Acta Cryst. (1993). D**49**, 9–12

On the Application of Phase Relationships to Complex Structures. XXXII. A Small Protein at Low Resolution

By M. Mukherjee* and M. M. Woolfson

*Department of Physics, University of York, York YO*1 *5DD, England*

(*Received* 17 *May* 1992; *accepted* 11 *September* 1992)

Abstract

The direct-methods program *SAYTAN* is applied to data at various restricted resolutions for a small protein. It is shown that useful sets of phases can be obtained even down to 3 Å resolution. Conventional figures of merit are not very discriminating for the phase sets developed, but modified figures of merit seem capable of selecting the better phase sets, at least for those generated from 2 Å or higher resolution data.

Introduction

It was shown by Woolfson & Yao (1990) that a straightforward application of the direct-methods program *SAYTAN* could solve the protein, avian pancreatic polypeptide (aPP) (Glover *et al.*, 1983). This protein was not only small (36 amino-acid peptide plus Zn plus 80 H_2O in the asymmetric unit, space group $C2$ with $a = 34.18$, $b = 32.92$, $c = 28.44$ Å, $\beta = 105.3°$) but also contained a fairly heavy atom and had data to 0.98 Å resolution. The difficulty of solving a protein structure obviously becomes greater when the protein is of larger size and contains no heavy atoms, and when the data is of lower resolution. Here we examine the last of these conditions – that of lower resolution. In order to appreciate clearly the influence of changing the resolution, and avoid the confusion caused by different structures and data sets, we have used artificially truncated aPP data.

Application of *SAYTAN*

In all our computer experiments we started with pseudo-random phases generated by a magic-integer series (White & Woolfson, 1975). At each resolution 1000 trials were used and, as proposed by Woolfson & Yao (1990), the phases of the 50 largest E's were kept fixed until the last cycle of refinement when they were allowed to relax to fit in with the other phase values. The version of *SAYTAN* which was used was that for which only quartet terms corresponding to small E's are used (Debaerdemaeker, Tate & Woolfson, 1988). The use of a tangent-formula weighting scheme as given by Hull & Irwin (1978) makes the refinement quite stable.

For resolutions lower than 2 Å, refinement of initially random phase sets by *SAYTAN* alone does not give a mean phase error (MPE) less than 70°. For these resolutions we found it beneficial to begin with five cycles of parameter-shift refinement as described by Debaerdemaeker & Woolfson (1983). In this application phases are changed one at a time by ±45° and tested against minimization of the function

$$S = \sum_{\mathbf{h}} |E(\mathbf{h}) - (T_\alpha/Q_t)\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|^2 \qquad (1)$$

where T_α is the sum of all the triple-phase invariants with the current phases and Q_t is a theoretically derived value for the sum of all the quartets (Debaerdemaeker, Tate & Woolfson, 1988). The phase which is accepted at each step is that corresponding to the shift of +45, 0 or −45° which gives the lowest value of S. After the parameter-shift refinement *SAYTAN* is used in the normal way.

Results

In Table 1 we show the results of our experiments giving the MPE's for various resolutions. Without introducing the parameter-shift cycles the MPE's at resolution 2 Å and higher were all >70° so the benefit of front-ending *SAYTAN* with parameter shift is evident. Maps with MPE's in the 65–70° range are sometimes adequate for the approximate fitting of models, although it is also possible to improve the maps by various methods (*e.g.* Wang, 1985; Zhang & Main, 1990*a*,*b*; Shiono & Woolfson, 1992).

These results show that it is probably better to start with data of the highest possible resolution from the outset rather than adopt the strategy of first aiming for a low-resolution phase set and then entering a phase-extension process. It is almost as much work to derive the 15 sets of phases at 3 Å resolution with MPE's ≃ 69° as to find the 11 sets at 1 Å

[*] Present address: Indian Association for the Cultivation of Science, Jadavpur, Calcutta 700 032, India.

0907-4449/93/010009-04$06.00

Table 1. *A summary of results in applying SAYTAN to aPP data at different resolutions*

NREF is the number of reflections used, $E_{\min}$ is the minimum E used, NREL is the number of linking three-phase relationships and MPE is the mean phase error.

