RESEARCH PAPERS

Acta Cryst. (1996). D52, 893–900

Application of Known X-ray Phases in the Crystallographic Study of a Small Protein

F. Mo,[a]* R. H. Mathiesen,[a] B. C. Hauback[b] and E. T. Adman[c]

[a]*Institutt for fysikk, Norges Teknisk-Naturvitenskapelige Universitet, N-7034 Trondheim, Norway,* [b]*Fysikkavdeligen, IFE, Pb 40, N-2007 Kjeller, Norway, and* [c]*Department of Biological Structure SM-20, and Department of Biochemistry, University of Washington, Seattle WA 98195, USA*

(Received 2 January 1996; accepted 25 March 1996)

Abstract

Phase information, assumed known from three-beam diffraction experiments, has been used successfully as input to direct methods to re-determine the crystal structure of rubredoxin. With data at 1.54 Å resolution, starting sets containing 26 single phases, or alternatively 45 triplet phases, in both cases known with a mean error of ±22.5°, were sufficient to initiate solution of the structure. Conventional figures of merit were employed in the early stages to reject the majority of the incorrect phase models. The presence of a FeS_4 cluster in the structure was used in the interpretation of the initial maps. Phase sets including 2500 E's with a mean single phase error $\langle\Delta\varphi\rangle \simeq 70°$ or a mean triplet phase error $\langle\Delta\Phi_3\rangle \simeq 80°$, both relative to the model from the crystallographic refinement, could be refined and expanded by Fourier recycling using the *SAYTAN* formalism. Several parameters have been varied to study their influence on phase expansion and refinement.

1. Introduction

Over the past decade a physical approach to structure-factor phase acquisition has developed very strongly, both theoretically and experimentally, following the first reports on phase effects in three-beam diffraction experiments with mosaic crystals (Post, 1977; Chang, 1982; Thorkildsen & Mo, 1982, 1983; Gong & Post, 1983). Diffraction experiments with crystals of myoglobin (Hümmer, Schwegle & Weckert, 1991) and hemoglobin (Chang, King, Huang & Gao, 1991) have indicated that even proteins are within reach of physical phase estimation (PPE). More recently Weckert, Schwegle & Hümmer (1993) have provided compelling evidence of successful PPE with protein crystals. They reported estimation of about 80 triplet phases from a crystal of tetragonal lysozyme, unit-cell volume $V = 237\,180$ Å^3, with a mean phase deviation less than 20° compared to the calculated structure-factor phases of the refined structure model. This work has been extended later to include about 600 estimated triplet phases (Weckert, 1996).

Since complete phase sets cannot be measured, the usefulness of PPE for macromolecular work depends on how the phase sets can be extended and refined. In this paper we assume that a certain number of known phases acquired from three-beam diffraction experiments, or by other methods, for instance from multiple anomalous dispersion, are available and can be used as a large starting set for further expansion by direct methods. Under this assumption, primary questions for study are: is there an optimal resolution (range) for the data to obtain a structure solution? What is the optimal strategy for phase expansion? Which reflections are most efficient and how many known phases are needed for promoting phase expansion? What are the reliable figures of merit for identifying the best phase models in the early stages? The answers to some of these questions are clearly dependent on structural complexity and space-group symmetry. Several studies of reciprocal-space methods for the solution of macromolecular structures have been made from different starting principles (Sayre, 1972, 1974; Agarwal & Isaacs, 1977; Tsoucaris, 1970; de Rango *et al.*, 1985; Bricogne, 1984, 1993; Sjölin & Svensson, 1993; Woolfson & Yao, 1990; Sheldrick, 1990; Sheldrick, Dauter, Wilson, Hope & Sieker, 1993). There are now also powerful methods combining reciprocal-space refinement with real-space filtering (Miller *et al.*, 1993; Weeks, DeTitta, Hauptman, Thuman & Miller, 1994).

2. Scope of the present work

In the present study we have used rubredoxin from *Desulfovibrio vulgaris* (RdDv) as a test structure. RdDv crystallizes in space group $P2_1$, with $a = 19.993$, $b = 41.505$, $c = 24.404$ Å, $\beta = 107.6°$ and $Z = 2$ (Adman, Sieker, Jensen, Bruschi & LeGall, 1977). It is a small protein containing 52 amino-acid residues including an FeS_4 cluster, a total of 395 non-H atoms. The final model from the crystallographic study at 1.5 Å resolution contains, in addition, one sulfate group and 180 water molecules of varying occupancies, summing up to the equivalent of 98 fully occupied water positions (Adman, Sieker & Jensen, 1991). A more extensive

data set comprising 26237 reflections to 0.92 Å resolution has also been collected (Sheldrick *et al.*, 1993). Available crystals of RdDv had large mosaic spread (FWHM values from ω-scans of the best specimen were in the range 0.05–0.25°) and contained several individuals; therefore, attempts at PPE carried out on beamline X-3A2, NSLS, Brookhaven, were unsuccessful. In the absence of physically estimated values, starting phases were assumed known, and Cu $K\alpha$ data to 1.54 Å resolution ($2\theta_{max} = 60°$) was used. The structure of this protein has been re-determined by more conventional direct methods (Sheldrick *et al.*, 1993), and also by the combined reciprocal-space/real-space shake-and-bake method (Hauptman, 1995). However, both studies were based on data to 0.92 Å resolution.

