Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the application of phase relationships to complex structures. XXXVI: some experiments with a small protein without heavy atoms

M. Mukherjee,[a] S. Ghosh[a] and M. M. Woolfson[b]*

[a]Department of Solid State Physics, Indian Association for the Cultivation of Science, Calcutta 70032, India, and [b]Department of Physics, University of York, York YO1 5DD, England

Received 15 January 1998
Accepted 1 May 1998

The direct-methods program *MULTAN*88 has been applied to a known protein, ribonuclease (RNAP1), containing 808 non-H atoms, including five S atoms, plus 83 ordered solvent water molecules. Phase sets with mean phase errors between 69 and 75° were selected by modified figures of merit for trials with the full data at 1.17 Å resolution and also with restricted data at 1.25 and 1.5 Å resolution. These figures of merit had previously only been applied to protein structures containing heavy atoms, and this is the first demonstration of their usefulness with no heavy atom present. An initial set of 1091 phases from a 1.17 Å trial was developed by an objective procedure to give the full structure with a residual of 0.21, which agrees well with the published structure.

1. Introduction

In recent years there have been some significant advances in developing direct methods for the *ab initio* solution of protein structures. The first demonstration that this was possible was given by Woolfson & Yao (1990), with the solution of the small protein aPP (avian pancreatic polypeptide) by the direct-methods program *SAYTAN*. This structure, the asymmetric unit of which contained 36 amino acids plus 80 H_2O molecules and a Zn atom had data to resolution 0.97 Å. This was less than a factor of two greater in size than some of the larger 'small' structures which had previously been solved by direct methods. It was only possible to find the solution because the structure was already known, since the conventional *MULTAN* figures of merit were unable to pick out the good phase sets. Other successful demonstrations of protein-structure solution followed, *e.g.* Sheldrick *et al.* (1993), Miller *et al.* (1993), Weeks *et al.* (1993), Sheldrick & Gould (1995), Hauptman (1995) and Mukherjee & Woolfson (1993, 1995).

During the earlier years of this development it became accepted as a rule-of-thumb condition by workers in this field that for success it was essential to have good-quality data of resolution 1.2 Å or better. However, it has been shown (Mukherjee & Woolfson, 1993) that, for aPP, phase sets with 69° mean phase error (MPE) could be obtained for 315 reflections at 3.0 Å resolution and, perhaps more significantly, 54° MPE for 556 reflections at 1.77 Å resolution. It was subsequently shown (Mukherjee & Woolfson, 1995) that useful starting sets of phases could be obtained for 2-Zn insulin, even at 2.25 Å resolution. Of some importance in these two investigations was the development of revised figures of merit which were able to distinguish the better sets of phases. Yet another successful figure of merit was developed by Mishnev & Woolfson (1994) although, since it requires the calculation of an E map for each phase set, it is

Table 1
Results of applying *MULTAN*88 to RNAP1 at different resolutions.

Resolution (Å)	NREF†	NREL‡	KMIN§	TOTSET¶	NG††	LMPE (°)‡‡
1.17	800	7091	0.30	431	12	70.00
1.25	800	7112	0.30	431	23	69.20
1.50	800	7894	0.30	431	11	73.80

† The number of reflections with large $|E|$. ‡ The number of linking three-phase relationships. § The minimum value of κ for any three-phase relationship. ¶ The total number of phase sets generated. †† The number of sets with mean phase error less than 75°. ‡‡ The lowest mean phase error in the NG sets.

somewhat expensive to apply. Thus, the first step on the path to solution of these structures could be made by an objective procedure and no prior knowledge was required.

All previous investigations in which we have been concerned have involved structures containing one or more heavy atoms, and we have always thought that this was an important, perhaps even necessary, factor in the successful application of both *SAYTAN* and the revised figures of merit, especially at lower resolutions. We have therefore been exploring what would be possible when there is no heavy atom in the structure and our latest results are presented here.

2. Phase determination for RNAP1

The structure of ribonuclease, RNAP1, (Bezborodova *et al.*, 1988) contains 96 amino acids (808 non-H atoms, including five S atoms) together with 83 ordered water molecules. The space group is $P2_1$ with unit-cell parameters a = 32.01, b = 49.76, c = 30.67 Å, β = 115.83° and Z = 2. The results reported here were obtained by applying *MULTAN*88 (Debaerdemaeker *et al.*, 1988), using the weighting scheme devised by Hull & Irwin (1978), to the 800 reflections with largest $|E|$ values selected from the 23853 independent reflections within the 1.17 Å resolution limit of the data. The triple phase relationships used in the phasing procedure were limited to those with $\kappa \geq 0.3$ where

$$\kappa(\mathbf{h}, \mathbf{k}) = 2\sigma_3\sigma_2^{-3/2}|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \quad (1)$$

