the main effect associated with the sign of each single centric phase that has been permuted. The LLG average, μ^+, and its associated variance V^+ is computed for those nodes in which the sign of the phase under test is +. The calculation is then repeated for those nodes in which the same sign is − to give the corresponding μ^-, and variance V^-. The t statistic is then

$$t = |\mu^+ - \mu^-|/(V^+ + V^-)^{1/2}. \qquad (1)$$

The use of the t test enables a sign choice to be derived with an associated significance level. This calculation is repeated for all the single-phase indications and is then extended to combinations of two and three phases. A further extension to acentric phases is straightforward by employing two signs to define the phase quadrant. In general, only relationships with associated significance levels <2–10% are used but this is sometimes relaxed with sparse data sets.

(vii) The phase relationships obtained from the t test are used to identify the best nodes (in the sense of minimum phase error). The optimum method for this is a high-dimensional Fourier transform as used in the *BUSTER* program (Bricogne, 1993), which exploits the periodicity of phase angles and their phase relationships. Here we employ a simpler, but effective, algorithm:

(*a*) The one-, two- and three-phase relationships are derived from the t test, and given an associated weight proportional to their significance level.

(*b*) Each node is assigned a score related to the measure of agreement between the phase relationships weighted by the associated significance levels from (*a*) and the phases in the set itself.

(*c*) The eight sets from (*b*) having the highest scores are kept. Alternatively, the phase relationships from the t tests can be used to determine unique phases for one or more basis set reflections and a new phasing tree is started using these values.

(viii) The tree building and pruning procedure is continued until most large unitary structure factors have significant phase indications. Potential maps are generated as *centroid maps* by means of a Sim-type filter in which each reflection is given a coefficient $|U_{\mathbf{h}}|^{\text{obs}}$, a phase from $U_{\mathbf{h}}^{\text{ME}}$ and an associated weight $w_{\mathbf{h}}$ (Bricogne & Gilmore, 1990) computed as follows:

$$w_{\mathbf{h}} = \tanh(N/\varepsilon_{\mathbf{h}}|U_{\mathbf{h}}|^{\text{obs}}|U_{\mathbf{h}}^{\text{ME}}|) \quad \text{for } \mathbf{h} \text{ centric} \qquad (2)$$

$$w_{\mathbf{h}} = I_1(X_{\mathbf{h}})/I_0(X_{\mathbf{h}}) \quad \text{for } \mathbf{h} \text{ acentric}, \qquad (3)$$

where

$$X_{\mathbf{h}} = (N/\varepsilon_{\mathbf{h}})|U_{\mathbf{h}}|^{\text{obs}}|U_{\mathbf{h}}^{\text{ME}}|, \qquad (4)$$

N is the number of atoms in the unit cell and $\varepsilon_{\mathbf{h}}$ is the statistical weight.

In practice, for both structures presented here, the procedures used were quite automatic; the only user decisions were the resolution limit to use and the number of nodes to generate. Variations of these parameters still produced good phase sets so that the results are not unduly sensitive to initial phase choices. The total computer time involved was less than one hour on a 15 processor workstation network for both structures.