Two program packages were employed for extension and refinement of the phases: *MULTAN*78, which was modified locally to (*a*) accept up to 3000 *E*'s (*b*) develop and store up to 100000 triplet phase relationships (TPR's), and (*c*) accept phase triplets with user-defined weights as input in the *SIGMA*2 routine; and *MULTAN*88 *E*, which employs the *SAYTAN* formalism of Woolfson (Debaerdemaeker, Tate & Woolfson, 1985, 1988). The particular version of this program that was used allows processing up to 100000 TPR's, ranked according to weight.

3. Expansion of the phase set

3.1. *Work with single phases*

The most common approach to PPE involves the study of three mutually interacting beams, **H**, $-$**L** and **L**$-$**H**, which may give information on the sum of the parent phases, $\Phi_3 = \varphi(\mathbf{H}) + \varphi(-\mathbf{L}) + \varphi(\mathbf{L} - \mathbf{H})$, a three-phase structure invariant (SI). However, if the secondary, $-L$, and the coupling, $L-H$, reflections are related by symmetry, the sum $\varphi(-\mathbf{L}) + \varphi(\mathbf{L} - \mathbf{H})$ is known, and the measurement then provides the single phase $\varphi(\mathbf{H})$ directly. The required symmetry is satisfied if a one-phase structure seminvariant (SS) is chosen as the primary phase $\varphi(\mathbf{H})$. For a one-phase SS all reflections L can be identified in any space group (Giacovazzo, 1980). In space group $P2_1$, the only class of one-phase SS are the reflections $2h\ 0\ 2l$. Setting $-L = \bar{h}k\bar{l}$ and $L - H = \bar{h}k\bar{l}$ gives $\varphi(-\mathbf{L}) = -\varphi(\mathbf{L} - \mathbf{H})$ for $k = 2n$, and $\varphi(-\mathbf{L}) = -\varphi(\mathbf{L} - \mathbf{H}) + \pi$ for $k = 2n + 1$. In both cases the experiment gives $\varphi(\mathbf{H})$ directly.

The first attempts at phase extension based on the use of $2h\,0\,2l$ as the primary reflections H were unsuccessful. This was not unexpected, as only about 15 $2h\,0\,2l$ reflections interacted strongly in many TPR's and contributed appreciably to the generation of new phases. As a more general explanation one recalls that $2h\,0\,2l$ represents one centrosymmetric projection, hardly sufficient as a basis for developing the complete complex structure. In space groups of higher symmetry with several classes of one-phase SS, phase expansion from a starting set of different one-phase SS is more likely to be useful.

Information on single phase values can be obtained from multiple isomorphous replacement and/or multiple anomalous dispersion. This provided further motivation to study the characteristics of starting sets comprising only single (known) phases. The next trials were based on starting sets including the most active $2h\,0\,2l$ reflections assigned the correct phase, 0 or π, and in addition general reflections from the bottom of the convergence map, either assigned one of the special values $m \cdot \pi/2$, or centred in the correct phase quadrant, *i.e.* with a mean phase error ±22.5°. Phase expansion was first attempted within a set of the 1000 largest *E*'s. The minimal starting set required to effect phasing of the full data set comprised 24 general reflections, 14 $2h\,0\,2l$ reflections, three reflections for fixing the origin, one for enantiomer discrimination and one weak link reflection to be permuted in steps of 90° over the phase plane. Thus, the total number of starting set phases was 43, of which 38 were assumed known. The procedure generated six distinct phase models. They were ranked according to the mean phase error $\langle \Delta\varphi \rangle$ defined as,

$$\langle \Delta\varphi \rangle = 1/n \sum_H \min \left\{ [|\varphi_g(\mathbf{H}) - \varphi_c(\mathbf{H})|], [2\pi - |\varphi_g(\mathbf{H}) - \varphi_c(\mathbf{H})|] \right\}, \quad (1)$$

where $\varphi_g(\mathbf{H})$ is the generated phase, $\varphi_c(\mathbf{H})$ is the phase calculated with the parameters from the crystallographic refinement, and n is the number of phases. The best model had $\langle \Delta\varphi \rangle = 55°$. The E map calculated for this model from the set of 1000 *E*'s showed one very dense maximum and weaker maxima attributable to bonded S atoms. The information content of this map was inferior to maps calculated from larger sets of E in the range 2500–3000, or 5–6 × number of non-H atoms of protein +98 water molecules.

As the next step we wanted to optimize and extend the generation of new phases, assuming that this would involve an intermediate step of phase expansion within a subset of the largest *E*'s. An optimum-sized subset should give a minimal $\langle \Delta\varphi \rangle$ after phase expansion and refinement from the start g set. The existence of an optimal size of this subset is anticipated, as an undersized set of *E*'s will involve fewer TPR's and, therefore, a less reliable estimate for each phase. The result of this study is summed up in Fig. 1. For sets up to about 300 *E*'s, $N_E \simeq 300$, the generation of new phases is unstable, as evidenced by the strong variation in $\langle \Delta\varphi \rangle$ for a small change in N_E. From $N_E \simeq 300$ the mean phase error decreases, goes through a slight minimum at $N_E \simeq 500$, and increases slowly again with increasing N_E. Although apparently not very critical, a

set with about 500 E's can be taken as an optimal subset within which new phases are generated and refined prior to further expansion within a larger set of E's. With 500 E's corresponding to an $E_{\min} = 1.54$ there were 5040 TPR's, *i.e.* a ratio 10:1 for $N_{\mathrm{TPR}}:N_E$, and $\langle\Delta\varphi\rangle = 51.3°$ for the best model. Further expansion among the 2500 largest E's ($E_{\min} = 0.875$) gave a final $\langle\Delta\varphi\rangle = 61.8°$. From a comparison of maps calculated from sets of E's of variable size, the information content appeared optimal for $N_E \simeq 2500$. Presumably, this number is determined by several parameters like resolution of the data, space-group symmetry, the minimum allowed TPR weight, $\kappa_{\min}$; the size of structure and whether or not it contains heavy atoms.

