Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Ab initio structure determination of a small protein, rubredoxin, by direct methods

Monika Mukherjee

Department of Solid State Physics, Indian Association for the Cultivation of Science, Jadavpur, Calcutta-700032, India

Correspondence e-mail: sspmm@iacs.ernet.in

Received 28 May 1998
Accepted 3 December 1998

The direct-methods program *SAYTAN* has been applied successfully to a known protein, rubredoxin, which contains 52 amino-acid residues including an FeS_4 unit, a sulfate ion and 102 solvent water molecules. Starting with initially random phases, useful sets can be obtained from multiple trials and selected by figures of merit at different resolutions. Phase extension followed by weighted Fourier recycling reveals a recognizable structure of rubredoxin. The model is refined against 1 Å resolution data to an R factor of 14.5% using the program *SHELXL*93.

1. Introduction

Direct methods, in the form of easily used computer packages, have been successful in solving the great majority of small molecular structures in a relatively straightforward manner. But the *ab initio* phasing of proteins, *i.e.* constructing the three-dimensional structure solely from a single set of X-ray diffraction data by direct methods, still poses a lofty challenge. The problem associated with the application of direct methods to proteins is the inevitable weakening of the probability distributions of individual phase relations owing to the large number of atoms in the unit cell, and the situation is further complicated by the lack of atomic resolution data for most of the proteins. Recent developments in the existing methodologies with the Sayre tangent formula as incorporated in *SAYTAN* (Debaerdemaeker, Tate *et al.*, 1988), the minimal principle (Hauptman, 1991) as implemented in *Shake-and-Bake* (Weeks *et al.*, 1994) and the improved protein-data collection strategies employing synchrotron radiation (Helliwell *et al.*, 1993) have shown considerable promise for the application of direct methods in protein crystallography.

The Sayre tangent formula *SAYTAN* (Debaerdemaeker, Tate *et al.*, 1988), as embodied in the program package *MULTAN*88 (Debaerdemaeker, Germain *et al.*, 1988), can be expressed as

$$\varphi(\mathbf{l}) = \text{phase of } [t(\mathbf{l}) - 2Kq(\mathbf{l})], \tag{1}$$

where K is a scaling factor and $t(\mathbf{l})$ and $q(\mathbf{l})$ correspond to the contributions from phase triplets and quartets, respectively. A set of phases satisfying equation (1) minimizes the least-squares residual for the system of Sayre equations (Sayre, 1952). The minimization condition is

$$\delta R/\delta\varphi(\mathbf{l}) = 0 \text{ for all } \mathbf{l}, \tag{2}$$

where

$$R = \sum_h |E(\mathbf{h}) - [K/g(\mathbf{h})] \sum_k E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|^2. \tag{3}$$

Table 1
Summary of the results of applying *SAYTAN* and *PSM* to rubredoxin at different resolutions.

N_{rel} is the number of three-phase relationships, N_G is the number of good sets generated with MPE < 71° and LMPE is the lowest MPE for the N_G good sets. In each case 1000 trials were made.

		Refined by *SAYTAN*		Refined by *PSM* and *SAYTAN*	
Resolution (Å)	N_{rel}	N_G	LMPE	N_G	LMPE
1.0	8412	30	41.1	32	41.0
1.25	13495	35	47.3	39	49.2
1.5	27335	39	47.1	43	49.0
1.77	21602	27	59.7	20	58.7
2.0	26439	6	67.8	8	64.8
2.25	18879	1	71.1	11	70.8

$E(\mathbf{h})$, $E(\mathbf{k})$ and $E(\mathbf{h}-\mathbf{k})$ are the normalized structure factors and $g(\mathbf{h})$ can be determined on theoretical grounds. Woolfson & Yao (1990) first demonstrated the potential of the direct-methods program *SAYTAN* in solving aPP (avian pancreatic polypeptide), a small protein (36 amino-acid residues plus 80 waters and a Zn atom) with a previously known structure (Glover *et al.*, 1983). Other successes in the direct-methods phasing of proteins include redetermination of aPP from truncated 3 Å resolution data (Mukherjee & Woolfson, 1993), 2-Zn insulin (Mukherjee & Woolfson, 1995), crambin (Weeks *et al.*, 1995), toxin II (Smith *et al.*, 1997) and the previously unknown structure E*r*-1 (Anderson *et al.*, 1996).