3.2. *Omp F porin*

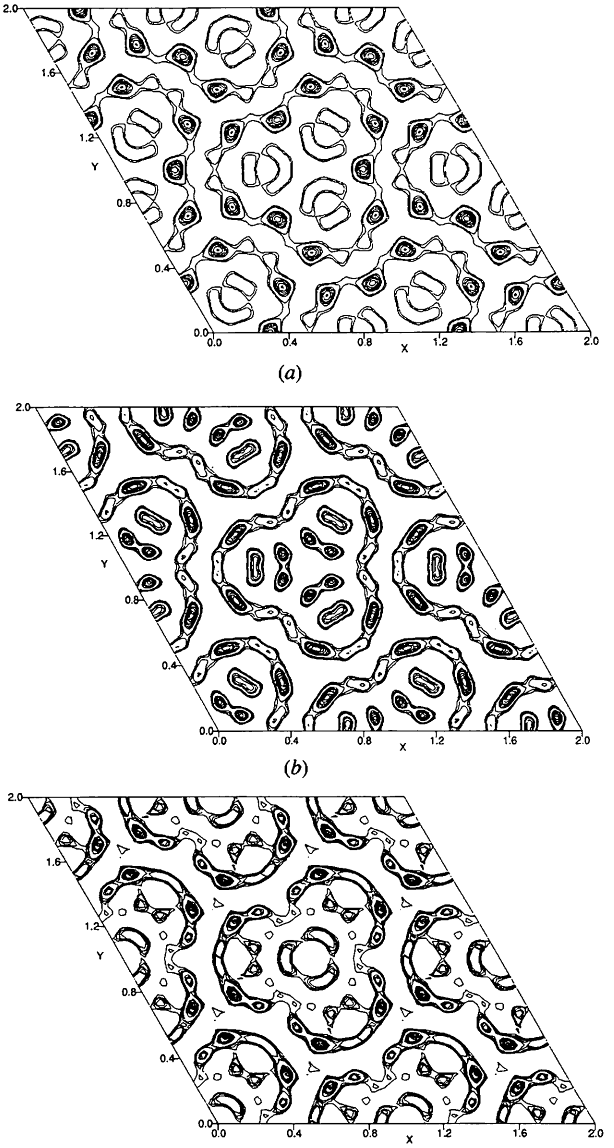
In the plane group $p31m$, all the reflections are structure seminvariants and the group defines the origin. Also, all reflections except the ($h0$) are acentric. To commence phasing, the top five centric reflections were given permuted phases so generating 32 nodes, each of which was subjected to entropy maximization. Analysis of the LLG gave unambiguous phase assignments for three of these, of which the indication for reflection 1 (50) ($= \pi$) was the strongest. All three indications were correct; this is in sharp contrast with the use of the traditional Σ_1 formula in *MITHRIL*, where two of the three phases were incorrectly predicted. Accordingly, a new starting point was defined by fixing the phase of reflection (50) ($|U_{\mathbf{h}}^{\text{obs}}| = 0.103$) with an angle of π. This single phased reflection comprised the basis set for the root node of a phasing tree. Three reflections were then selected for phase permutation *via* the algorithm of optimum second-neighbourhood enhancement (Bricogne, 1993; Gilmore, Bricogne & Bannister, 1990). These were the $(4\bar{2})$, $(6\bar{3})$ and (40) with corresponding U magnitudes of 0.068, 0.058 and 0.048, respectively. The first two reflections are acentric and were given permuted phases of $\pm\pi/4, \pm3\pi/4$ while the latter is centric and was given the values 0 or π, thus generating 32 nodes, each of which was subjected to constrained entropy maximization. The basis set resolution at this point was 12 Å. The LLGs were subjected to statistical analysis and scoring as described above. Three nodes 29, 31 and 32 had scores much higher than any other phase sets and thus survived the analysis; they had corresponding basis set phase errors of 13, 49 and 27° with LLGs of 0.023, 0.007 and 0.033, respectively. It is not possible to choose between these nodes on the grounds of LLG or score alone, although node 31 is less strongly indicated and each bifurcation of the phasing process that these represent must be followed. Fig. 1(*a*) shows the potential map for the node having a mean phase error of 13°. Even at this stage with only four unique reflections in the basis set, the molecular outline is quite clear. The correlation coefficient using as coefficients $U^{\rm obs}_{\bf h} w_{\bf h}$ and $U^{\rm ME}_{\bf h}$ to the full resolution of the data was 0.93.

