Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A direct-method *ab initio* phasing of a protein, cupredoxin amicyanin, at 1.31 Å resolution

Monika Mukherjee,[a] Susim Maiti,[a] Soma Ghosh[b] and M. M. Woolfson[c]*

[a]Department of Solid State Physics, Indian Association for the Cultivation of Science, Calcutta-700032, India, [b]Department of Physics, St Xavier's College, Calcutta-700016, India, and [c] Department of Physics, University of York, York Y01 5DD, England

Correspondence e-mail: mmw1@york.ac.uk

Received 29 January 2001
Accepted 20 June 2001

The direct-methods program *MULTAN*88 has been applied successfully to redetermine the structure of a protein, cupredoxin amicyanin, containing 808 non-H atom sites, one Cu atom and 132 ordered water molecules in the asymmetric unit using data at 1.31 Å resolution. Starting with initially random phases, useful phase sets selected by figures of merit could be obtained from multiple trials. The *E* maps corresponding to the best eight phase sets in order of combined figures of merit (CFOM2) revealed a distorted tetrahedral geometry around the Cu site. The phase estimates from the metal and a few neighbouring atoms in the initial *E* map corresponding to the set with the highest CFOM2 could be improved by the density-modification procedure *PERP* and led to an interpretable electron-density map.

1. Introduction

Developments in protein data-collection techniques using synchrotron-radiation sources, advances in computer speed and efficient computational methods have made it possible to determine the *ab initio* structures of protein by multisolution direct methods using either real-space or reciprocal-space approaches when the diffraction data extend to atomic resolution. However, the crystals of macromolecules rarely yield X-ray diffraction data beyond 1.2 Å resolution, the threshold for atomic resolution in biomolecules (Sheldrick *et al.*, 1993). The lack of high-resolution data for macromolecules hinders not only atomic modelling but also *ab initio* phasing of native crystals by probabilistic direct methods. Present-day direct-method procedures (Debaerdemaeker *et al.*, 1988; Hauptman, 1995; Sheldrick & Gould, 1995; Burla *et al.*, 1999, 2000) have been remarkably successful in phasing native protein structures with as many as ∼2000 independent non-H atoms in the crystallographic asymmetric unit. Thus, structure solution of proteins by direct methods from a single native data set has been successful with atomic resolution data, *i.e.* 1.2 Å or better (Smith *et al.*, 1997; Mukherjee, 1999; Mukherjee *et al.*, 1999; Parisini *et al.*, 1999; Schneider *et al.*, 2000). With data of resolution less than 1.2 Å, a poor interpretability of the resulting *E* map owing to lack of connectivity between the amino-acid residues has been reported by Sheldrick & Gould (1995) and Mukherjee *et al.* (2000). Recently, *ab initio* solution of a protein, a serine protease inhibitor from the leech *Hirudo medicinalis*, with unknown structure and without any metal atoms has been reported by Usón *et al.* (1999) using either 1.4 Å resolution room-temperature data or 1.2 Å resolution low-temperature data.

In the present paper, we have redetermined the structure of cupredoxin amicyanin, a protein located in the periplasm of a

Table 1
Results of applying *MULTAN*88 to cupredoxin for various NREF with 1.31 Å data.

In each case 1000 trials were made.

NREF†	NREL‡	KMIN§	TOTSET¶	NG††	LMPE1 (LMPE2)‡‡ (Å)
1000	12777	0.20	1000	15	66.3 (67.8)
1500	42714	0.20	1000	8	68.0 (69.2)
2000	99464	0.20	1000	11	68.6 (69.2)

† No. of reflections with largest $|E|$. ‡ No. of three-phase relationships. § Minimum value of K for any three-phase relationship, where $K(h,k) = 2\sigma_3\sigma_2^{-3/2}|E(h)E(k)E(h-k)|$, $\sigma_n = \sum_{j=1}^{N} Z_j^n$ and where the jth of N atoms in the unit cell has atomic number Z_j. ¶ Total No. of phase sets generated. †† No. of phase sets with mean phase error $\leq 72^\circ$. ‡‡ LMPE1, the lowest mean phase error for the NG good sets; LMPE2, the lowest mean phase error for the enantiomorph for the NG good sets.