In the first part of this work, calculation of normalized structure factors was based on protein +90 water molecules, the latter corresponding approximately to the equivalent number of fully occupied solvent sites (Adman *et al.*, 1991). The number of atoms used for normalization will affect the weight κ for a TPR,

$$\kappa(\mathbf{H}, \mathbf{L}) = 2\sigma_3\sigma_2^{-3/2}|E(\mathbf{H})E(-\mathbf{L})E(\mathbf{L}-\mathbf{H})|, \quad (2)$$

which may in itself influence the development of a larger phase set. To gain some insight, useful at least in the present work, phase expansion and refinement were also performed with E's based on protein only. In this case a smaller starting set was adequate to accomplish a successful phase expansion. The starting set comprised two $2h\,0\,2l$, 24 general and one weak link reflection, in addition to the four reflections for defining origin and the enantiomer, a total of 31 reflections, of which 26 were assumed known in phase. In all later tests with single phases, calculation of E's was based on the protein only. Starting sets were first expanded and refined within a subset of the 500 largest E's using

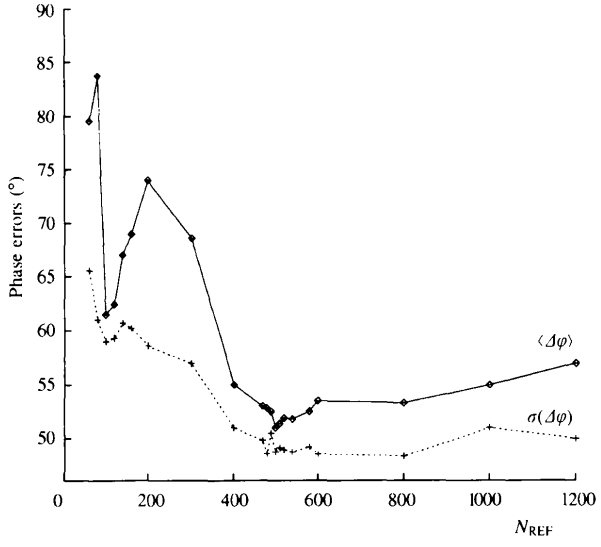

Fig. 1. Mean phase error $\langle\Delta\varphi\rangle$ with standard deviation $\sigma(\Delta\varphi)$ as a function of the number of phases in initial subset.

*MULTAN*78, the number of TPR's being about 5000. *MULTAN*88 E was used for further phase development within the 2500 largest E's, and in these runs $\kappa_{\min}$ was adjusted by the program to keep the number of TPR's < 100 000. The actual value of $\kappa_{\min}$ is critical also in the sense that it affects the weights of the phases and, therefore, also the number of rejected phases, φ_r. Increasing $\kappa_{\min}$ will reduce the number of TPR's and increase φ_r.

Up until this point $\langle\Delta\varphi\rangle$ had been used as a monitor to optimize the generation of new phases and to study the effect of alternative bases for calculating E's. It cannot be used as a figure of merit (FOM) when the target structure is unknown. FOM's calculated by *MULTAN*88 (Debaerdemaeker, Germain *et al.*, 1988) are ABSFOM, PSIZERO, RESID.

$$\mathrm{ABSFOM} = \frac{\sum_{\mathbf{H}} \alpha(\mathbf{H}) - \sum_{\mathbf{H}} \alpha(\mathbf{H})_{\mathrm{ran}}}{\sum_{\mathbf{H}} \alpha(\mathbf{H})_{\mathrm{est}} - \sum_{\mathbf{H}} \alpha(\mathbf{H})_{\mathrm{ran}}}, \quad (3)$$

where $\alpha(\mathbf{H}) = 2\sigma_3\sigma_2^{-3/2}E(\mathbf{H})|\sum_{\mathbf{L}} E(-\mathbf{L})E(\mathbf{L}-\mathbf{H})|$, $\alpha(\mathbf{H})_{\mathrm{ran}} = \langle\alpha(\mathbf{H})^2\rangle^{1/2} = \{\sum_{\mathbf{L}} \kappa^2(\mathbf{H}, \mathbf{L})\}^{1/2}$ and $\alpha(\mathbf{H})_{\mathrm{est}}$, the estimated $\alpha(\mathbf{H})$ with true phases is as defined by Germain, Main & Woolfson (1970). For a good set of phases ABSFOM will approach 1.0, values in the range 1.0 ± 0.2 are commonly obtained in small-structure work.

$$\mathrm{PSIZERO} = \frac{\sum_{\mathbf{H}} \left| \sum_{\mathbf{L}} E(-\mathbf{L})E(\mathbf{L}-\mathbf{H}) \right|}{\sum_{\mathbf{H}} \alpha(\mathbf{H})_{\mathrm{ran}}}. \quad (4)$$