The present paper reports the *ab initio* solution of rubredoxin (RXN) derived from the bacterium *Desulfovibrio vulgaris* and comprising 393 protein atoms including an FeS_4 unit, a sulfate ion and 102 water molecules in the asymmetric unit using the program *SAYTAN* followed by least-squares refinement of the model. Structure redetermination at different resolutions was undertaken in order to study the effect of data resolution on the direct-methods protein phasing and the possibility of recognizing good phase sets from the modified figures of merit. The data for rubredoxin used in the present analysis, consisting of $|F|$ and $\sigma(|F|)$ for 18532 unique reflections at 1 Å resolution, have been obtained from the Protein Data Bank (PDB codes: R8RXNSF, 8RXN). The experimental details of data collection using synchrotron radiation and X-rays from a sealed tube and the structure of rubredoxin (unit-cell parameters $a = 19.97$, $b = 41.45$, $c = 24.41$ Å and $\beta = 108.30°$; space group $P2_1$) at 1 Å resolution have been reported in the literature (Dauter *et al.*, 1992). The known structure was not, however, used in the current phasing and the structure-refinement processes except for the mean phase error (MPE) calculation.

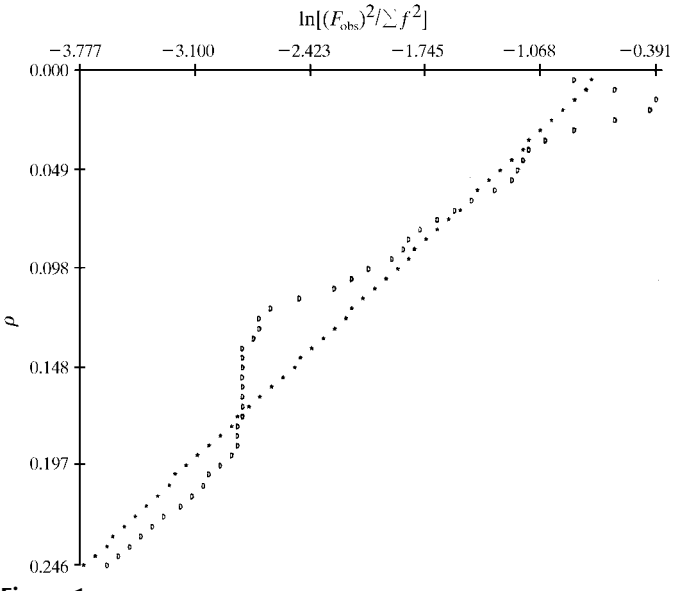

Figure 1
Wilson plot for rubredoxin at 1 Å resolution.

2. *Ab initio* phase determination

Normalized structure-factor magnitudes $|E|$ for RXN were obtained *via* a Wilson analysis, with an estimate of the overall temperature factor of 6.2 Å^2. The deviations of experimental points from linearity in the low-resolution regions of the Wilson plot (Wilson, 1942), based on 18532 reflections at 1 Å resolution (Fig. 1), were insignificant. The 800 largest and 100 smallest *E*s selected from the 1 Å resolution data were input into *SAYTAN*. Triplet relationships linking the largest *E*s and having $\kappa > 0.2$ were considered where

$$\kappa(\mathbf{h},\mathbf{k}) = 2\sigma_3\sigma_2^{-3/2}|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \quad (4)$$