number of methylotropic bacteria, from a single data set extending to 1.31 Å resolution. The crystals of amicyanin isolated from *Paracoccus denitrificans* contain 808 protein atoms (105 amino-acid chain of amicyanin; 11.5 kDa molecular mass), 132 ordered water molecules and one Cu atom in the asymmetric unit. The data used in the present analysis were obtained from the Protein Data Bank (PDB code 1aac) and consist of 21 131 unique reflections in the resolution range 8.0–1.31 Å. The crystal system is monoclinic, space group $P2_1$, with unit-cell parameters a = 28.95 (3), b = 56.54 (6), c = 27.55 (3) Å, β = 96.38 (5)°. The experimental details of data collection using graphite-monochromated X-rays λ = 1.5418 Å) from a Rigaku RU-200 generator are given by Cunane *et al.* (1996). However, the knowledge of the known structure of cupredoxin was not utilized at any stage either of structure solution by the direct method or of phase improvement by the density-modification procedure. The known structure was used to calculate MPE1 and MPE2 (the mean phase error from the phases of the published structure and its enantiomorph, respectively) and MCC (the map correlation coefficient), quantities which are given just to illustrate the effectiveness of our procedures.

In the present study with 1.31 Å resolution data, the main chain and some parts of the remaining structure could be found from the maps obtained by the direct-method phasing. In order to be able to complete the protein model building with the 1.31 Å data, we improved the initial phases obtained from the direct method by a density-modification approach.

2. The *ab initio* structure solution

In describing our structure solution as *ab initio* we mean that we have used single-wavelength native data without assuming any knowledge of the structure other than its general composition. Starting with a few thousand of the largest normalized structure factors from the 1.31 Å resolution data, the procedure essentially involves a search for probable solutions selected by the discriminating figures of merit (FOMs) that were devised by Mukherjee & Woolfson (1993, 1995). For each of several runs of *MULTAN*88 (Debaerdemaeker *et al.*, 1988) with random initial phases developed by

Table 2
A selection of figures of merit for various phase sets generated at 1.31 Å resolution.

Good sets are marked with an asterisk.

SET	ABSM†	PSIM‡	RESM§	CFOM2¶	DEVI††	MPE1 (MPE2)
101	0.898	0.287	24.62	1.734	42.37	85.2 (82.8)
271	1.282	0.742	35.84	2.502	27.35	68.6 (68.2)*
275	0.921	0.354	26.60	1.797	42.02	85.4 (84.2)
313	0.894	0.276	25.50	1.689	42.64	85.3 (84.1)
450	1.292	0.778	35.34	2.577	25.48	69.3 (68.2)*
487	0.887	0.304	26.44	1.692	42.62	84.5 (84.4)
503	1.325	0.763	35.46	2.593	25.67	67.8 (66.3)*
505	0.903	0.317	24.82	1.775	41.77	82.4 (83.5)
527	1.293	0.736	35.76	2.509	26.37	68.3 (68.3)*
541	0.886	0.303	26.57	1.686	42.47	84.4 (83.9)
570	1.301	0.727	35.24	2.522	26.24	67.9 (67.5)*
622	0.912	0.320	24.36	1.803	42.44	85.5 (85.4)
745	1.277	0.783	37.59	2.501	25.60	69.0 (70.0)*
768	1.305	0.744	36.04	2.564	27.39	68.7 (68.0)*
813	1.286	0.751	35.63	2.524	26.31	66.8 (67.9)*

† $s/s_{\rm exp}$. ‡ $\sum_l |\sum_k E(k)E(h-k)|/s$. § $\sum_\alpha |\{[\alpha(h)]/s\} - \{[\alpha(h)_{\rm est}]/s_{\rm est}\} \times 100$, where $\alpha(h) = |\sum_k E(k)E(h-k)|$, $s = \sum_h \alpha(h)$; the subscript exp corresponds to the value for the true phases and the summation over h is for large E values and that over l for small E values. ¶ $w_1[({\rm ABSM} - {\rm ABSM_{min}})/({\rm ABSM_{max}} - {\rm ABSM_{min}})] + w_2[({\rm PSIM} - {\rm PSIM_{min}})/({\rm PSIM_{max}} - {\rm PSIM_{min}})] + w_3[({\rm RESM} - {\rm RESM_{min}})/({\rm RESM_{max}} - {\rm RESM_{min}})]$, where the subscripts max and min correspond to the maximum and minimum values for the 1000 phase sets and the weights are set at $w_1 = w_2 = w_3 = 1.0$. †† $\langle \min(|\Phi_{3,i}|, |180 - \Phi_{3,i}|)\rangle_i$, where $\Phi_{3,i}$ is the value in degrees of the ith three-phase invariant.