Thus, all three surviving nodes were kept and the phase-determination process was continued for two further levels of the phasing tree, extending the resolution limit to 9 Å, as summarized in Table 3. At this stage, the preferred potential map is shown in Fig. 1(*b*) and the true map using the image-derived phases of Sass *et al.* (1989) is shown for comparison purposes in Fig. 1(*c*). At this resolution, the preferred map has a basis set mean absolute phase error of only 9°. With only minor details, there is an essential correspondence with this map and one computed with all correct angles from the image data with a correlation coefficient of 0.94. Extension to 6 Å *via* level 5 of the phasing tree resulted in increased phase errors – the best solution now had an error of 23° but the maps were substantially correct, with no important extra detail appearing when compared to Fig. 1(*c*); there was merely an increase in the number of contour levels. This is not surprising given the weak amplitudes of the 6 Å data and the corresponding difficulties in phasing them.

3.3. *Halorhodopsin*

Surprisingly, because it is a centric projection, halorhodopsin was more difficult to phase accurately than the porin. Two methods were used each with complementary strengths:

(i) An origin was first defined by fixing the phase of the (21) reflection ($|U^{\rm obs}_{\bf h}| = 0.172$) with a phase angle of 0°. This was chosen to match that of the published structure and it comprised the basis set for the root node of the phasing tree. One reflection is sufficient to defined the origin for this plane group. Six reflections with a maximum resolution of 15 Å were then selected for phase permutation. These are listed in Table 4. All the reflections are centric and so were given the phase values 0 or π, thus generating $2^6 = 64$ nodes. As before, each node was subjected to entropy maximization and its corresponding LLG evaluated. Those eight nodes with the maximum scores were kept; the one with the third highest LLG had a corresponding basis set phase error of 0°. The centroid potential map corresponding to this node is shown in Fig. 2(*a*); it has a



(*a*)

(*b*)

(*c*)

Fig. 1. Centroid maps for Omp F porin: (*a*) based on a basis set of four unique reflections with a mean phase error of 13° at 12 Å and a correlation coefficient of 0.93; (*b*) based on ten basis set reflections with a mean phase error of 9° at 9 Å and a correlation coefficient of 0.94; (*c*) a true map based on image-derived phases.

Table 3. *The phasing tree for Omp F porin*

Level	Permute (or fix)	Number of nodes generated	Number of nodes kept after analysis	Mean phase errors of nodes kept (°)	Map correlation coefficients	Notes
1	(5 0)	1	1	0		
2	(4 −2)	32	3	13	0.95	12 Å resolution, best solution ranked second
	(6 −3)			49	0.53	
	(4 0)			32	0.79	
3	(5 −1)	64 × 3	6	25	0.81	12 Å resolution
	(3 0)			18	0.85	
	(6 −2)			9	0.93	
	(2 0)			56	0.45	
				40	0.55	
				26	0.78	
4	(5 −2)	16 × 6	8	30	0.76	9 Å resolution
	(8 −4)			38	0.71	
				45	0.65	
				24	0.80	
				16	0.87	
				31	0.76	
				17	0.88	
				9	0.94	
5	(7 0)	32 × 8	8	51	0.65	6 Å resolution
	(2 −1)			57	0.59	
	(11 −4)			32	0.80	
				23	0.87	
				29	0.84	
				27	0.85	
				26	0.86	
				23	0.87	

Table 4. *The phasing tree for halorhodopsin*

Level	Permute	Number of nodes generated	Number of nodes kept after analysis	Mean phase errors of nodes kept (°)	Map correlation coefficients	Notes
1	(2 1)	1	1	0		Origin
2	(2 2)	64	8	27	0.79	15 Å resolution, best solution ranked third
	(2 0)			24	0.17	
	(1 1)			10	0.91	
	(3 3)			67	0.26	
	(6 3)			17	0.83	
	(6 0)			74	0.22	
				0	0.96	
				7	0.32	
3	(6 5)	32 × 8	8	41	0.66	11 Å resolution, best solution ranked second
	(8 0)			39	0.75	
	(8 1)			50	0.70	
	(5 5)			32	0.74	
	(6 4)			43	0.69	
				21	0.82	
				31	0.79	
				65	0.28	

correlation coefficient, as defined in the previous section, of 0.91. The phasing process was continued for another level extending the resolution to 11 Å, as summarized in Table 2, by permuting the phases of five further reflections. The best potential map obtained from this approach is shown in Fig. 2(*b*); it is ranked second in terms of LLG and has a mean phase error of 21° with a correlation coefficient of 0.82. For comparison, the true map using the image-derived phases of Havelka *et al.* (1993) is shown in Fig. 2(*c*). Attempts to increase the resolution to 6 Å were only partially successful because the reflections that we wished to phase had such low U magnitudes. Mean basis set phase errors increased to 23° for the preferred node and the corresponding maps were much poorer in quality, showing less accurate detail than those at 11 Å, although the gross features of the structure were still present.