and

$$\sigma_n = \sum_{j=1}^{n} Z_j^n. \quad (5)$$

The contributions of the small quartet terms only were included throughout the process. Starting with the random phases generated by a magic integer series (White & Woolfson, 1975), 1000 trials were made at various resolutions obtained by suitably truncating the data. Keeping the phases of the 50 largest *E*s fixed until the last cycle of refinement, when they were relaxed to fit in with the other values, phase refinements were carried out using *SAYTAN* alone or five cycles of the parameter-shift method (*PSM*; Debaerdemaeker & Woolfson, 1983) followed by *SAYTAN* in the normal way. The tangent-formula weighting scheme (Hull & Irwin, 1978) made the phase refinement quite stable. The use of *PSM* prior to *SAYTAN* enabled any initial phase to jump out from a local minimum and seemed to offer some advantage at lower resolutions (Table 1). In the application, the phases are changed one at a time by ±45° and tested against minimization of equation (3). At each step, the phase giving the lowest value of R corresponding to the shift of +45, 0, −45° was accepted. It should be noted that the *Shake-and-Bake* (Weeks *et al.*, 1994) phase-determination procedure invokes a similar parameter-shift routine for phase refinement using the minimal principle.

The results of *ab initio* phase refinement of rubredoxin at different resolutions are summarized in Table 1. A recent experiment on the application of direct methods to ribonuclease, a protein without any heavy atoms, has shown that

phase extension and refinement based on a set with an initial mean phase error (MPE) of 70° can lead to a map for automated development of the protein model (Mukherjee *et al.*, 1999). The quality of the maps could, however, be improved by various methods (Wang, 1985; Zhang & Main, 1990*a*,*b*). The phase sets with MPE < 71° are designated good sets (N_G) in Table 1. To account for the possible shift of origin in space group $P2_1$, the MPE for each set was calculated with respect to 400 equally spaced points along the *b* axis and the lowest value of MPE was considered. The lowest mean phase errors (LMPE) in the resolution range 1.0–1.5 Å varied from 41 to 47° (Table 1) with the number of N_G being 30–40; the corresponding MPE value reported by Sheldrick *et al.* (1993) at 1.1 Å resolution was 59°.

Since a prior knowledge of the protein structure is necessary for MPE calculation, the key factor for a successful application of direct methods to unknown proteins is the identification of the correct solution from suitable figures of merit (FOM; Gilmore *et al.*, 1991; Mishnev & Woolfson, 1994). Though the conventional *MULTAN* FOMs are not useful for macromolecular structures, the modified figures of merit (Mukherjee & Woolfson, 1993, 1995) seem capable of discriminating better phase sets from poorer ones in rubredoxin. The FOMs and the MPEs corresponding to both enantiomorphs (MPE1, MPE2) for selected phase sets at 1.0 Å and 1.5 Å resolutions are given in Tables 2(*a*) and 2(*b*), respectively. Representative plots of various FOMs against MPE, based on the top 60 sets at 1 Å resolution arranged in order of MPE, are shown in Figs. 2(*a*), 2(*b*) and 2(*c*). An examination of the results indicate that good phase sets can be recognized by the large values of ABSM, PSIM and especially CFOM2.

