field can only enhance the level of service now provided to the crystallographic community. It is likely that fresh financial and legal problems will emerge as new distributional media are utilized, but continued vigilance by the oversight mechanism set in place since the Twelfth General Assembly may be expected to maintain the present robust state of IUCr crystallographic publishing and financial solvency.

References

ABRAHAMS, S. C. (1982). *Acta Cryst.* A38, 1-2.
ABRAHAMS, S. C. (1987). *Acta Cryst.* A43, C300.
AMERICAN MATHEMATICAL SOCIETY (1983). *Not. Am. Math. Soc.* 30, 715-724.
BARSCHALL, H. H. (1986). *Physics Today,* December, pp. 34-36.
Communicated Abstracts of the Fourteenth International Congress of Crystallography (1987). *Acta Cryst.* A43, C1-C351.
Crystallographic Databases (1987). Bonn, Cambridge, Chester: IUCr Data Commission.
EWALD, P. P. (1948). *Acta Cryst.* 1, 1-2.
HAWKINS, D. T. (1980). *Acta Cryst.* A36, 475-482.
INTERNATIONAL UNION OF CRYSTALLOGRAPHY (1948). Statutes and By-Laws. *Acta Cryst.* 1, 275-276.
INTERNATIONAL UNION OF CRYSTALLOGRAPHY (1983). Appendices to the Agenda of the Twelfth General Assembly. *Acta Cryst.* A39, 444-465.
INTERNATIONAL UNION OF CRYSTALLOGRAPHY (1986). Report of the Executive Committee for 1984. *Acta Cryst.* A42, 58-79.
JUDD, P. (1983). *Bull. Am. Phys. Soc.* 28, 957 (B1-B57).
Notes for Authors (1983). *Acta Cryst.* A39, 174-186.
PEPINSKY, R. (1952). Editor. *Computing Methods and the Phase Problem in X-ray Crystal Analysis.* Pennsylvania State College: X-ray Crystal Analysis Laboratory.
PEPINSKY, R., ROBERTSON, J. M. & SPEAKMAN, J. C. (1961). Editors. *Computing Methods and the Phase Problem in X-ray Crystal Analysis.* Pergamon: Oxford.

On the Application of Phase Relationships to Complex Structures. XXVII. Phase Extension for Small Proteins

BY M. M. WOOLFSON

Department of Physics, University of York, York YO1 5DD, England

AND YAO JIA-XING

Institute of Physics, Beijing, China

Abstract

An example is given of phase extension for a small protein, avian pancreatic polypeptide, by the use of the Sayre-equation tangent formula (SETF). Initial data were the phase estimates of 129 reflexions with large values of $|E|$ within the 3 Å resolution sphere. The mean error of these phases, estimated by a combination of isomorphous replacement and anomalous scattering, was 26·5°. Random values were then given to 1371 other phases out to 1 Å resolution and refinement was carried out with SETF. In 20 trials, 11 gave mean phase errors less than 34° for all 1500 reflexions with the best set having a mean phase error of 31·9°. Maps computed with these phases showed the general form of the molecule.

Introduction

Much thought is being given to the application of direct methods to protein-structure problems. The most successful ideas so far have involved some combination of information from physical methods, such as the heavy-atom method, isomorphous replacement or anomalous scattering, with direct-methods theory (*e.g.* Hauptman, 1982; Karle, 1984, 1986; Fan Hai-fu, 1983). However, there is a tendency to illustrate ideas with ideal calculated data for known structures and, until recently, the only unknown protein structures to have been solved using direct methods have involved starting from heavy-atom positions found by using *MULTAN* (Wilson, 1978). However, there is an interesting example of the application of

anomalous-scattering data with direct methods on the solution of Cd, Zn metallothionein (Furey, Robbins, Clancy, Winge, Wang & Stout, 1986).

A critical step in many protein structure solutions is that of phase extension from phases estimated at low resolution from physical methods – very often with very low accuracy (root mean square error typically ~45°). A large number of procedures for phase extension and refinement have been reported. At one extreme there are theoretical reciprocal-space direct-methods approaches such as that of Tsoucaris (1970) using the maximum-determinant method and Bricogne (1984) using the ideas of maximum entropy. There are also real-space methods of density modification, the best known at present being that of Wang (1984) who uses density flattening outside an envelope identified as containing the protein molecule. At the other extreme there are density-fitting methods where a model is fitted to the low-resolution density on a trial-and-error basis (Agarwal & Isaacs, 1977).

An early approach of the direct-methods variety was that of Sayre (1972, 1974) who extended and refined phases on the basis that true phases should satisfy a system of Sayre equations. The method was a brute-force one requiring prodigious use of computer time although the effort required might seem more reasonable with the present availability of supercomputers.