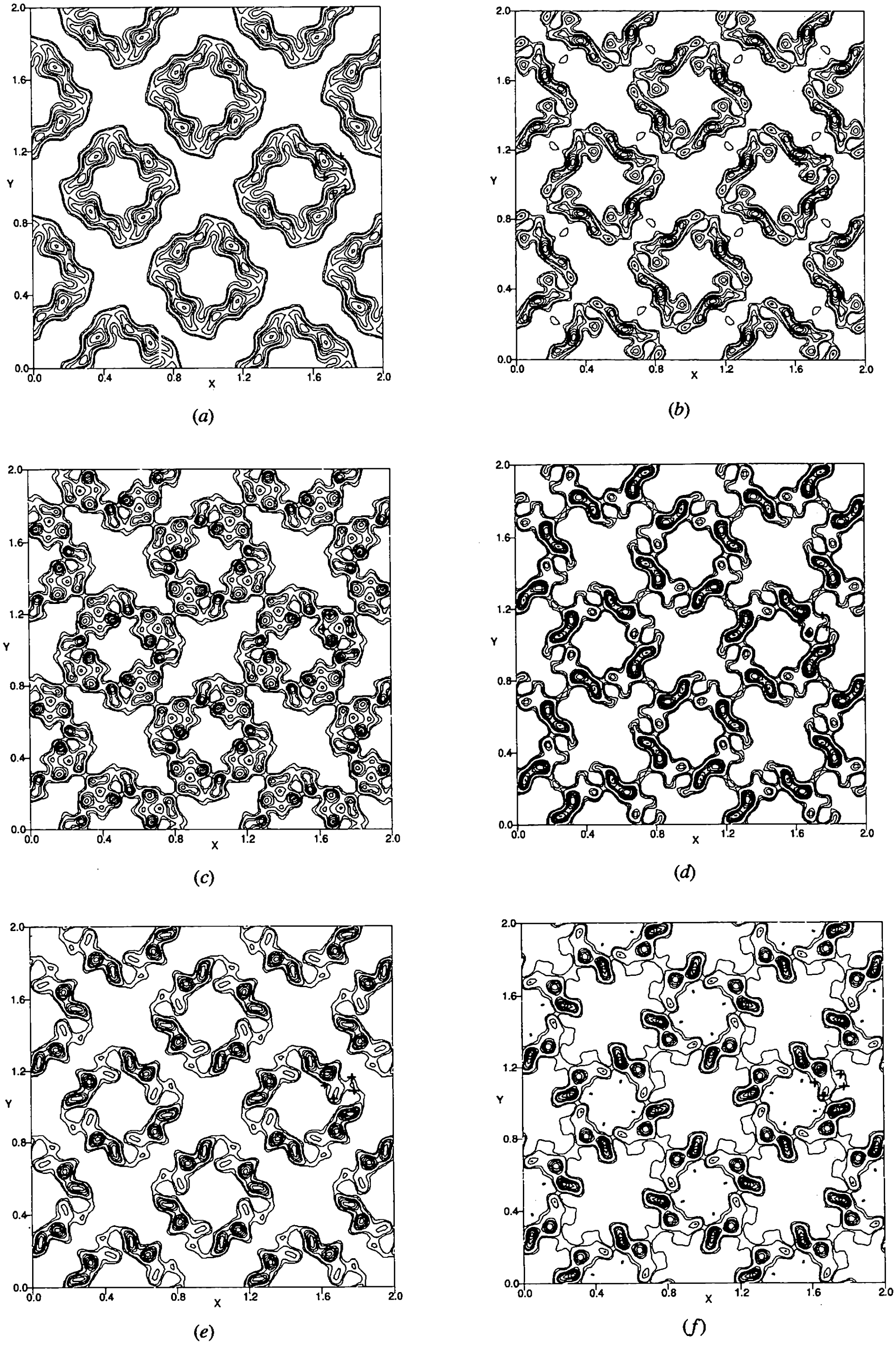


Fig. 2. Centroid maps for halorhodopsin: (*a*) based on a basis set of 7 unique reflections with a mean phase error of 0° at 15 Å, the correlation coefficient is 0.9; (*b*) based on 12 basis set reflections with a mean phase error of 21° at 11 Å, the correlation coefficient is 0.82; (*c*) a true map based on image-derived phases; (*d*) based on a basis set of 8 unique reflections with a mean phase error of 0° at 15 Å, the correlation coefficient is 0.97; (*e*) based on a basis set of 16 unique reflections with a mean phase error of 11° at 15 Å, the correlation coefficient is 0.90; (*f*) based on a basis set of 28 unique reflections with a mean phase error of 9° at 6 Å, the correlation coefficient is 0.87. The crosses mark the positions of the helices (from Havelka, Henderson, Heymann & Osterhelt, 1993).

(ii) As an alternative approach, the origin was left undefined initially and eight reflections with a maximum resolution of 15 Å were given permuted phases. This generated 256 nodes. Analysis of the associated LLGs using the t test gave unambiguous and correct single phase indications for six of these reflections, all of which were structure seminvariants. In addition, the phase of reflection 6, $\varphi(6)$, was indicated to be $\pi+\varphi(7)$. Fixing the origin by setting $\varphi(6)=\pi$ to match the correct structure resulted in defining $\varphi(7)$ also. At this point, eight reflections were uniquely phased and the associated centroid map is shown in Fig. 2(*d*). The correlation coefficient at this point was 0.97. These phases were used to define the root node of a new phasing tree. A further eight reflections in the resolution range 6–10 Å were now permuted. The lowest U magnitude in this set was 0.006. This corresponds to an E of 0.24, which is a magnitude normally inaccessible to routine direct methods. From LLG analysis, all eight reflections were unambiguously phased and of these only one was wrong. The corresponding map is shown in Fig. 2(*e*).