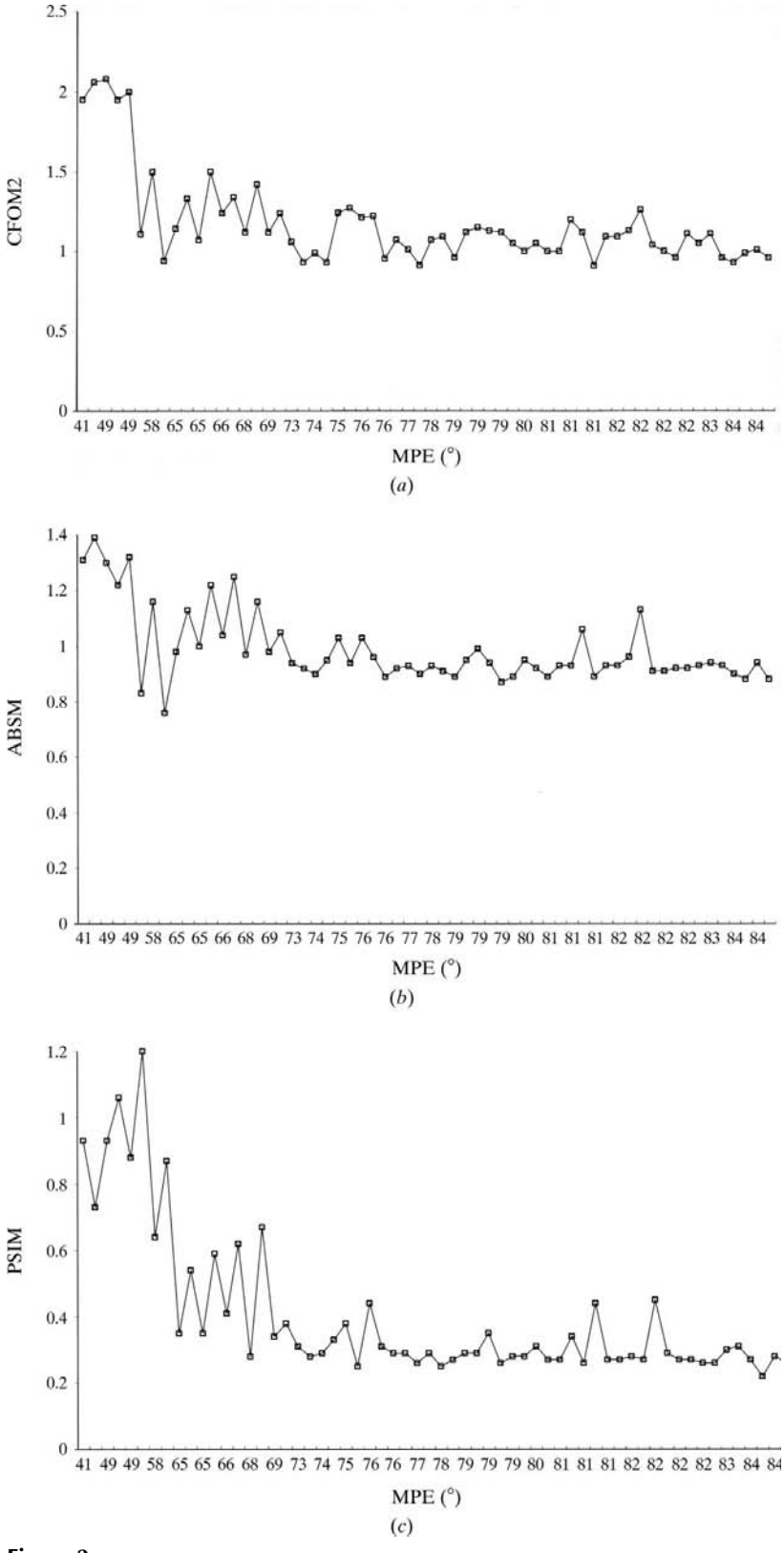

Figure 2
(*a*) Plot of figure of merit CFOM2 against mean phase error (MPE). (*b*) Plot of figure of merit ABSM against MPE. (*c*) Plot of figure of merit PSIM against MPE.

3. Phase extension

Starting with *m* known phases ranked in order of associated magnitudes of tangent-formula indications (α) and other ($M - m$) reflections with large *E* values, phase extension and refinement were carried out following the general procedure of Woolfson & Yao (1988). 700 *ab initio* phases corresponding to set number 338 in Table 2(*a*) (map correlation coefficient 0.47), obtained from a previous run of *SAYTAN* with 1.0 Å data, were extended to 2000 reflections. During the process of refinement, phases of the initial 700 reflections were kept fixed with unit weight. Trials with various values of *m* (300–700), *M* (1000–3000) and different initial random phases for ($M - m$) reflections had a marginal effect on the phase extension and refinement processes. An *E* map was calculated with the phases of 2000 reflections and interpreted by the *Peak-Search* routine of *SAYTAN*. Using the largest 100 peaks as input,

Table 2
A selection of FOMs for rubredoxin.