Resolution (Å)	NREF	$E_{\min}$	NREL	Refinement process	Result [MPE (°)]	Minimum MPE (°)
1.0	800	1.7	9726	*SAYTAN*	11 sets [~40]	38
1.5	650	1.4	11620	*SAYTAN*	29 sets [~50]	48
1.77	556	1.3	14217	*SAYTAN*	16 sets [~55]	54
2.0	600	1.0	26323	*SAYTAN*	30 sets [~64]	62
2.25	350	1.2	6809	Parameter shift and *SAYTAN*	12 sets [~65]	63
2.5	300	1.14	5841	Parameter shift and *SAYTAN*	8 sets [~69]	68
3.0	315	0.9	9841	Parameter shift and *SAYTAN*	15 sets [~69]	69

resolution with MPE's ≃ 40°. However, as was shown by Woolfson & Yao (1988, 1990), phase extension works quite well with *SAYTAN*, either in going from low to high resolution or, alternatively, increasing the number of phases determined within a fixed resolution.

Improved figures of merit

In presenting their results Woolfson & Yao (1990) remarked that they were only able to know that they had obtained good phase sets for aPP because it was a known structure. This was actually a rather pessimistic statement; if the figures of merit were examined carefully then those corresponding to the best sets of phases were distinguishable but the problem was that their values were not those usually associated with a good phase set.

The first conventional figure of merit we consider is:

$$\mathrm{ABSFOM} = \frac{\sum_{\mathbf{h}} \alpha(\mathbf{h}) - \sum_{\mathbf{h}} \alpha(\mathbf{h})_{\mathrm{ran}}}{\sum_{\mathbf{h}} \alpha(\mathbf{h})_{\mathrm{exp}} - \sum_{\mathbf{h}} \alpha(\mathbf{h})_{\mathrm{ran}}} \qquad (2)$$

where $\alpha(\mathbf{h}) = |\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|$ with the current phases and the subscripts ran and exp correspond to theoretical values with random phases and true phases respectively. For a good set of phases the expectation value of ABSFOM is 1.0 and for small structures values of 0.9–1.2 are usually found. However, if we take the 2.0 Å run for aPP it is found that the best phase sets, with MPE's between 62 and 64°, have ABSFOM greater than 3.0 and even completely wrong phase sets have ABSFOM's of 2.4–2.8. The reason for this is not hard to find. For large structures individual triple-phase relationships have large variances and while the distributions of their values are not random, the values of the two terms in the divisor of (2) may be similar; they differ by a factor of two for the 2 Å aPP example. Alternatively, the very process of refining phases with a tangent formula, even *SAYTAN*, tends to create phase sets which satisfy the three-phase relationships too well – and in the case of proteins far too well since the relationships hold so poorly with correct phases.

In order better to appreciate the significance of the degree of oversatisfaction of the relationships we have found a simpler figure of merit to substitute for ABSFOM which eliminates the random component. This is

$$\mathrm{ABSM} = \sum_{\mathbf{h}} \alpha(\mathbf{h}) / \sum_{\mathbf{h}} \alpha(\mathbf{h})_{\mathrm{exp}}. \qquad (3)$$

This figure of merit, which should have a value of 1.0 for correct phases, has values just over 2.2 for the 2 Å best solutions for aPP. Although the minimum value of ABSM for the 1000 generated phase sets is 1.80, nevertheless this narrow range of values does identify the good sets quite reliably.

The next conventional figure of merit is that which depends on small E's:

$$\mathrm{PSIZERO} = \sum_{\mathbf{l}} |\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{l}-\mathbf{k})| / \sum_{\mathbf{l}} [\sum_{\mathbf{k}} |E(\mathbf{k})E(\mathbf{l}-\mathbf{k})|^2]^{1/2} \qquad (4)$$

where the summation over $\mathbf{k}$ is for the large E's whose phases are being determined while the summation over $\mathbf{l}$ is for small E's. A small numerator for this expression indicates that Sayre's equation is holding for the small E's while the divisor is an expectation value of the numerator if random phases are used. A good set of phases usually has a small value of PSIZERO, values between about 1.0 and 1.6 being normal. For the 2 Å phase sets for aPP better solutions give PSIZERO somewhat less than 0.8 but some incorrect phase sets give 0.02 while other incorrect sets give about 0.9. We have found that the following expression gives a more discriminating figure of merit

$$\mathrm{PSIM} = \sum_{\mathbf{l}} |\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{l}-\mathbf{k})| / \sum_{\mathbf{h}} \alpha(\mathbf{h}) \qquad (5)$$

where the summation over $\mathbf{h}$ is for the reflections being phased. The rationale here is that when the

Table 2. *A selection of figures of merit for various phase sets generated for aPP at* 1.5 Å *resolution*

For comparison the figures of merit are also given for phases derived by calculation from the final refined structure (indicated as 'true').