This criterion, introduced by Cochran & Douglas (1957), is a test on weak $E(\mathbf{H})$ for which the main contribution in Sayre's equation comes from large terms $E(-\mathbf{L})E(\mathbf{L}-\mathbf{H})$. Thus, the inner summation is made over large terms, while the outer summation involves very small or zero $E(\mathbf{H})$'s. The numerator is sensitive to the phase values of the pairs $E(-\mathbf{L})E(\mathbf{L}-\mathbf{H})$. A correct phase set will give a low PSIZERO since $E(\mathbf{H})$ is small. Increasing phase errors will increase PSIZERO. Good phase sets for small structures commonly give PSIZERO in the range 1.0–1.5.

$$\mathrm{RESID} = \left[\frac{\sum_{\mathbf{H}} |\mathrm{sc}\,\alpha(\mathbf{H})_{\mathrm{est}} - \alpha(\mathbf{H})|}{\sum_{\mathbf{H}} \alpha(\mathbf{H})_{\mathrm{est}}} \right] \times 100. \quad (5)$$

RESID is a conventional FOM, it is smaller the closer the experimental α's follow the estimated values. In the presence of heavy atoms sc differs from unity, and is calculated as $\mathrm{sc} = \min\{1.3, [\max(\mathrm{ABSFOM}, 1.0)]^{1/2}\}$. For correct solutions in small-structure work RESID may be around 20.

All trials with single phases had produced six different phase models. These models were developed as described above. The final expansion involved

Table 1. *Distribution of FOM's, number of rejected phases φ_r in the set of 2500 E, and mean phase error $\langle\Delta\varphi\rangle$ for six models developed from single phases*

Model	ABSFOM	PSIZERO	RESID	φ_r	$\langle\Delta\varphi\rangle$
1	0.31	0.93	44.0	94	89.8
2	4.01	4.93	133.4	10	89.2
3	0.19	1.00	50.6	76	91.3
4	4.09	4.99	137.3	18	90.0
5	4.63	5.21	158.0	16	69.0
6	4.03	4.84	135.9	13	89.8

2700 E's. Phases were then refined among the 2500 best phase determined E's (with highest α_{cst}), and maps were calculated based on the latter group of E's. Table 1 gives the values of three of the FOM's, the number of rejected phases φ_r and $\langle\Delta\varphi\rangle$ for all models. According to the first four figures the six models are roughly divided in two groups: 1 and 3 have much lower FOM's and, most importantly, a much larger number of rejected phases (large φ_r) than the other four models, implying that the phase refinement was unsuccessful. Therefore, the former two were discarded. For the remaining models all FOM's are distinctly different from the values for small structures. According to the calculated $\langle\Delta\varphi\rangle$, 5 is the best choice, however, based on their FOM's alone it is not possible to rank models 2, 4, 5 and 6, and so all were examined further. It is a common observation in work with macromolecules that the standard FOM's are not well suited for identifying the best structure model(s), at least in the early stages of structure development (Woolfson & Yao, 1990). This problem will be discussed in more detail in the section on triplet phases. In the E maps only 5 and 6 had one very dense peak with several other peaks of less density in its vicinity. RdDv contains an Fe atom which should appear in a chemically sensible environment in the E map. Model 5 had four maxima of secondary density in an approximately tetrahedral arrangement about the densest peak. Two of these peaks were ascribed tentatively to S atoms, thus defining a molecular fragment of three atoms (1 Fe + 2 S). In the map of 6 the geometry of the peaks surrounding the strongest one was not in obvious agreement with bonding, and only the strongest peak was included as one Fe atom.

In *MULTAN*88 E a molecular fragment in known orientation and position can be incorporated by calculating a factor $q(\mathbf{H}, \mathbf{L})$ which will then modify the weight factor κ of the individual TPR's accordingly (Main, 1976). The result is to enhance those TPR's that receive the largest contribution from the fragment, and to sharpen and shift the estimate of new phases. The recycling process was carried out as follows.

(*a*) Calculate κ_{new} from new fragment for all TPR's among the set of 2700 E's.

(*b*) Use original phases for subset of 500 largest E's to retain origin and enantiomorph. Develop and refine phases among the 2500 best phase-determined E's using *SAYTAN*. Allow also a group of 500 phases to refine during last six cycles.

(*c*) Calculate a new E map from 2500 E's and search for larger fragment.

(*d*) Return to (*a*).

Table 2. *Work with single phases, model 5*

Development of FOM's, number of rejected phases φ_r, number of atoms included from map prior to this cycle, and mean single phase error $\langle\Delta\varphi\rangle$ after this cycle.

Cycle	ABSFOM	PSIZERO	RESID	φ_r	No. of atoms	$\langle\Delta\varphi\rangle$
0	4.630	5.214	158.02	16	0	69.0
1	4.972	5.023	173.25	16	3	59.8
2	4.542	4.413	159.72	3	5	54.4
3	4.106	4.172	140.54	3	11	52.1
4	3.841	4.112	128.38	5	15	50.5
5	3.619	3.938	118.59	4	18	49.9
6	3.535	3.824	114.79	3	20	49.1
9	2.228	2.877	56.84	3	50	42.4
12	1.398	2.139	34.00	2	82	38.6
15	0.879	1.767	30.70	3	150	30.7
18	0.684	1.557	32.85	1	211	26.4

After cycle 1 the two most dense maxima in the map for 5 confirmed the tetrahedral arrangement that was observed after cycle 0, and the two peaks were included as two additional S atoms. The map for 6 showed several subsidiary peaks, but the geometry was not in agreement with expected bonding features. This model was, therefore, not pursued further, and work was concentrated on 5. After cycle 1 two of the three FOM's that normally should decrease for a correct partial structure, had increased. However, from cycle 2 onwards (FeS_4 cluster complete) all FOM's decreased with increasing structure fragment, see Table 2. Progress was slow during the next five to six cycles, mainly because a very cautious strategy was adopted: only maxima satisfying connectivity and rather conservative bonding criteria were accepted as new atomic sites, and they were not refined.