(*a*) At 1 Å resolution

Set	ABSM†	PSIM‡	RESM§	CFOM2¶	MPE1†† (°)	MPE2†† (°)
62	1.16	0.64	30.91	1.50	58.27	59.29
77	0.94	0.31	27.64	1.06	73.37	80.11
100	0.88	0.25	26.20	1.03	84.77	83.55
139	1.22	1.06	31.64	1.95	48.56	50.69
312	1.22	0.59	31.32	1.50	64.54	71.02
338	1.30	0.93	31.68	1.95	41.41	52.81
395	0.90	0.27	27.54	0.97	82.73	83.59
422	1.30	0.93	30.06	2.08	49.14	50.66
560	0.93	0.31	26.47	1.15	80.45	79.25
725	0.91	0.28	27.01	1.04	81.86	83.42
738	1.39	0.73	29.57	2.06	45.99	53.18
806	1.32	0.88	30.72	2.00	48.54	51.65

(*b*) At 1.5 Å resolution.

Set	ABSM†	PSIM‡	RESM§	CFOM2¶	MPE1†† (°)	MPE2†† (°)
143	1.72	0.96	41.43	2.31	53.64	52.47
205	1.62	0.98	44.01	1.75	57.99	55.18
241	1.45	0.27	43.46	0.76	85.22	86.68
264	1.92	0.85	38.99	2.78	47.66	55.80
304	1.63	0.99	41.63	2.25	55.56	57.56
500	1.69	0.71	43.28	1.57	84.39	82.72
545	0.61	0.65	38.50	1.66	78.81	81.69
565	1.74	0.98	40.93	2.43	51.89	54.36
838	1.95	0.86	39.47	2.72	47.10	54.53
862	1.90	0.80	38.90	2.70	52.36	53.39
970	0.54	0.76	39.53	1.57	83.93	85.10

† ABSM = $s/s_{\rm exp}$. ‡ PSIM = $\sum_{\mathbf{l}} |\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|/s$. § RESM = $100 \times \sum_{\alpha} |\{[\alpha(\mathbf{h})]/s\} - \{[\alpha(\mathbf{h}_{\rm est})]/s_{\rm est}\}$, where $\alpha(\mathbf{h}) = |\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|$, $s = \sum_{\mathbf{h}} \alpha(\mathbf{h})$, the subscript exp corresponds to the value for the true phases and the summations are over **h** for large *E*s and **l** for small *E*s. ¶ CFOM2 = $w_1[(\text{ABSM} - \text{ABSM}_{\rm min})/(\text{ABSM}_{\rm max} - \text{ABSM}_{\rm min})] + w_2[(\text{PSIM} - \text{PSIM}_{\rm min})/(\text{PSIM}_{\rm max} - \text{PSIM}_{\rm min})] + w_3[(\text{RESM} - \text{RESM}_{\rm min})/(\text{RESM}_{\rm max} - \text{RESM}_{\rm min})]$, where the subscripts max and min correspond to the maximum and minimum values for the 1000 phase sets and the weights are set at $w_1 = w_2 = w_3 = 1.0$. †† MPE1 is the mean phase error and MPE2 is the mean phase error for the inverted phases.