At this point, there was sufficient detail in the centroid map to make use of the $P(\delta q)$ function (Bricogne, 1984; Gilmore, Bricogne & Bannister, 1990). Although the method is described in detail in these two references, some discussion concerning its use may be helpful. $P(\delta q)$ can be used with any maximum entropy map $q^{\rm ME}(\mathbf{x})$. A set of reflections is chosen using the standard criteria and their phases permuted. Each phase permutation gives rise to a difference map $\delta q(\mathbf{x})$, which is a Fourier synthesis using the coefficients $U_{\mathbf{h}}^{\rm obs} - U_{\mathbf{h}}^{\rm ME}$, where $U_{\mathbf{h}}^{\rm ME}$ is the Fourier coefficient of $q^{\rm ME}(\mathbf{x})$. $P(\delta q)$ is calculated for each phase permutation from

$$\int_{V} [\delta q^2(\mathbf{x})/q^{\rm ME}(\mathbf{x})]\,{\rm d}^3\mathbf{x}. \qquad (5)$$

A minimum value of $P(\delta q)$ is expected for the correct phase set. This technique has several useful features:

(*a*) Each phase permutation requires only one Fourier synthesis and a map division and so is very fast, although computer time is not a problem with these data.

(*b*) It acts as a pre-filter for entropy maximization; the reflections to be permuted are first subjected to this filter and only those with a certain minimum $P(\delta q)$ (typically 10–25%) are passed on for entropy maximization.

(*c*) The likelihood function uses only moduli, but $P(\delta q)$ incorporates phases into the calculation. It therefore acts as a useful tool in exploring structure-factor space from the current node.