A turning point was reached when a fragment consisting of about 25 atoms, or 6% of the protein, had been localized. From then on identification of new atomic positions in the E maps was easier, all FOM's indicated improvement, but RESID showed a minimum for a fragment of about 150 atoms and then increased slightly. The FOM's may also be influenced by the choice of κ_{min} and the resulting φ_r. It is probable that refinement of atomic positional and displacement parameters would have accelerated the process, but this was not explored. The mean phase error $\langle\Delta\varphi\rangle$ decreased throughout the recycling process, and was 26.4° after cycle 18 when 211 atoms had been included in the molecular fragment, ABSFOM = 0.68, PSIZERO = 1.56, and RESID = 32.85. At this point the backbone atoms of 41 of the 52 residues had been identified, the r.m.s. deviation from the refined positions for all 211 atoms $\langle(\Delta r)^2\rangle^{1/2} = 0.18$ Å. It was

concluded that a correct solution of the phase problem had been established beyond doubt, and further work to obtain the complete structure was deemed unnecessary. A summary of the development during recycling of model 5 is given in Table 2.

Despite the high FOM values in the early stages it is clear that the parent E maps contain significant information. Fig. 2(*a*) shows a section of the first E map for model 5 near the Fe position before any atoms have been included. Fig. 2(*b*) shows the same section after cycle 18 with 211 atoms in the molecular fragment. The latter map gives positions of atoms identified prior to cycle 18 in the range 0–0.375 Å from this layer. In addition to the peaks corresponding to Fe and S6, there are many similar features in the two maps, for instance a band of maxima at about constant x

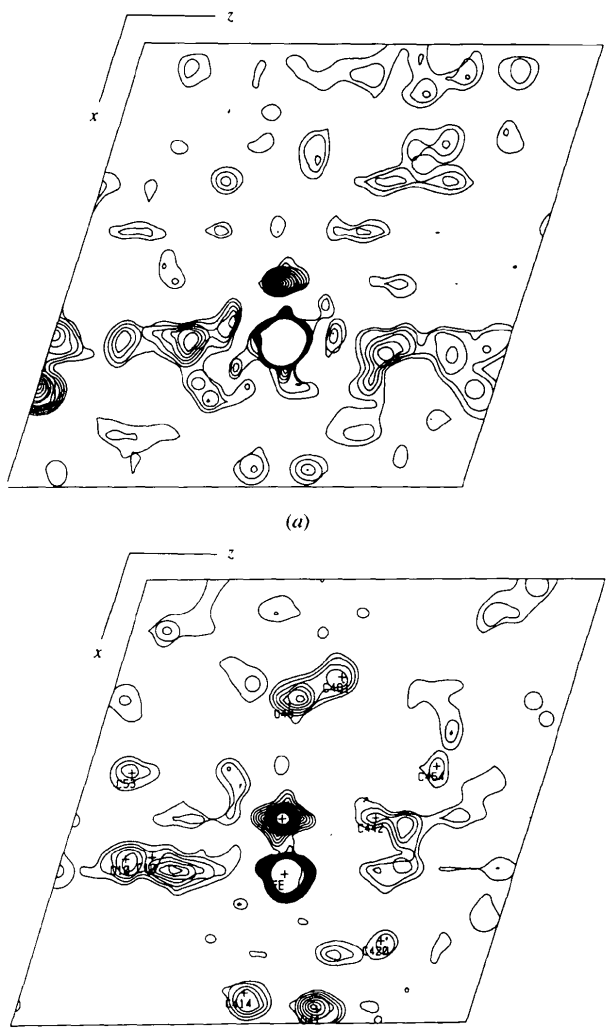

Fig. 2. (*a*) First map based on 2500 E, developed from 26 known single phases, no atoms included. (*b*) Map from 2500 E after cycle 18, based on 211 atoms in the fragment.

including the Fe peak. Six of the ten protein C or O atoms fall in maxima or regions of increased density in the first E map. The high content of structure information in the E maps suggests that a more radical approach could have been followed in building up the molecular fragment.