Set number	ABSM	PSIM	RESM	CFOM	D (°)	MPE (°)
9	1.86	0.50	30.2	2.53	26.7	52.7
100	1.60	0.49	34.5	1.36	33.0	81.2
102	1.83	0.51	30.3	2.17	28.9	52.4
130	1.35	0.53	37.2	0.12	37.1	82.5
132	1.82	0.50	30.6	2.21	28.9	52.5
181	1.51	0.52	36.1	0.64	35.7	84.4
185	1.63	0.42	34.1	1.97	34.9	80.0
186	1.84	0.50	30.0	2.32	29.2	52.0
200	1.57	0.52	34.8	0.98	34.6	79.5
201	1.49	0.55	36.9	0.32	32.8	81.6
255	1.85	0.51	31.0	2.13	27.0	51.7
True	1.64	0.41	18.8		35.5	

values of $\alpha(\mathbf{h})$ are large then it is normally found that a pattern of phases is established which tends to give a larger value for the numerator of (5). By looking for smaller values of PSIM we are discriminating against small values of the numerator which arise just because the phases are fairly random and the divisor is small. It must be said that this argument is based on experience gained over many years of developing direct methods and cannot be justified analytically but the experience does seem to point in the right direction in this case.

The better phase sets tend to have lower values of PSIM although, as will be seen from Table 2, this is not universally true. Once again, although the values found for the new figure of merit are somewhat constrained they do help to indicate correctly the better phase sets.

The final conventional figure of merit is

$$\mathrm{RESID} = \frac{\sum_{\mathbf{h}} |\alpha(\mathbf{h}) - \alpha(\mathbf{h})_{\mathrm{exp}}|}{\sum_{\mathbf{h}} \alpha(\mathbf{h})_{\mathrm{exp}}} \times 100 \qquad (6)$$

which depends on how well individual values of α agree with their expectation values. Because the values of $\alpha(\mathbf{h})$ derived from *SAYTAN* are much larger than their theoretical expected values very large values of RESID are found – for example, greater than 100 for the better phase sets and between 80 and 140 in general. We have compensated for the abnormally large values of $\alpha(\mathbf{h})$ by using

$$\mathrm{RESM} = \sum_{\mathbf{h}} \left| \frac{\alpha(\mathbf{h})}{s} - \frac{\alpha(\mathbf{h})_{\mathrm{exp}}}{s_{\mathrm{exp}}} \right| \times 100 \qquad (7)$$

instead of RESID. Here $s = \sum_{\mathbf{h}} \alpha(\mathbf{h})$ and s_{exp} is the expected value of s with true phases.

The values of RESM are not scaled to give an expectation value valid for all structures but, for the 2 Å phase sets we are considering here, the values vary between 44.9 and 51.7 with better phase sets having values from 44.9 to 45.3.

Another figure of merit which was computed (although not used in the automatic selection of the better phase sets) was

$$D = \langle \min(|\Phi_{3,i}|, |180 - \Phi_{3,i}|) \rangle_i \qquad (8)$$

where $\Phi_{3,i}$ is the value in degrees of the ith three-phase invariant, which was given by Woolfson & Yao (1990) to distinguish sets of phases which give enantiomorph discrimination. The average is over the values of all the three-phase invariants; if its value is small then the invariants are all close to 0 or π and the enantiomorph will be poorly indicated. As a rule of thumb any value of D over 15° represents satisfactory enantiomorph discrimination.

A selection of results for 1.5 Å, including some better solutions, is shown in Table 2. The combined figure of merit CFOM is defined in the usual *MULTAN* and *SAYTAN* way

$$\begin{aligned} \mathrm{CFOM} = {} & w_1 \frac{\mathrm{ABSM} - (\mathrm{ABSM})_{\min}}{(\mathrm{ABSM})_{\max} - (\mathrm{ABSM})_{\min}} \\ & + w_2 \frac{(\mathrm{PSIM})_{\max} - \mathrm{PSIM}}{(\mathrm{PSIM})_{\max} - (\mathrm{PSIM})_{\min}} \\ & + w_3 \frac{(\mathrm{RESM})_{\max} - \mathrm{RESM}}{(\mathrm{RESM})_{\max} - (\mathrm{RESM})_{\min}} \end{aligned} \qquad (9)$$

where the subscripts max and min correspond to the maximum and minimum values for the 1000 phase sets and the weights are set at $w_1 = w_2 = w_3 = 1.0$. A phase set with the best value for all three figures of merit would have CFOM = 3.0.