the magic integer series (White & Woolfson, 1975) followed by phase refinement using the Hull & Irwin (1978) weighting scheme, 1000 phase sets were generated. Table 1 shows the results of *ab initio* phase refinements of cupredoxin with the 1000, 1500 and 2000 largest E values, with different seeds for the random-number generator and the value of KMIN (the quantity governing the lowest acceptable variance for a triple-phase relationship) taken as 0.20. The trial with the 1000 largest E values yielded 15 good solutions having MPE < 72°; the corresponding numbers with the 1500 and 2000 largest E values were 8 and 11, respectively. A selection of figures of merit and mean phase errors (MPE) with the 1000 largest E values is shown in Table 2, an examination of which reveals that good phase sets could be recognized by the large values of ABSM, PSIM and, especially, of CFOM2. Good sets having MPEs less than 72° are marked with asterisks. The top eight good sets, selected on the basis of CFOM2, were used to compute E maps that were interpreted by the peak-search routine of *MULTAN*88. All these maps showed a distorted tetrahedral arrangement of ligands around the Cu site. The set with the highest CFOM2 and lowest mean phase error in Table 2 (set No. 503) was selected for subsequent analysis. The highest peak in the E map was assumed to be the Cu atom, while four other peaks approximately 2.2 Å away from the highest one and displaying a distorted tetrahedral geometry around it were taken to be two S and two N atoms. Using the top 100 peaks, an iterative process of weighted Fourier syntheses followed by peak-search selection was carried out, increasing the number of peaks by 50 in each cycle. Once 800 peaks had been identified, the process of model building was carried out by successive difference Fourier maps, taking account of acceptable bond-distance and bond-angle criteria.

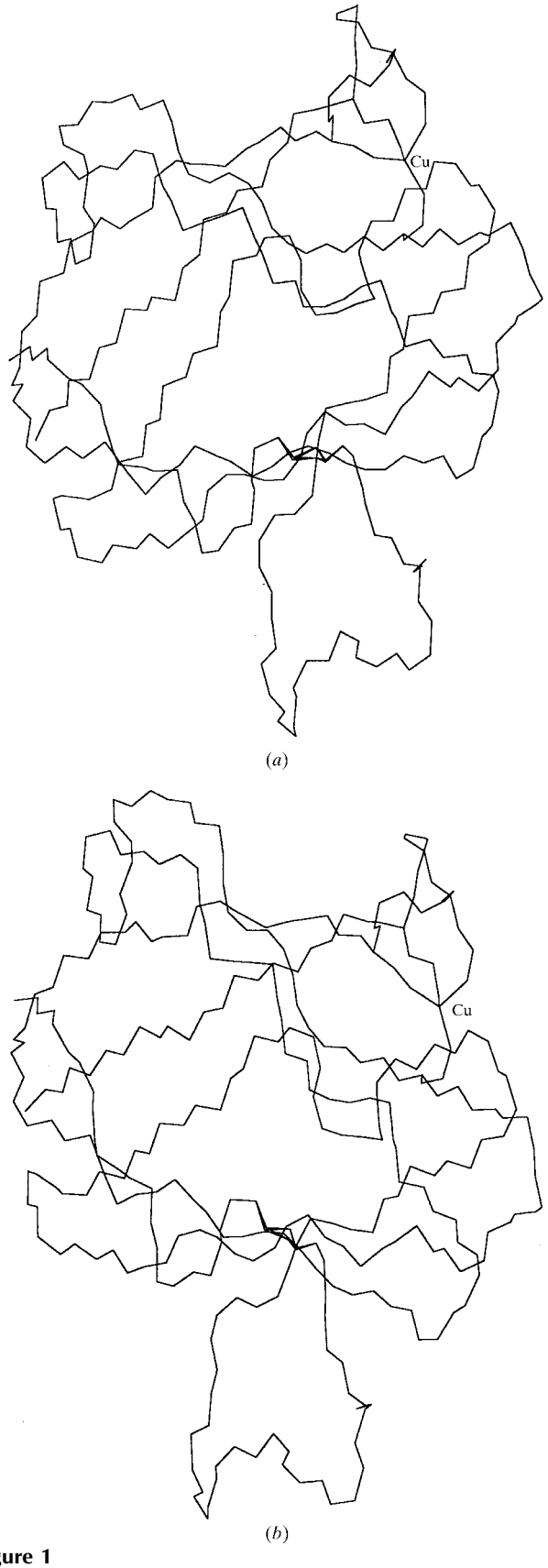

Figure 1
(*a*) Molecular view of the main-chain of cupredoxin amicyanin based on the coordinates obtained from the present work. (*b*) Molecular view of the main-chain of cupredoxin amicyanin based on the coordinates obtained from Cunane *et al.* (1996).