3.2. *Work with triplet phases*

In relation to PPE a more realistic situation involves the use of triplet phases Φ_3. The primary diffracted intensity in the case of three interacting beams can be written as, $I(\mathbf{H}) \propto k|F(\mathbf{H})|^2[y_B]$, where y_B is the dynamical correction to the kinematically diffracted intensity. Information on the triplet phase is contained in the correction y_B which can be given a simple analytical form in the centrosymmetric case following a set of first order approximations that have been described previously (Mo, Hauback & Thorkildsen, 1988),

$$y_B = [1 - 2Q^{1/2}PR_F \cos\Phi_3(1/s_L) + QP^2R_F^2(1/s_L)^2]. \quad (6)$$

Q contains both universal and experimental constants, P describes polarization for the three-beam case in question within the frame of the approximations, and s_L = signed distance of the secondary reciprocal lattice node L from the Ewald sphere. For centrosymmetry the phase information resides in the backgrounds of the intensity profile, away from $s_L = 0$. For $\Phi_3 \simeq +\pi/2$ or $-\pi/2$ the diffraction power is symmetric in s_L, and the phase information lies in the relative heights of the profile extrema at $s_L = 0$. Expressions for general phase values have been developed by Hümmer & Billy (1986) from higher order approximations to dynamical plane-wave theory, and by Thorkildsen (1987) from spherical wave theory. Of particular interest for the present analysis is the parameter $R_F = |F(-\mathbf{L})|\,|F(\mathbf{L}-\mathbf{H})|/|F(\mathbf{H})|$. A large R_F will enhance the phase signal and in general make it more amenable to measurement. Reflection triplets were accepted based on the following criteria.

(*a*) They should appear at the bottom of the convergence map of *MULTAN* (Germain, Main & Woolfson, 1971).

(*b*) $\{|F(-\mathbf{L})|, |F(\mathbf{L}-\mathbf{H})\}$ among top p% of $|F|$ at a resolution of 1.54 Å.

(*c*) $R_F \geq R_{F\,\mathrm{lim}}$.

General triplet phases were introduced with a mean phase error ±22.5°. Normalized structure factors were calculated based on two different models: protein +90 water molecules and protein alone, and were tested in separate runs. As in the study with single phases, the best results were obtained with E's based on protein alone. In this case, and with $p = 7\%$ and $R_{F\,\mathrm{lim}} = 300$, 45 triplet phases were required to obtain complete phase development within the subset

comprising the 500 largest E's. The acceptance criteria implied five weak links in the convergence map, leading finally to a total 40 different phase models (run I). With E's based on protein +90 H_2O and $p = 4\%$, 52 triplet phases were needed, and with seven weak links a total 256 models had to be examined (run II). All models were first developed and refined within the subset of 500 largest E's using the modified version of *MULTAN*78. Subsequent expansion among 2700 E's, and refinement involving the 2500 best determined E's of this set was carried out with *MULTAN*88 E.

3.3. *Use of FOM's as reject criteria*

FOM's for the 40 phase sets from run I were in the following ranges: ABSFOM 3.27–3.86, PSIZERO 4.64–5.63, and RESID 87.0–106.8. Previous experience had suggested that a better approach could be to use the standard FOM's as criteria for rejecting the worst phase models. In the study with single phases in the starting set it was found that the best models, as judged by φ_r and $\langle\Delta\varphi\rangle$, had RESID > 100. Similar observations have been reported by Mukherjee & Woolfson (1993). Of the 40 models, 27 had RESID < 100, but also generally low values for the other two FOM's. They were rejected, leaving for further study 13 models with narrower ranges for ABSFOM and PSIZERO, 3.77–3.86, and 5.09–5.63, respectively. As a check on this criterion for rejection, and perhaps to enable further reduction of the number of phase models, the mean triplet phase error $\langle\Delta\Phi_3\rangle$ was calculated for all 40 models and their enantiomorphs, a total 80 models, relative to the refined rubredoxin structure. In each case about 53 000 triplets were used for the calculations. Triplet phases are independent of the choice of origin and $\langle\Delta\Phi_3\rangle$ should allow a reliable identification of the best models. Work with single phases had shown that a model with mean phase error $\langle\Delta\varphi\rangle = 69.0°$ for 2500 E's could be successfully refined, and recycling initiated on a small three-atom fragment led in 17 cycles to a partial structure comprising 211 non-H protein atoms. In the run started from triplet phases 78 of the 80 models had $\langle\Delta\Phi_3\rangle$ in the range 81.5–83.1°, and only two models, 32 and its enantiomer, had $\langle\Delta\Phi_3\rangle < 80°$. Assuming that 32 or its enantiomer were promising models to be retained for further study, both the product ABSFOM × RESID and PSIZERO appeared to be useful reject criteria. All models with PSIZERO normalized = PSIN > 1.00 were rejected, leaving nine models. The E maps for these models were searched (*PEKPIK*) for features relating to the expected topology of the FeS_4 cluster. Only the maps for models 12, 29 and 32 showed one strong peak with four weaker ones within reasonable distance, and work was concentrated on these three models. A five-atom fragment was fitted to the pattern of peaks in each of the maps, and recycled as described above. $\langle\Delta\Phi_3\rangle$ for 32 was thereby reduced to 76.6° from its initial value 79.2°, for the other two models $\langle\Delta\Phi_3\rangle$ increased slightly to 82.6° (12) and 81.9° (29). The new map for the 32 enantiomer showed several maxima in chemically acceptable positions, and seven new atoms could be added to the five-atom fragment. Maxima in the map for 12 did not confirm the presence of an FeS_4 cluster, or provide evidence of other structure elements, and in the next cycle a reduced fragment Fe + 2S was used. Maxima in the map for 29 had shifted, and as no structure fragment was obvious, this model was abandoned. Recycling was continued for both models 12 and 32. After three more cycles it was clear that 12 did not progress towards a solution, and only the 32 enantiomer was processed further. After nine refinement cycles a fragment comprising 100 atoms had been identified. The work with single phases had shown that a fragment of RdDv this size is far more than required to ensure further expansion, and recycling was discontinued at this point. In the recycling process all FOM's decreased towards normal values, and $\langle\Delta\Phi_3\rangle$ was reduced to 56.0°. By applying an appropriate origin shift to the coordinates of the model, also the development in the mean single phase error could be followed. After cycle 10, $\langle\Delta\varphi\rangle$ was 36.0° compared with 64.0° at the beginning. The development for a selection of parameters during recycling of enantiomer model 32 is shown in Table 3.