five cycles of weighted Fourier syntheses followed by a peak search were carried out, increasing the number of peaks by 50 in each step. The *R* factor (based on *E* magnitudes) for 300 peaks in the last cycle with 18532 reflections at 1 Å resolution was 40.2%. The correspondence between the peaks in the last *E* map and the known atomic positions (Dauter *et al.*, 1992) were as follows. The top 99/100, 196/200, 293/300, 372/400 and 405/480 peaks were within 0.5 Å of the true atomic positions. Starting with set number 264 in Table 2(*b*) for 1.5 Å resolution and repeating the process of phase extension to full 1 Å data followed by weighted Fourier recycling, the final *E* map showed an *R* value of 39.9% for 18532 data and had 98/100, 197/200, 294/300, 375/400 and 403/480 top peaks within 0.5 Å of the true atomic sites. The connectivity of peaks in the final *E* map looked promising for the protein model building. At this stage, the least-squares refinement of the model structure from *SAYTAN* was considered. It should be noted that the structure of rubredoxin has been determined by direct methods (Sheldrick *et al.*, 1993), real/reciprocal space recycling (Sheldrick & Gould, 1995) and the combined reciprocal-space/direct-space *Shake-and-Bake* procedure (Hauptman, 1995), based on 0.92 Å resolution data. With 1.5 Å resolution data, the structure could also be solved using a small number of known phases from the three-beam diffraction experiments as input for the direct method (Mo *et al.*, 1996) and from the Patterson interpretation of the FeS_4 cluster followed by *E* Fourier recycling (Sheldrick *et al.*, 1993).

4. Least-squares refinement of the model

The largest 480 peaks of the last *E* map were used as the initial starting model for the least-squares refinement with *SHELXL*93 (Sheldrick, 1993). The first peak in the final *E* map was assumed to be the Fe atom, while four other strong peaks ∼2.2 Å distant from the first peak and displaying an approximately tetrahedral geometry were considered to be four S atoms. The remaining peaks were taken to be C atoms. All atoms were assigned with a unit occupancy factor and an isotropic temperature factor (*U*) of 0.07 $Å^2$, the value obtained from the Wilson plot. No stereochemical restraint was applied during refinement. From 18532 reflections at 1 Å resolution, 10% of the data was excluded and used to calculate a free *R* value (Brünger, 1992). At the beginning of the refinement, the residual was 0.388 and the free *R* value was 0.495. 12 least-squares cycles of positional and individual-atom isotropic temperature-factor refinement using the Konnert & Hendrickson conjugate-gradient algorithm (CGLS) reduced the residual and the free *R* value to 0.208 and 0.261, respectively. At this stage, the S atom of the SO_4^- ion with four bonding distances ∼1.45 Å and the methionine S atom could easily be identified from the electron-density map. The atoms with only one bonding distance ∼1.2 Å were relabelled as oxygen (C═O) and those having no bond indication in the connectivity list were considered to be O atoms of the solvent water molecules. The iterative process of rebuilding and refinement of the model structure consisted of two steps. While the atoms involved in chemically unacceptable interatomic contacts (<1.0 Å) and exhibiting high thermal parameters (>0.9 $Å^2$) were excluded from the current list, new atoms were introduced in the model from the $(F_o - F_c)$ difference Fourier map. Similar procedures for automatically rejecting and adding atoms to the protein model *via* iterative

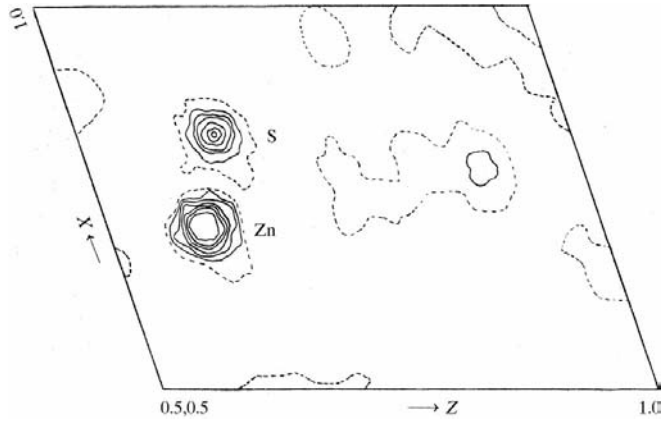

Figure 3
A section of electron-density map with refined phases at 1.0 Å.