In protein crystallography, model building from the electron-density map is a time-consuming critical step and human intervention is often regarded as necessary in order to trace the connectivity in the polypeptide chains. In the present case, the main-chain structure of the protein was found; the views of the main-chain structure of cupredoxin as obtained from the direct method and that from the published coordinates (Cunane *et al.*, 1996) are given in Figs. 1(*a*) and 1(*b*), respectively. (The r.m.s. coordinate agreement, using *LSQKAB*, between the two coordinate sets is 0.23 Å for the main chains.)

Owing to missing atoms and noise in the electron-density maps, side-chain tracing and completing the protein model was not possible or would have been very difficult, time-consuming and only partially successful. To overcome this problem, we decided to make use of a density-modification procedure to improve further the phase estimates.

3. Density modification to improve the starting phases

With the advancement of powerful computers, a number of density-modification techniques have been developed over the last few years to improve the starting phases without introducing any model bias, *e.g.* solvent flattening (Wang, 1985), histogram matching (Zhang & Main, 1990) and the application of the Sayre equation (Sayre, 1972). In the present case, the program *PERP* (phase extension and refinement program), developed by Refaat *et al.* (1996*b*), has been used for density modification. The program includes several separate density-modification processes: SE, Sayre's equation refinement (Refaat *et al.*, 1995); LE, low-density elimination (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993); HM, histogram matching (Zhang & Main, 1990); DM, double histogram matching involving the local maximum density (Refaat *et al.*, 1996*a*); SQ, squaring of density (taking the phases of the squared current density); DV, double histogram matching involving the local density variance (Refaat *et al.*, 1996*a*). All the histogram-matching methods (HM, DM and DV) automatically include solvent flattening. Given a set of initial phases, the program goes through cycles of density modification and finding new phases. In the present version of *PERP*, the original LE component was replaced by an improved version (Matsugaki & Shiono, 1999).

The results of several trials of density modification for cupredoxin with *PERP* are summarized in Table 3. In the first trial, the initial phases were calculated from the Cu position alone obtained from set 503 of the 1000-reflection *MULTAN*88 run. In the next five trials, neighbouring light atoms obtained directly from the *E* map corresponding to the initial phase set of *MULTAN*88 were added stepwise – first (Cu, 1S), then (Cu, 2S) followed by (Cu, 2S, 1N), (Cu, 2S, 2N) and (Cu, 2S, 2N, 25C). To monitor the progress of density modification, MPEs were calculated allowing for the necessary origin shift to give the best agreement between the calculated phases and either the published phases (MPE1) or the published phases subtracted from 360° (for the enantiomorph MPE2). The map correlation coefficient (MCC), which was strongly correlated with the MPE and provided a very reliable

indication as to whether a correct solution was emerging or not, was calculated after each cycle. Usually, an MCC value greater than 0.50 corresponds to a basically correct solution. In Table 3, the initial MPEs (MPE1 and MPE2) were close, indicating the presence of a possible pseudo-centre of symmetry. Although it has been observed that tangent-formula refinement tends to push phases towards a pseudo-centrosymmetric solution, especially in the presence of heavy atoms, phase refinement by a few cycles of *PERP* effectively broke the pseudosymmetry.

An analysis of the results in Table 3 shows that except in trial 1, 7660 phases of reflections with $|E| > 1.0$ were satisfactorily refined. After 200 cycles of refinement, the starting MPE values of 68 $\pm$ 4° dropped to ~35 $\pm$ 2.5° and the MCC increased from ~0.30 to ~0.78. Thus, both the MPE and MCC values substantially improved and enantiomorph discrimination became very positive with *PERP* refinement. It has been previously observed that the use of several methods of density modification in sequence, as happens in *PERP*, is more effective than the use of any single method (Mukherjee *et al.*, 2000). With more reflections included in the *PERP* phase extension, *i.e.* for $|E| > 0.9$ (9431 reflections), both the MPE and MCC values converge well, as shown in Table 3. The results of *PERP* refinement using coordinates from Cunane *et al.* (1996) in Table 4 indicate that even with the published coordinates of cupredoxin, phase refinement starting with the Cu atom alone is not very effective. For the other trials, however, results are quite similar to those of Table 3. The superposition of the electron-density map prepared using the program *O* (Jones *et al.*, 1991) around the metal position as obtained from the present experiment and the backbone of the reported structure (Cunane *et al.*, 1996) is shown in Fig. 2.