Table 3. *Work with triplet phases, model 32 enantiomer*

Development of FOM's, number of rejected phases φ_r, number of atoms included from map prior to this cycle, mean single phase error $\langle\Delta\varphi\rangle$ and mean triplet phase error $\langle\Delta\Phi_3\rangle$ after this cycle.

Cycle	ABSFOM	PSIZERO	RESID	φ_r	No. of atoms	$\langle\Delta\varphi\rangle$	$\langle\Delta\Phi_3\rangle$
0	3.804	5.178	104.35	168	0	64.0	79.2
1	4.086	4.522	129.06	36	5	57.6	76.6
2	3.599	4.061	110.65	20	12	51.7	73.8
3	3.198	3.720	93.43	28	19	49.3	71.3
4	3.021	3.610	85.50	22	22	48.9	71.4
5	2.837	3.467	78.09	25	26	48.6	70.4
6	2.321	3.158	52.75	81	34	46.0	69.5
7	1.847	2.703	32.82	165	47	42.8	66.1
8	1.437	2.510	24.65	422	57	39.7	61.6
9	1.200	2.229	27.55	279	72	38.8	60.2
10	1.020	1.768	30.26	140	100	36.0	56.0

The 256 models from run II represented a more demanding test of the usefulness of our empirical reject criteria. Of these models, 137 had RESID < 100 and were rejected. Ranges in ABSFOM, PSIZERO and RESID for the remaining 119 models were: 4.30–4.99, 4.80–5.53 and 100.0–119.6, respectively. Another 69 models with PSIN > 1.00 were rejected, and E maps for the remaining 50 models were calculated and searched for maxima corresponding to a complete or partial FeS_4 cluster. Four maps showed one strong plus four weaker

maxima, 14 others had only three weaker maxima in addition. In the four most promising models, 29, 69, 142 and 207, a start fragment of Fe + 4S could be tentatively identified. The four models were examined by recycling and phase refinement in the same manner as described already. Only models 29 and 207 were found to yield a promising structure fragment, 207 corresponds to the enantiomer of 29 with a shift of origin. Model 29 was expanded to include 54 atoms before recycling was discontinued. At this point $\langle\Delta\varphi\rangle$ had been reduced from 69.4 to 56.4°, and $\langle\Delta\Phi_3\rangle$ from 80.7 to 75.9°.

3.4. *Faulty FOM's*

Although standard FOM's were used successfully in this work to reject the majority of the poor phase models the situation is not satisfactory. The three indicators ABSFOM, PSIZERO and RESID have anomalously high initial values. They all have either $\alpha(\mathbf{H})_{\text{est}}$, or $\alpha(\mathbf{H})_{\text{ran}}$ or both parameters in the denominator. For a given set of $\kappa(\mathbf{H}, \mathbf{L})$ it was shown by Germain *et al.* (1970) that an estimate of $\alpha(\mathbf{H})$ can be written as,

$$\alpha^2(\mathbf{H})_{\text{est}} = \sum_i \kappa^2(\mathbf{H}, \mathbf{L}_i) + \sum_{i_1}\sum_{\neq i_2} \kappa(\mathbf{H}, \mathbf{L}_{i_1})\,\kappa(\mathbf{H}, \mathbf{L}_{i_2}) \times \frac{I_1[\kappa(\mathbf{H}, \mathbf{L}_{i_1})]}{I_0[\kappa(\mathbf{H}, \mathbf{L}_{i_1})]}\,\frac{I_1[\kappa(\mathbf{H}, \mathbf{L}_{i_2})]}{I_0[\kappa(\mathbf{H}, \mathbf{L}_{i_2})]}, \quad (7)$$

where $I_0[\kappa(\mathbf{H}, \mathbf{L}_i)]$ and $I_1[\kappa(\mathbf{H}, \mathbf{L}_i)]$, are modified Bessel functions.

For a large structure, in particular in the early stages when there are relatively few TPR's with large variances, both sums in (7) tend to be very small as each individual $\kappa(\mathbf{H}, \mathbf{L}_i)$ is a small quantity. Also $\alpha(\mathbf{H})_{\text{ran}}$ is initially very small as can be seen from the definition following (3). This explains directly the large values of PSIZERO and RESID, in ABSFOM the denominator is the difference between two initially small quantities. As an increasing fraction of the structure is incorporated, the weight factor κ for each individual TPR is modified, and the FOM's are expected to approach more normal values.

The three FOM's apparently develop differently under recycling. In the present work, ABSFOM and RESID increased in the first refinement cycle to reach a maximum, and then decreased as expected with further expansion of the structure fragment, *cf.* Tables 2 and 3. PSIZERO decreased from the beginning. The difference in behaviour may be a result of the particular strategy used for increasing the structure fragment. In this work a very cautious procedure was employed, thus in run I, the fragment of model 32 was increased by no more than three to seven atoms in each of the first four cycles. It is probable that if more atoms had been incorporated in the fragment during the first cycles, all FOM's would have decreased from the beginning. In any case it can be stated that the common FOM's have anomalously high values initially, but they can be used as rough reject criteria in the early critical stages of recycling and phase refinement.