least-squares minimization followed by difference Fourier synthesis or *E*-map recycling have been described by Lamzin & Wilson (1993) and Sheldrick (1997). The iterative process of rebuilding and refinement of the structure continued until convergence (no additional atom could be included), when 80 atoms in the initial model had been replaced by 61 new ones leaving a total of 461 atomic sites. Of these, 384 sites were finally identified as protein atoms and the remaining 77 sites belonged to the O atoms of the solvent water molecules. H atoms were not located or placed at the calculated positions. In the following cycles of full-matrix least-squares refinement, anisotropic temperature factors were used for the atoms of the FeS_4 unit while other atoms were treated isotropically. The residual and free R value dropped to 0.157 (for 15937 reflections) and 0.176 (for 1589 reflections), respectively. The final residual was 0.145 for 11972 data with $F_o > 4\sigma(F_o)$ and U_{iso} values for the protein and the solvent atoms lying in the ranges 0.05–0.22 and 0.14–0.71 Å^2, respectively. The maximum residual electron density in the final difference Fourier map was 0.68 e Å^{-1}. The bond lengths and angles of the refined structure, with Fe—S distances ranging between 2.22 and 2.35 Å, agree well with the reported values for rubredoxin (Dauter *et al.*, 1992). A section of the electron-density map with refined phases at 1.0 Å is given in Fig. 3, while comparisons of molecular views of the rubredoxin model based on the present work and that of Dauter *et al.* (1992) are shown in Figs. 4(*a*) and 4(*b*).

It is worth noting that starting with a good phase set followed by phase extension and *E*-map calculation with data beyond 1.2 Å resolution, our attempts to build the protein model and subsequently refine it through least-squares were not successful. Although unrestrained least-squares refinement with the top 400 peaks of the *E* map input into *SHELXL*93 reduced the residual by ∼0.20, the model building could not be completed owing to poor connectivity in the resulting electron-density maps.