and

$$\sigma_n = \sum_{j=1}^{N} z_j^n.$$

We adopted the procedure suggested by Woolfson & Yao (1990) and subsequently used with 2-Zn insulin (Mukherjee & Woolfson, 1995), where the phases of the 50 reflections with the largest $|E|$ values were kept at their initial random values until the very last cycle of tangent-formula refinement. The modified figures of merit proposed by Mukherjee & Woolfson (1993) were then applied to the developed phase sets and this confirmed that they were as successful with this non-heavy-atom-containing structure as with applications to 2-Zn insulin and rubredoxin (Mukherjee, unpublished results). These figures of merit are:

$$\mathrm{ABSM} = s/s_{\mathrm{exp}}, \quad (2)$$

where

$$s = \sum_{\mathbf{h}} \alpha(\mathbf{h}), \;\; \alpha(\mathbf{h}) = \Big|\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k})\Big|$$

and the subscript exp indicates the theoretically expected value with true phases;

$$\mathrm{PSIM} = \Big[\sum_{\mathbf{l}} \Big|\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{l}-\mathbf{k})\Big|\Big]/s, \quad (3)$$

where the summation over **l** is for small $|E|$'s;

$$\mathrm{RESM} = \sum_{\mathbf{h}} \left| \frac{\alpha(\mathbf{h})}{s} - \frac{\alpha(\mathbf{h})_{\mathrm{exp}}}{s_{\mathrm{exp}}} \right|. \quad (4)$$

These three figures of merit are combined to give a combined figure of merit, CFOM2, constructed in a similar way to the original CFOM in *MULTAN*88. However, the term involving PSIM was arranged to take larger values of PSIM as indicating good phase sets, although theory might suggest otherwise, based on our experience with aPP and 2-Zn insulin. Another figure of merit which was computed but not used for the automatic selection of the better phase sets was

$$D = \langle \min(|\varphi_{3,i}|, |180 - \varphi_{3,i}|)\rangle_i, \quad (5)$$

where $\varphi_{3,i}$ is the value in degrees of the ith three-phase invariant. This was suggested by Woolfson & Yao (1990) to distinguish phase sets which give good enantiomorph discrimination; it is found, in general, that this requires a value of D greater than about 15°.

Table 1 shows the results for *ab initio* phase determination at resolutions of 1.17, 1.25 and 1.50 Å. Phase sets with a MPE of less than 75° were taken as 'good' sets, although the real test of 'good' or 'bad' is whether or not it leads to the structure solution.

All attempts in many trials to obtain MPE < 75° for resolution less than 1.5 Å were unsuccessful. The better phase sets were indicated by higher values of CFOM2 as shown in Tables 2, 3 and 4 for the three resolutions. In addition, the values of D indicate that there were no problems of enantiomorph discrimination.

3. Phase extension

As previously stated, the real test of the quality of an initial phase set is whether or not it leads to the structure. To test this for RNAP1 we selected phase set 113 (Table 2) with MPE 70°, and went through a procedure for phase extension similar to that described by Woolfson & Yao (1990) for aPP. Although for RNAP1 800 reflections with large $|E|$ values are selected by *MULTAN*88, phases were found for only 727 of them, since 9% of those least well linked to the others were discarded – a usual *MULTAN* process. The value of $\alpha(\mathbf{h})$ is a measure of the reliability of a determined phase, at least theoretically, and phases for the 300 reflections with the largest α values were input into *MULTAN*88 as 'known' phases and a further 900 reflections were selected to be phased. In the case of aPP, the procedure gave virtually the same result for every trial and

Table 2
A selection of figures of merit for phase sets at 1.17 Å.

Better sets are marked **.

Set number	ABSM	PSIM	RESM	MPE (°)	*D* (°)	CFOM2
100	4.23	0.770	34.52	85.3	32.0	0.18
113	4.56	0.838	30.46	70.1	28.10	2.29 **
145	4.81	0.749	31.33	84.4	24.80	1.75
177	4.56	0.837	31.59	71.0	26.40	2.02 **
217	4.55	0.865	30.99	71.5	26.40	2.38 **
300	4.50	0.769	31.02	83.5	29.60	1.45
395	4.61	0.816	30.65	70.8	28.70	2.13 **
402	4.68	0.836	30.93	71.8	26.90	2.37 **
425	4.62	0.835	31.56	71.3	26.70	2.10 **
428	4.50	0.825	30.93	83.6	30.20	1.25