The characteristic of all these methods is that they are normally of the single-solution variety. From the starting point an objective procedure is followed which leads to the final result. If that result is unsatisfactory then one might, in an *ad hoc* way, modify the procedure to get a different answer, but rarely does this seem to be done in practice. Our experience with direct methods applied to small structures suggests that single-solution methods are unlikely to be successful in general. The strength of modern direct methods resides in their multi-solution nature, in that for difficult structures one may simply increase the number of trials to make the probability of netting the correct solution reasonably large.

An illustration of this is the fragment development procedure of Yao Jia-xing (1983). A time-honoured and successful way of developing a complete structure from a fragment is that given by Karle (1968). The fragment from an E map is used to generate reliable phase estimates for a relatively small number of reflexions, these are inserted into a cyclic process of phase development using the tangent formula, and the newly estimated phases are used to compute a new E map. Starting with a small fragment the pattern is that a somewhat larger fragment is obtained, which can then be used as a new starting point; repeating this process eventually gives the complete structure.

By contrast Yao Jia-xing's (1983) approach is to take the same set of reliable phase estimates from the initial fragment and to assign random phase values to all other reflexions whose phases are required. In a small number of trials the *complete* structure is usually obtained in one stage. More importantly, examples are found where the Karle procedure is ineffective but the Yao Jia-xing (1983) method will work – albeit not in one stage.

Fragment development is a process which builds from partial to complete information and this also applies to phase extension and refinement. Based on this kind of background and experience we have made trials of phase extension using a multi-solution direct-methods procedure.

Phase extension by *SAYTAN*

Debaerdemaeker, Tate & Woolfson (1988) have reported the development of the *SAYTAN* procedure, based on the Sayre-equation tangent formula (SETF), which is incorporated in *MULTAN*87. The SETF is a phase developing formula giving phases tending to satisfy a set of Sayre equations. In view of the work by Sayre (1972, 1974) it seemed to be sensible to try the SETF to develop phases; the SETF is easy and fast to apply and it is possible to make many trials.

The test structure we used was aPP (avian pancreatic polypeptide) solved by Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983) to 1 Å resolution. The details of the crystal are:

asymmetric unit contents:
36 amino-acid peptide + Zn + 80H$_2$O;
space group $C2$, $Z = 4$;
$a = 34 \cdot 18$, $b = 32 \cdot 92$, $c = 28 \cdot 44$ Å, $\beta = 105 \cdot 30°$.

It is a favourable case to take since, although there are 382 non-hydrogen atoms in the asymmetric unit, the data extend out to a resolution of about 1 Å. In fact phase extension from 2·04 to 1·37 Å was carried out for this structure (Tickle, 1981) using the tangent formula with a weighting scheme suggested by Hull & Irwin (1978). Starting with phases from the solved structure at 2·04 Å, obtained by isomorphous replacement and anomalous scattering measurements from an HgCl$_2$ derivative, Tickle (1981) used a controlled application of the weighted tangent formula and, in his paper, presented weights for the phase estimates in the region of phase extension but no information about phase errors.

For our part we have started with phase information at much lower resolution (3 Å), extended to the observed limit of resolution (1 Å) and give mean phase errors ($|\langle \Delta \varphi \rangle|$) for our various trials. The steps in the process, which is quite automatic, are as follows:

(1) Calculate normalized structure factors using the Wilson-plot method.

Table 1. *Figures of merit and mean phase error for 20 trials of phase extension for aPP*

Set	ABSFOM	PS10	RESID	CFOM	ERROR
1	1·341	2·823	35·50	0·966	32·1
2	1·070	4·161	31·27	1·995	32·6
3	0·947	4·645	29·07	2·546	32·3
4	0·788	5·405	30·01	2·200	32·9
5	0·880	4·881	28·51	2·671	31·9
6	1·365	2·398	36·79	0·638	41·8
7	1·054	2·926	37·77	0·261	84·0
8	0·888	4·825	28·42	2·700	33·0
9	= Set 8				
10	1·087	2·431	35·54	0·918	88·1
11	1·374	2·258	36·37	0·766	33·7
12	0·951	4·584	28·80	2·624	32·4
13	0·598	2·141	32·07	1·783	78·1
14	1·131	3·904	32·17	1·774	36·6
15	1·111	4·007	32·13	1·774	32·6
16	0·875	4·948	28·75	2·598	36·0
17	0·846	5·026	28·77	2·581	32·1
18	= Set 16				
19	1·432	2·355	37·29	0·521	32·9
20	1·150	3·903	33·38	1·442	50·5

(2) Select 1500 of the largest ($|E| \geq 1\cdot548$) and 100 smallest E values from the 16 538 reflexions within 1 Å resolution data.

(3) Set up a $\sum_2$ list of 64 237 phase relationships linking the large E values and 12 862 contributors to the right-hand side of Sayre's equation for the small E's.