The technique preferentially selects those phases that build density where it is already well defined. This is very appropriate in this situation where we have a clear envelope. It does, however, need to be used with care. In particular, $q^{\rm ME}(\mathbf{x})$ must have developed sufficient detail. Those $q^{\rm ME}(\mathbf{x})$ maps based on a very small basis sets or utilizing only small U magnitudes may have insufficient contrast for its successful use. In this case, one obtains a set of $P(\delta q)$ values that are virtually constant. It is also advantageous to choose reflections for which there is a small but finite extrapolated magnitude from the current $q^{\rm ME}(\mathbf{x})$.

A set of 12 reflections in the resolution range 6–15 Å were selected, giving rise to 4096 phase sets, each of which was filtered *via* $P(\delta q)$. The 100 phase sets with the lowest values of $P(\delta q)$ were kept and passed on to entropy maximization. The best eight nodes based on LLG estimates were kept. The centroid map corresponding to the best of these is shown in Fig. 2(*f*). The basis set reflections have a mean phase error of 9° for 28 reflections and a correlation coefficient of 0.90. Attempts to phase beyond this point were as unsuccessful as in (i).

4. Discussion

The results of Omp F porin are impressive and have surprised even the authors: after four levels of a phasing tree with basis set phases extending to 9 Å, there are eight nodes that survive likelihood analysis. One of these has a U-weighted mean absolute phase error of only 9°; the corresponding map is shown in Fig. 1(*b*) and comparison with the true map is really striking, as is to be expected with a correlation coefficient of 0.94. When one considers that this is an *ab initio* phasing of a membrane protein from electron crystallographic data in projection, this is a remarkable result.

There was a similar, if not so pronounced, success with the halorhodospin data set. At 15 Å, an envelope is found that strongly resembles the one found earlier in the multisolution investigation by Dorset (1995). Extension to higher resolution was also found to produce maps where most of the features of the α-helix bundle could be discerned, although some detail was still obscure when compared with the true map; in particular, the envelope area is rather smaller than expected and, whereas most of the density in the best maps is in the correct places, it can have false weights. The positions of four of the helices are unambiguously indicated, whilst the remainder are in regions of high density but unresolved. The development of further detail requires further phasing of very small U magnitudes, which will always be difficult.

Comparison of the ME approach with the methodology developed by Dorset (1995, 1996) is interesting. The latter uses the Sayre–Hughes equation for phase extension from a small starting set, coupled with phase annealing to refine phases and the Luzzati *et al.* (1988) flatness criterion of minimum $\langle\Delta\rho^4\rangle$ as a figure of merit. For the case of halorhodopsin, this method yielded a basis set of 20 strong reflections at 15 Å, of which 6 were incorrect (this corresponds to a mean phase error of 54°) and, at 10 Å, 6 were wrong out of 23 (a mean phase error of 46°). It is clear that the ME–likelihood formalism is working better, but the use of phase annealing does allow incorrect phases to be refined. This is currently rather difficult and unreliable in the ME–likelihood approach as presented here. In terms of map quality, which must always be the final arbiter, the two methods produce maps that are not too dissimilar, although the centroid maps have a larger dynamic range.

At this point, the use of entropy as a figure of merit needs discussion. As with all the ME *ab initio* structure determinations we have carried out (see, for example, Gilmore, Bricogne & Bannister, 1990; Gilmore, 1996), entropy, S, was a poor indicator of phase correctness for both structures. This is not surprising: the true role of entropy in this probabilistic context is, to quote Bricogne (1984), 'a quantitative measure of the extent to which the range of structures which can be generated with any likelihood can be narrowed down'. It *can* be used as a measure of map flatness or dynamic range but entropy is calculated from the maximum entropy map, $q^{\rm ME}({\bf x})$, not the potential map itself and, furthermore, there is no reason why the correct structure should exhibit an entropy maximum, except perhaps at very low resolution; quite often, the correct phase set has an entropy minimum rather than a maximum.