only a few were run. For this larger structure it was found again that wide variations in the number of 'known' phases (*m*) and extended phases (*M*) made little difference to the results obtained. With the determined phases of 1091 reflections (1200 less 9%), an *E* map was calculated and interpreted by the peak-search procedure in *MULTAN*88. Starting with the top 50 peaks, an iterative process of Sim-weighted Fourier synthesis and peak-search selection was carried out, increasing the number of peaks selected by 50 in each iteration. At each stage *R* factors were found for five ranges of $(\sin\theta/\lambda)^2$, defined by the limits 0.00, 0.04, 0.07, 0.11, 0.15 and 0.18 for the structure factors calculated with the currently selected peaks. For the first cycle these *R* values were 0.63, 0.60, 0.56, 0.56, 0.53 and in the final cycle 0.39, 0.41, 0.43, 0.45, 0.44 after which there was no further improvement. At this stage the 700 determined coordinates were checked against the published structure and the numbers of correctly positioned peaks (within 1.0 Å of the true position) were 99/100, 199/200, 296/300, 396/400, 495/500, 592/600 and 681/700.

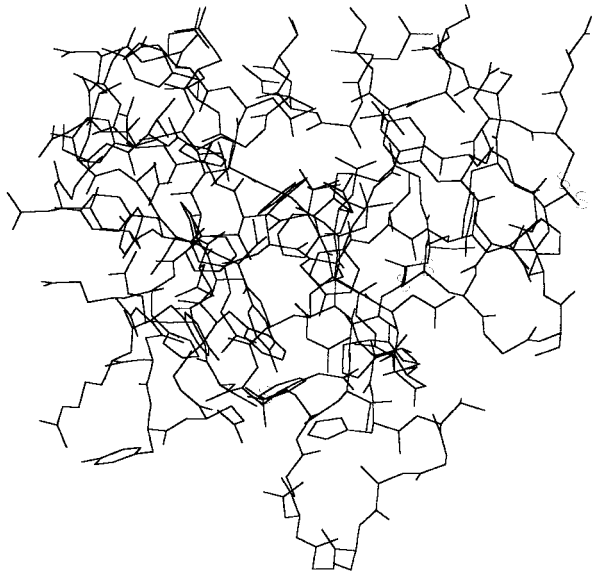
4. Structure refinement

With the 700 peaks from the last *E* map the least-squares refinement of the model was started using the Konnert–Hendrickson conjugate algorithm in *SHELXL*93 (Sheldrick, 1993). During the refinement, stereochemical constraints were applied to some of the amino-acid side chains. As the refinement progressed, new atoms and solvent water molecules indicated by difference maps were introduced into the model. The iterative process of building and refining the structure was continued until convergence was achieved. At this stage all 808 protein atoms and 83 water molecules were identified, all with unit occupancies. The final *R* value was 0.21 for 22971 observed *F*s without hydrogen included and with isotropic thermal parameters (*U*) in the range 0.07 to 0.58. The maximum residual electron density in the final difference map was 0.65 e Å^{-3}. The refined bond lengths and angles agree well with those of the published structure; the projected structure produced by this work is shown in Fig. 1(*a*) and may be compared with that of the published structure in Fig. 1(*b*).

5. Comments and conclusions

There is no doubt that, starting with the use of *MULTAN*88 as described, this small protein could have been solved *ab initio* without any prior knowledge of the structure. However, it must be noted that several different combinations of the number of large $|E|$s and the minimum value of κ were tried before a successful combination was found leading to reasonable phase sets. In that sense, we did use knowledge of the structure. On the other hand, it could be argued that since the procedures we describe are automated and quick to apply, a few abortive attempts are affordable if eventually one is successful. It can be stated quite categorically that this particular structure is amenable to solution by *MULTAN*88;

(*a*)



(*b*)

Figure 1
(*a*) Molecular view of RNAP1 based on the refined atomic coordinates of the present work ($R = 0.21$). (*b*) Molecular view of RNAP1 based on the published coordinates of Bezborodova *et al.* (1988). S denotes position of the S atom.

Table 3
A selection of figures of merit for phase sets at 1.25 Å.

Better sets are marked **.

Set number	ABSM	PSIM	RESM	MPE (°)	D (°)	CFOM2
80	4.50	0.874	31.59	71.8	27.4	1.67 **
100	4.30	0.783	33.62	79.5	32.10	0.24
245	4.34	0.751	29.29	74.1	33.40	1.04 **
271	4.54	0.864	31.68	70.6	26.20	1.62 **
300	4.41	0.798	33.24	81.7	29.60	0.62
304	4.51	0.859	32.40	70.7	27.40	1.41 **
323	5.15	0.790	31.49	69.3	22.0	1.81 **
406	4.54	0.863	30.11	70.4	26.70	1.98 **
429	4.46	0.804	30.43	71.6	30.40	1.31 **

Table 4
A selection of figures of merit for phase sets at 1.5 Å.

Better sets are marked **.