4. Conclusions

In the present work we have shown that a small protein like rubredoxin could be solved with data at a resolution of 1.5 Å, starting from a set of 26 single phases, alternatively from a set of 45 triplet phases, in both cases known with an averge error ±22.5° which can be obtained in three-beam diffraction experiments. From recent *ab initio* studies of small proteins it was observed that a practical lower limit in resolution for successful structure solution is about 1.2 Å at present (Sheldrick *et al.*, 1993; Weeks *et al.*, 1995). A small set of physically estimated triplet phases contains significant information that would be helpful in direct methods with data at any resolution, and seems to be essential at lower resolution. It is very likely that 1.5 Å is not a lower limit, but the effect of reducing the resolution further was not studied in this work. An interesting result is that phase sets including 2500 E with a mean phase error $\langle\Delta\varphi\rangle \simeq 70°$ or a mean triplet phase error $\langle\Delta\Phi_3\rangle \simeq 80°$, both relative to the model from the crystallographic refinement, can be refined and expanded successfully in a Fourier recycling process. From the beginning of recycling the E maps contain much structure information and a more radical strategy may be advantageous in building up a molecular fragment. The presence of an FeS_4 cluster in this structure was used in the interpretation of the initial maps, and may also have been of importance for the fact that quite small sets of known phases were sufficient for solving the phase problem.

Work is in progress to explore the potential of this method for small proteins with no heavy atoms using data at about 1.5 Å resolution.

We are grateful to Professor M. M. Woolfson and Dr C. Tate for making the program *MULTAN*88 *E* available to us and for several valuable discussions on technical aspects. Norges Forskningsråd is thanked for Grants 424.93/021 and 101166/432 in support of this work.

References

Adman, E. T., Sieker, L. C. & Jensen, L. H. (1991). *J. Mol. Biol.* **217**, 337–352.
Adman, E. T., Sieker, L. C., Jensen, L. H., Bruschi, M. & LeGall, J. (1977). *J. Mol. Biol.* **112**, 113–120.
Agarwal, R. C. & Isaacs, N. W. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2835–2839.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Chang, S.-L. (1982). *Phys. Rev. Lett.* **48**, 163–166.

Chang, S.-L., King, H. E. Jr, Huang, M.-T. & Gao, Y. (1991). *Phys. Rev. Lett.* **67**, 3113–3116.
Cochran, W. & Douglas, A. S. (1957). *Proc. R. Soc. London Ser. A*, **243**, 281–288.
Debaerdemaeker, T., Germain, G. Main, P., Refaat, L. S., Tate, C. & Woolfson, M. M. (1988). *MULTAN*88. *Computer programs for the automatic solution of crystal structures from X-ray diffraction data*. University of York, York, England.
Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1985). *Acta Cryst.* A**41**, 286–290.
Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1988). *Acta Cryst.* A**44**, 353–357.
Germain, G., Main, P. & Woolfson, M. M. (1970). *Acta Cryst.* B**26**, 274–285.
Germain, G., Main, P. & Woolfson, M. M. (1971). *Acta Cryst.* A**27**, 368–376.
Giacovazzo, C. (1980). *Direct Methods in Crystallography*. London: Academic Press.
Gong, P. P. & Post, B. (1983). *Acta Cryst.* A**39**, 719–724.
Hauptman, H. A. (1995) *Acta Cryst.* B**51**, 416–422.
Hümmer, K. & Billy, H. (1986). *Acta Cryst.* A**42**, 127–133.
Hümmer, K., Schwegle, W. & Weckert, E. (1991). *Acta Cryst.* A**47**, 60–62.
Main, P. (1976). *Crystallographic Computing Techniques,* edited by F. R. Ahmed, pp. 97–105, Copenhagen: Munksgaard.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Mo. F., Hauback, B. C. & Thorkildsen, G. (1988). *Acta Chem. Scand. A*, **42**, 130–138.
Mukherjee, M. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 9–12.
Post, B. (1977). *Phys. Rev. Lett.* **39**, 760–763.
Rango, C., de, Mauguen, Y., Tsoucaris, G., Dodson, E. J., Dodson, G. G. & Taylor, D. J. (1985). *Acta Cryst.* A**41**, 3–17.
Sayre, D. (1972). *Acta Cryst.* A**28**, 210–212.
Sayre, D. (1974). *Acta Cryst.* A**30**, 180–184.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Sjölin, L. & Svensson, L. A. (1993). *Acta Cryst.* D**49**, 66–74.
Thorkildsen, G. (1987). *Acta Cryst.* A**43**, 361–369.
Thorkildsen, G. & Mo, F. (1982). *Abstr. of the 7th Eur. Crystallogr. Meet.* (ECM-7), Jerusalem, 1982, p. 6.
Thorkildsen, G. & Mo, F. (1983). *Abstr. of the 8th Eur. Crystallogr. Meet.* (ECM-8), Liege, 1983, p. 258.
Tsoucaris, G. (1970). *Acta Cryst.* A**26**, 492–499.
Weckert, E. (1996). *Acta Cryst.* Submitted.
Weckert, E., Schwegle, W. & Hümmer, K. (1993). *Proc. R. Soc. London Ser. A*, **442**, 33–46.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.
Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.