The Bayesian score $NS + LLG$, which can be used as a composite figure of merit instead of likelihood for individual nodes was equally recalcitrant. In this case, there are two problems both concerned with N:

(i) The definition of N itself: in this work, it is usually taken to be the number of atoms in the unit cell, although it can be treated as a refinable parameter and optimized *via* likelihood. However, the accurate definition of N at 6 Å is difficult.

(ii) N itself is critical: if it is too large, the entropy measure swamps the LLG and, if it is too small, the entropy does not make a significant contribution to the Bayesian score.

In our experience, $NS + LLG$ is much more useful in cases where a large, albeit approximate, basis set is available from, for example, MIR phases in X-ray protein crystallography.

The use of the Luzzati criterion is not unrelated to using entropy as a figure of merit and, as discussed in the previous paragraph, entropy is at best only suitable as a secondary indicator in *ab initio* studies, and so the Luzzati test could be unreliable. In view of the comments concerning entropy, why does it work at all?

(i) The Luzzati criterion is applied to potential maps where negative pixels can be present, not maximum entropy ones where positivity is imposed.

(ii) It is applied at very low resolution where the flatness of a potential map can discriminate between possible phase choices.

(iii) It is a different function to entropy (minimum $\langle \Delta\rho^4 \rangle$ as against maximum $-\sum_i p_i \log p_i$).

In the case of the porin, the ME formalism produces much lower phase errors and more accurate maps than the Luzzati-annealing method. In this case, all the nodes on the phasing tree have very similar entropies, which would indicate that the Luzzatti criterion is an insensitive figure of merit here, which is born out by the results. However, phase annealing may prove to be a very useful adjunct to the ME–likelihood method and this needs to be tested in this context. Why the porin should be easier to solve is not clear – it could reflect exploitation of the threefold axis by the membrane molecule or simply good luck; data quality is unlikely to be the reason.

Although the inherent capability of electron-micrographic techniques for providing high-resolution images of a crystalline object is an advantage not enjoyed in any other branch of crystallography, it is, nevertheless, important to consider what can be achieved if only diffraction data are available. For example, before these trial experiments on representative membrane protein data sets, it was not known that phase determination could be so effective at such low resolutions. This is because, in X-ray crystallography, much of the two-angle intensity data needed for definition of the molecular envelope often are simply not recorded due, for example, to geometrical limitations imposed by the apparatus. Carter *et al.* (1990) and Fan, Hao & Woolfson (1991) have mentioned, however, that, since most of the diffracted energy from protein crystals is contained at low resolution, the phase relationships between these low-angle reflections should not be any less valid than those, for example, from small molecules, even though there are fewer interactions than would be found at atomic resolution. Phase extensions at low angle (Gilmore, Shankland & Fryer, 1993; Dorset, Kopp, Fryer & Tivol, 1995; Dorset, 1996), therefore, have proven to be quite effective, yielding results with accuracy quite consistent with the findings of earlier low-resolution extensions with X-ray data (Reeke & Lipscomb, 1969; Podjarny, Schevitz & Sigler, 1981). In agreement with previous observations, therefore, there appear to be two regions of the diffraction pattern from proteins equally amenable to phase determination, *i.e.* at very high and very low resolutions [see Weeks, Hauptman, Smith, Blessing, Teeter & Miller (1995) for the former case]. The criteria for finding a correct structure correspond, respectively, to the Cochran (1952) condition for 'peakiness' or the exact opposite of density flatness or smoothness (Luzzati, Mariani & Delacroix, 1988). The intermediate region near 5 Å, however, is a boundary zone where most difficulties with macromolecular phasing occur (Podjarny & Yonath, 1977; Dorset, Kopp, Fryer & Tivol, 1995) and it remains to be seen how well the ME formalism can span this region.