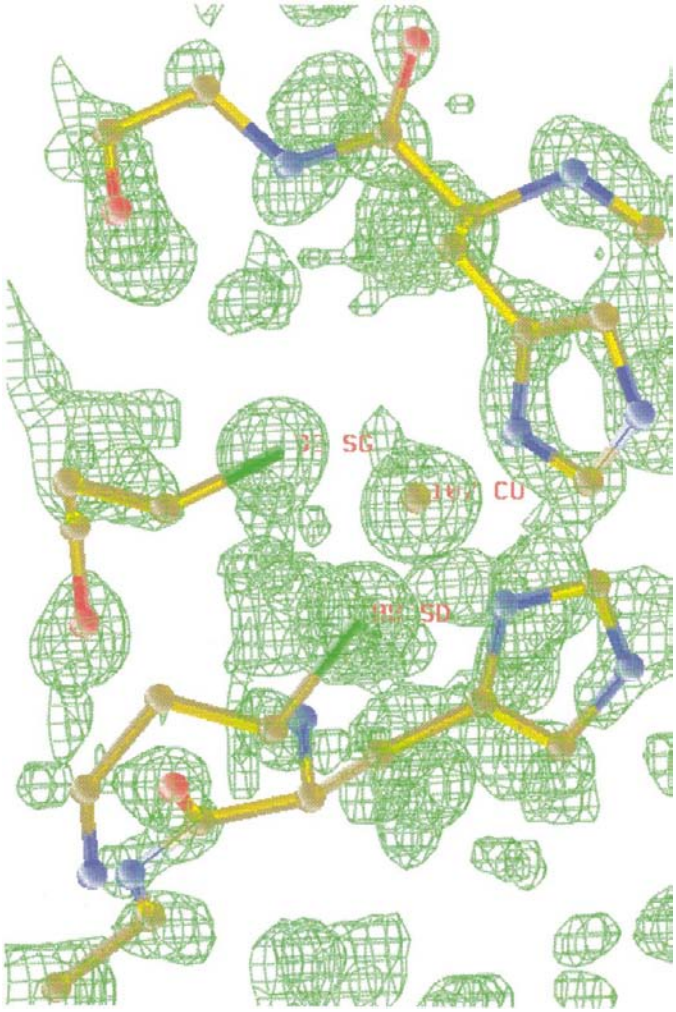

Figure 2
A section of the electron-density map showing the copper coordination sphere for the present work.

Table 3
Results of phase refinement using *PERP* with starting phases derived from coordinates obtained using *MULTAN*88.

MPEs and MCCs were calculated against the refined model. MCC is the standard map correlation coefficient.

		MPEs		MCC	
No. of reflections	Starting phases with coord.	Initial MPE1 (MPE2)	Final MPE1 (MPE2)	Initial	Final
7660 ($E > 1.0$)	Cu	71.6 (71.6)	80.5 (67.1)	0.280	0.382
7660 ($E > 1.0$)	Cu + 1S	74.5 (69.3)	83.5 (33.0)	0.313	0.793
7660 ($E > 1.0$)	Cu + 2S	75.1 (71.1)	83.3 (34.4)	0.290	0.781
7660 ($E > 1.0$)	Cu + 2S + 1N	75.4 (70.9)	83.6 (32.9)	0.291	0.799
7660 ($E > 1.0$)	Cu + 2S + 2N	75.8 (69.0)	83.4 (36.5)	0.320	0.760
7660 ($E > 1.0$)	Cu + 2S + 2N + 25C	77.6 (64.6)	83.6 (32.6)	0.300	0.799
9431 ($E > 0.9$)	Cu + 2S + 1N	76.6 (71.5)	84.4 (37.5)	0.288	0.757

4. Concluding remarks

Our success in solving the cupredoxin structure, which was by objective procedures that should be applicable to similar but unknown structures, depends on three important factors. The first of these is that a very basic direct-methods approach (*MULTAN*88 in this case, but it could be another) is capable of generating some phase sets that are good enough to indicate some prominent features of the structure. The determination of not only the Cu position but also the positions of some small number of coordinated atoms was important in this case. This was also the case for the solution of pseudoazurin, a copper-containing protein that was solved at 1.55 Å resolution (Mukherjee *et al.*, 2000). However, the presence of a heavy atom is not absolutely necessary to obtain the solution of a small protein (~1000 non-H atoms) if good data of high resolution is available (*e.g.* Mukherjee *et al.*, 1999).