We also show in Table 3 some extracts from results obtained at 2 and 2.25 Å; for 2 Å the values of CFOM close to 2.0 indicate several sets with mean phase errors between 61 and 63°. The values of D for these sets are all close to 20° indicating that the map should show good enantiomorph discrimination. For 2.25 Å resolution the value of CFOM does not clearly show the better solutions and, indeed, the highest value of CFOM is for a set of phases with a high MPE. It appears that the value of ABSM alone could pick up the best phase sets in this case, suggesting that a different weighting scheme for combining FOM's should have been used.

Table 3. *A selection of figures of merit for various phase sets generated for aPP at* 2.0 *and* 2.25 Å *resolution*

The values with 'true' phases (see Table 2) are also given.

Set number	ABSM	PSIM	RESM	CFOM	MPE (°)	*D* (°)
2 Å						
502	2.08	0.39	49.4	1.99	61.4	20.1
505	0.25	0.16	79.9	0.98	83.9	14.3
506	2.09	0.39	49.3	2.01	62.8	19.2
510	2.08	0.39	49.1	1.99	62.5	20.1
549	2.09	0.40	48.9	1.99	63.0	20.4
600	0.31	0.23	72.7	0.94	84.2	9.5
601	0.34	0.16	54.6	1.86	80.2	14.7
609	2.08	0.38	49.2	1.99	62.6	20.1
700	2.09	0.38	48.7	1.99	61.7	20.4
754	2.08	0.38	49.4	2.00	61.8	19.3
True	1.44	0.31	29.6			42.1
2.25 Å						
150	0.36	0.25	42.6	1.45	81.8	14.0
152	0.36	0.23	46.1	1.08	79.8	16.8
155	0.38	0.26	40.5	1.68	81.3	14.8
159	2.02	0.46	42.1	1.78	62.7	21.3
525	0.32	0.30	47.9	0.56	84.9	14.8
600	0.39	0.17	45.1	1.41	83.2	19.8
715	0.42	0.17	44.3	1.54	83.8	17.0
811	0.48	0.20	41.3	1.89	83.2	15.1
819	1.98	0.44	42.6	1.81	65.4	21.0
900	0.40	0.20	43.5	1.53	83.8	18.9
True	1.76	0.33	34.2			43.0

Concluding remarks

We have shown that by a judicious choice of figures of merit it is possible to develop and recognize phase sets with acceptable phase errors for a small protein structure at moderate resolutions. While we have demonstrated this down to 2 Å resolution it must be said that recognizing better phase sets for lower resolution has eluded us. This is a very stringent limit on the applicability of this kind of direct method as many proteins give data limited in resolution to 2.5 Å or even lower.

We are beginning experiments on the second limiting factor, the size of the structure, and we hope to report on this work in due course. However, one thing is already clear: conventional direct methods which operate with the *MULTAN/SAYTAN* philosophy have only a very limited contribution to make to protein crystallography and new ideas, perhaps coupled to the use of real-space methods and physical data such as that from anomalous scattering, are needed to make further progress.

We express our gratitude to the Science and Engineering Research Council who supported this work. Our thanks are also due to Drs Yao Jia-xing, C. Tate and L. Refaat for their help and advice.

References

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1988). *Acta Cryst.* **A44**, 353–357.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* **A39**, 193–196.

Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, T., Tickle, I. & Blundell, T. (1983). *Biopolymers*, **22**, 293–304.

Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* **A34**, 863–870.

Shiono, M. & Woolfson, M. M. (1992). *Acta Cryst.* **A48**, 451–456.

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

White, P. & Woolfson, M. M. (1975). *Acta Cryst.* **A31**, 53–56.

Woolfson, M. M. & Yao Jia-xing (1988). *Acta Cryst.* **A44**, 410–413.

Woolfson, M. M. & Yao Jia-xing (1990). *Acta Cryst.* **A46**, 409–413.

Zhang, K. Y.-J. & Main, P. (1990*a*). *Acta Cryst.* **A46**, 41–46.

Zhang, K. Y.-J. & Main, P. (1990*b*). *Acta Cryst.* **A46**, 377–381.