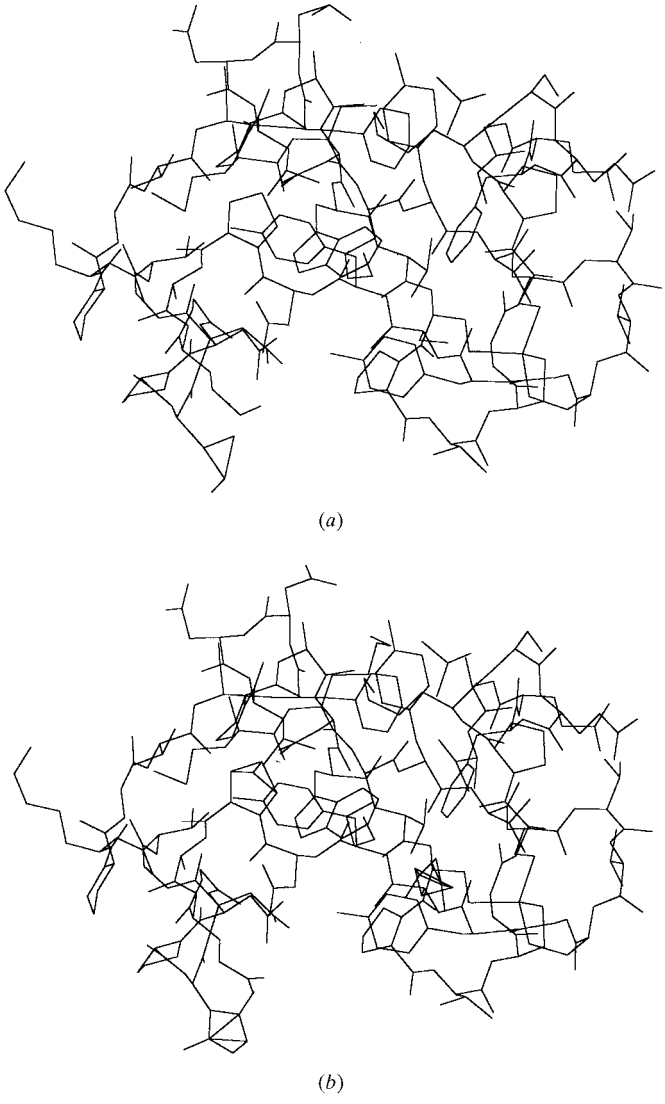

Figure 4
Molecular views of rubredoxin based on (*a*) the refined atomic coordinates of the present work, (*b*) the atomic coordinates of Dauter *et al.* (1992).

5. Concluding remarks

The *ab initio* phase determination of rubredoxin, a 393-atom protein including an FeS_4 unit and 102 solvent water molecules, at different resolutions clearly demonstrates the power of *SAYTAN* in its present form. With the atomic resolution data, the procedure described above seems capable of leading to an interpretable *E* map for rubredoxin to be followed by least-squares refinement of the model structure. Although the use of truncated high-resolution (1 Å) data for rubredoxin, which are intrinsically better than actual low-resolution data from a protein, and the presence of an FeS_4 cluster in rubredoxin appear to have reduced the scale of the problem, the ability to extract good phase sets even down to 2.25 Å resolution enhances confidence in our ongoing direct-method experiments with larger proteins without any heavy-atom or lower resolution data. The results of the investigation will be published in due course.

The author is indebted to Professor M. M. Woolfson for his encouragement and help in revising the manuscript. She is grateful to Dr A. K. Mukherjee for many helpful suggestions. Professor P. Lindley, co-editor of *Acta Crystallographica Section D*, and the referees are thanked for their valuable comments.

References

Anderson, D. S., Weiss, M. S. & Eisenberg, D. (1996). *Acta Cryst.* D**52**, 468–480.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.
Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* B**48**, 42–59.
Debaerdemaeker, T., Germain, G., Main, P., Refaat, L. S., Tate, C. & Woolfson, M. M. (1988). *MULTAN*88. *A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Universities of York, England and Louvain, Belgium.
Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1988). *Acta Cryst.* A**44**, 353–357.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
Gilmore, C. J., Henderson, A. N. & Bricogne, G. (1991). *Acta Cryst.* A**47**, 842–846.
Glover, I., Haneef, I., Pitts, J.-E., Wood, S. P., Moss, D., Tickle, I. J. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293–304.
Hauptman, H. A. (1991). *Crystallographic Computing 5: From Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. IUCr/Oxford University Press.
Hauptman, H. A. (1995). *Acta Cryst.* B**51**, 416–422.
Helliwell, J. R., Ealick, S., Doing, P., Irving, T. & Szebenyi, M. (1993). *Acta Cryst.* D**49**, 120–128.
Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* A**34**, 863–870.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Mishnev, A. F. & Woolfson, M. M. (1994). *Acta Cryst.* D**50**, 842–846.
Mo, F., Mathiesen, R. H., Hauback, B. C. & Adman, E. T. (1996). *Acta Cryst.* D**52**, 893–900.
Mukherjee, M., Ghosh, S. & Woolfson, M. M. (1999). *Acta Cryst.* D**55**, 168–172.
Mukherjee, M. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 9–12.
Mukherjee, M. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 626–628.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Sheldrick, G. M. (1993). *SHELXL*93. *Program for the Refinement of Crystal Structures.* University of Göttingen, Germany.
Sheldrick, G. M. (1997). *Methods Enzymol.* **277**, 319–343.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A, Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
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.
White, P. & Woolfson, M. M. (1975). *Acta Cryst.* A**31**, 53–56.
Wilson, A. J. C. (1942). *Nature (London)*, **150**, 151.
Woolfson, M. M. & Yao, J.-X. (1988). *Acta Cryst.* A**44**, 410–413.
Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.
Zhang, K. Y.-J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.
Zhang, K. Y.-J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.