The second important feature is that it must be possible to recognize the good phase sets from the very large number that may be generated. The traditional figures of merit that were developed with the earlier versions of *MULTAN* are not capable of doing this, but the newer FOMs developed by Mukherjee & Woolfson (1993, 1995) discriminate well

Table 4
Results of phase refinement using *PERP* (starting phases generated from coordinates of Cunane *et al.*, 1996).

No. of reflections with $\|E\| > 1.0$	Starting phases with coord.	MPEs		MCC	
		Initial MPE1 (MPE2)	Final MPE1 (MPE2)	Initial	Final
7660	Cu	71.3 (71.4)	77.5 (74.7)	0.282	0.270
7660	Cu + 1S	74.7 (68.7)	83.5 (33.0)	0.319	0.795
7660	Cu + 2S	76.0 (67.1)	83.3 (36.5)	0.347	0.760

between phase sets even when the 'good' sets have phase errors $\sim 70^\circ$. With the better phase sets found and some prominent features picked up in an E map, then a standard process of successive weighted Fourier syntheses selecting more and more peaks in each cycle usually gives a significant increase in the number of determined atomic positions. When we had found 800 top peaks in our maps we were able, with some effort, to trace the main chain, but further progress, although possible, was very tedious.

The final component of our success was the availability of a very effective density-modification procedure, *PERP*. We started with very minimal information from our initial E map, Cu and a few neighbouring atoms, and from this we objectively generated greatly improved phases and a high-quality map that would be amenable to automatic interpretation.

We shall be exploring how far this kind of approach can be taken. We have already shown that with a heavy atom present it is possible to solve structures with data of resolution lower than 1.2 Å, an often-quoted limit. We have also demonstrated that with atomic resolution data a small protein can be solved without the presence of a heavy atom. Since a large number of proteins contain S atoms, our future work will be directed towards studying the applicability of this approach in solving the structures of proteins where sulfur is the heaviest atom.

References

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (1999). *Acta Cryst.* A**55**, 991–999.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2000). *Acta Cryst.* A**56**, 451–457.

Cunane, L. M., Chen, Z.-W., Durley, R. C. E. & Mathews, F. S. (1996). *Acta Cryst.* D**52**, 676–686.

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-la-Neuve, Belgium.

Hauptman, H. (1995). *Acta Cryst.* B**51**, 416–422.

Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* A**34**, 863–870.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Matsugaki, N. & Shiono, M. (1999). Abstr. XVIIIth IUCr Congr., Abstract P12.02.011.

Mukherjee, M. (1999). *Acta Cryst.* D**55**, 820–825.

Mukherjee, M., Ghosh, S. & Woolfson, M. M. (1999). *Acta Cryst.* D**55**, 168–172.

Mukherjee, M., Maiti, S. & Woolfson, M. M. (2000). *Acta Cryst.* D**56**, 1132–1136.

Mukherjee, M. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 9–12.

Mukherjee, M. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 626–628.

Parisini, E., Francesco, C., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 1036–1040.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1996*a*). *Acta Cryst.* D**52**, 252–256.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1996*b*). *Acta Cryst.* D**52**, 1119–1124.

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.

Sayre, D. (1972). *Acta Cryst.* A**28**, 210–212.

Schneider, T. R., Kärcher, J., Pohl, E., Lubini, P. & Sheldrick, G. M. (2000). *Acta Cryst.* D**56**, 705–713.

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.

Shiono, M. & Woolfson, M. M. (1992). *Acta Cryst.* A**48**, 451–456.

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.

Usón, I., Sheldrick, G. M., de La Fortelle, E., Bricogne, G., Di Marco, S., Priestle, J. P., Grutter, M. G. & Mittl, P. R. E. (1999). *Structure*, **7**, 55–63.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.

White, P. & Woolfson, M. M. (1975). *Acta Cryst.* A**31**, 53–56.

Zhang, K. Y.-J. & Main, P. (1990). *Acta Cryst.* A**46**, 41–46.