Although it might be accepted that low-resolution phase extensions might be feasible, given the presence of an accurate basis phase set from the Fourier transform of an electron micrograph, the success of true *ab initio* determinations, demonstrated in this paper, was not necessarily expected. It is clear that the most success is experienced in the lowest-resolution range with some degradation of accuracy occurring as the resolution limit is extended. However, for the two examples investigated, much of the important structural detail is already manifested at about 10 Å so the determinations are still sufficient to discern the essential features of the protein without need of an electron micrograph. For these procedures to be made more effective, however, an efficient and accurate method for phase refinement must be found to improve the basis set after a given resolution shell is reached. While phase annealing has been shown earlier to be useful, the density flatness figure of merit monitored with it is not always capable of discerning phase changes in favour of the most correct value. The challenge of phase refinement, therefore must be faced in future work in this area.

Finally, the ME formalism can also use a non-uniform prior in the form of an envelope (Bricogne, 1984); this has been programmed and tested in this environment. However, this is an *ab initio* study and so the envelope cannot be considered known *a priori* but, even when it is included, preliminary calculations indicate that the method works no better. This is probably because the large U magnitudes that we are phasing generate sufficient contrast in the centroid maps, which makes the use of envelopes unnecessary. This may not always be the case, however, and high-resolution electron micrographs from sugar-embedded or frozen-hydrated preparations could give useful envelope information to assist difficult phase-determination processes.

This research was funded in part by grants from the Human Frontier Science Program and the National Institute of Medical Sciences (GM-46733).

References

Amos, L. A., Henderson, R. & Unwin, P. N. T. (1982). *Prog. Biophys. Mol. Biol.* **59**, 183–231.

Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.

Bricogne, G. (1988*a*). *Acta Cryst.* A**44**, 517–545.

Bricogne, G. (1988*b*). *Crystallographic Computing 4: Techniques and New Technologies*, edited by N. W. Isaacs & M. R. Taylor, pp. 60–79. IUCr/Oxford University Press.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.

Bricogne, G. & Gilmore, C. J. (1990). *Acta Cryst.* A**46**, 284–297.

Carter, C. W. Jr, Crumley, K. V., Coleman, D. E., Hage, F. & Bricogne, G. (1990). *Acta Cryst.* A**46**, 57–68.

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

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

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

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

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

Gilmore, C. J. (1984). *J. Appl. Cryst.* **17**, 42–46.

Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.

Gilmore, C. J., Bricogne, G. & Bannister, C. (1990). *Acta Cryst.* A**46**, 297–308.

Gilmore, C. J. & Brown, S. R. (1988). *J. Appl. Cryst.* **21**, 571–572.

Gilmore, C. J., Shankland, K. & Bricogne, G. (1993). *Proc. R. Soc. London Ser. A*, **442**, 97–111.

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

Havelka, W. A., Henderson, R., Heymann, J. A. W. & Oesterhelt, D. (1993). *J. Mol. Biol.* **234**, 837–846.

Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. (1986). *Ultramicroscopy*, **19**, 147–178.

Henderson, R. & Glaeser, R. M. (1985). *Ultramicroscopy*, **16**, 139–150.

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.

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

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

Podjarny, A. D. & Yonath, A. (1977). *Acta Cryst.* A**33**, 655–661.

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

Roth, M. (1991). *Crystallographic Computing 5: from Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Theirry, pp. 229–248. IUCr/Oxford University Press.

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

Schluenzen, F., Volkmann, N., Thygesen, J., Hansen, H. A. S., Harms, J., Bennett, W. S. & Yonath, A. (1994). Proceedings of the American Crystallographic Association Meeting, Atlanta, USA, TRN15.

Shankland, K., Gilmore, C. J., Bricogne, G. & Hashizume, H. (1993). *Acta Cryst.* A**49**, 493–501.

Subbiah, S. (1993). *Acta Cryst.* D**49**, 108–119.

Voigt-Martin, I. G., Yan, D. H., Gilmore, C. J., Shankland, K. & Bricogne, G. (1994). *Ultramicroscopy*, **56**, 271–288.

Voigt-Martin, I. G., Yan, D. H., Yakimansky, A., Schollmeyer, D., Gilmore, C. J. & Bricogne, G. (1995). *Acta Cryst.* A**51**, 849–868.

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.