Set number	ABSM	PSIM	RESM	MPE (°)	D (°)	CFOM2
15	4.47	1.193	33.68	74.4	26.0	1.94 **
75	4.07	0.852	30.02	84.5	37.90	1.42
100	4.23	0.990	31.61	84.1	33.90	1.60
105	4.52	1.140	31.66	75.2	28.50	2.17 **
187	4.34	1.034	35.01	84.3	31.10	1.35
212	4.17	0.954	33.02	83.9	33.30	1.29
303	4.67	1.128	31.21	75.4	27.3	2.39 **
351	3.81	0.556	27.45	83.7	40.10	1.00
384	4.49	1.115	33.55	73.8	27.60	1.86 **

however, it cannot be claimed that all structures of similar size and character would be similarly amenable. Crystallographers know from experience that structures similar in structural complexity may be very different in behaviour when it comes to applying methods of solution, so that no generalization can be made from a single example. Attempt to solve other similar, or perhaps larger, structures in the same way must now be performed in order to see if the success reported here can be repeated. Another point to be investigated is to see whether or not the 'good' phase sets we found at lower resolution are really 'good' in that they can lead to useful structural information.

An interesting philosophical question is why a protein, even a small one, can be solved at all by direct methods. For such structures, the individual three-phase relationships are very weak and most of the probability distributions of values of the three-phase invariants are fairly flat over the range from $-\pi$ to $+\pi$. It was shown by Fan & Woolfson (1991) that from a theoretical point of view, the weakness of the individual phase relationships is compensated by the larger number of relationships. The possibility of solving a structure should, in principle, be dependent on resolution rather than structural complexity, although in practice this may not be completely true. What is certainly true, however, is that for low-resolution data the number of strong relationships per reflection is less, which makes finding a solution much more difficult. When direct methods are applied to proteins the phase sets obtained tend to make the three-phase relationships much closer to 0 (modulo 2π) than they really are, and this is seen in those figures of merit which depend on the quality of the relationships. Thus the ABSFOM relationship in *MULTAN*88 and our replacement ABSM have values of unity for correct phases but 'good' phase sets can give values of 4 or even greater, as seen in Tables 2, 3 and 4.

A useful way to understand why it is that proteins can be solved by direct methods is to go back to the original approach of Cochran (1952, 1955) in the development of sign relationships and general phase relationships. He gave the condition for a good phase set as one which would give an electron-density map which concentrated the density around atomic positions and was flat elsewhere. This condition he expressed as maximizing $\int \rho^3 \mathrm{d}V$ and the general form of the three-phase relationship follows naturally from this. However, we may look at this the other way round and say that a map with a large value of $\int \rho^3 \mathrm{d}V$ would tend to give density clustered into small regions with little negativity, which is the condition for atomicity. Of course, a large value of the integral can occur in other ways – for example, by creating one, or a few, very large concentrations of density with regions of moderately negative density in between, but if many different phase sets are generated then some of them may have the desired characteristic of atomicity. Now we consider the Patterson function as the convolution of the electron density with its own inverse, giving the vector set of the atomic distribution, and consider the implication that every map generated in a direct-methods procedure has the same corresponding Patterson map. If the electron-density map has atomicity then the vectors between the main peaks are going to form a set having a large overlap with that of the correct structure; they will not be exactly the same because the atomicity will not be perfect, and there will be some negativity and peaks of varying height even for an equal-atom structure. The number of vectors is much greater than the number of peaks, and the obvious and efficient way to obtain this correspondence of the inter-peak vectors is if the main peaks of the trial map tend to form structural fragments similar to those of the true structure and correctly oriented, albeit in incorrect positions and not correctly linked to each other. If a large correct fragment exists and could be selected, then this can be the basis for elucidating the complete structure. These are the thought processes which are guiding our present investigations.

The programs described in this paper were implemented on Vax and Pentium machines and are completely portable. For further information please e-mail sspmm@iacs.ernet.in.

References

Bezborodova, S. I., Ermekbaeva, I. A., Shlyapnikov, S. V., Polyakov, K. M. & Bezborodov, A. M. (1988). *Biokhimiya*, **53**, 965–973.

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

Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.

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 & Louvain, Belgium.

Fan, H.-F. & Woolfson, M. M. (1991). *Z. Kristallogr.* **197**, 197–208.
Hauptman, H. (1995). *Acta Cryst.* B**51**, 416–422.
Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* A**34**, 863–870.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Mishnev, A. F. & Woolfson, M. M. (1994). *Acta Cryst.* D**50**, 824–846.
Mukherjee, M. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 9–12.
Mukherjee, M. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 626–628.
Sheldrick, G. M. (1993). *SHELXL*93. *Program for the Refinement of Crystal Structures.* University of Göttingen, Germany.
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.
Weeks, C. M., DeTitta, G. T., Miller, R., Hauptman, H. A. (1993). *Acta Cryst.* D**49**, 179–181.
Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst.* A**46**, 41–46.