(4) For the 129 large-E reflexions within 3 Å resolution assign the phases which were found in the original solution of the structure from a combination of single isomorphous replacement and anomalous scattering. The mean phase error of these assignments is 26·5°.

(5) To the 1371 other large-E-value reflexions assign 'random' phases generated by a magic-integer algorithm (Main, 1978).

(6) Set weights $W = 1$ for the 129 'known' phases, $W = 0\cdot25$ for the 1371 random phases.

(7) Refine phases by the SETF keeping the 129 known phases fixed. A new phase estimate is accepted when its calculated weight is greater than a cut-off value which is decreased cycle-by-cycle. The cut-off weight is $0\cdot8 \times (0\cdot85)^{n-1}$ where n is the cycle number.

(8) Repeat the whole process for 20 trials.

(9) Output an E map for the phase set with the best combined figure of merit (CFOM) and also a list of coordinates. Since the structure was known it was also possible to determine a mean phase error for each of the trials for all 1500 large-E reflexions.

The total run took about 20 min on a VAX8650/8550 cluster mainframe.

Results

In Table 1 the conventional figures of merit and also the mean phase error for each of the trials are shown. There were 11 phase sets with mean phase error less than 34°. The best set, with a mean phase error of 31·9°, was second best in order of CFOM. For set 8, that with the highest CFOM, 86 of the top 100 peaks identified by the routine SEARCH were within 0·4 Å of a true atomic position; for set 5, that with the lowest mean error, the corresponding figure was 85 atoms in the top 100 peaks – slightly worse. For structures of this size, however, it is probably better to look at the E-map density rather than interpreted peaks. A graphical display of the molecule superimposed on a density map gives confidence that a complete detailed structure solution would be possible starting from either set 8 or set 5.

Concluding remarks

The results with aPP are encouraging but it is too early to say that a routine procedure has been found for dealing with the solution of small proteins – especially as aPP is so favourable in terms of the quality of its data.

We can think of possible improvements in our procedure. For one thing the number of small E's used for aPP was far from optimal; Debaerdemaeker et al. (1988) have found that the best ratio of the number of contributors to small E's to the number of $\sum_2$ relationships for large E's is about 0·5 rather than the 0·2 we have used. Another possible improvement is to use much more of the information within the 3 Å region. We selected the phase estimates for the 129 largest E's without considering their reliability and ignored other available information from smaller (but not much smaller) E's with, perhaps, more reliable phase estimates.

Finally, in computing the E map we could have used more than 1500 terms by using the phase estimates for the 1500 to estimate phases for the extra 4678 reflexions down to $|E| \geq 1$.

Clearly there is far to go before any substantial success can be claimed – but the initial indications are good and there is much still to try.

References

AGARWAL, R. C. & ISAACS, N. W. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2835-2839.

BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410-445.

DEBAERDEMAEKER, T., TATE, C. & WOOLFSON, M. M. (1988). *Acta Cryst.* A**44**, 353-357.

FAN HAI-FU (1983). In Proc. IUCr Summer School on Crystallographic Computing, Kyoto, Japan.

FUREY, W. F., ROBBINS, A. H., CLANCY, L. L., WINGE, D. R., WANG, B. C. & STOUT, C. D. (1986). *Science*, **231**, 704–710.

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. (1982). *Acta Cryst.* **A38**, 289–294.

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

KARLE, J. (1968). *Acta Cryst.* **B24**, 182–186.

KARLE, J. (1984). *Methods and Applications in Crystallographic Computing*, edited by S. R. HALL & T. ASHIDA, pp. 120–140. Oxford: Clarendon Press.

KARLE, J. (1986). *Acta Cryst.* **A42**, 246–253.

MAIN, P. (1978). *Acta Cryst.* **A34**, 31–38.

SAYRE, D. (1972). *Acta Cryst.* **A28**, 210–212.

SAYRE, D. (1974). *Acta Cryst.* **A30**, 180–184.

TICKLE, I. J. (1981). *Refinement of Protein Structures*. Proceedings of the Daresbury Study Weekend, 15–16 November 1980, edited by P. A. MACHIN, J. W. CAMPBELL & M. ELDER. Warrington: SERC Daresbury Laboratory.

TSOUCARIS, G. (1970). *Acta Cryst.* **A26**, 492–499.

WANG, B. C. (1984). *Acta Cryst.* **A40**, C12.

WILSON, K. S. (1978). *Acta Cryst.* **B34**, 1599–1608.

YAO JIA-XING (1983). *Acta Cryst.* **A39**, 35–37.

Exact Random-Walk Models in Crystallographic Statistics. IV. P.D.F.'s of $|E|$ Allowing for Atoms in Special Positions

BY URI SHMUELI

School of Chemistry, Tel Aviv University, Ramat Aviv, 69 978, Tel Aviv, Israel

AND GEORGE H. WEISS

Division of Computer Research and Technology, Laboratory of Physical Sciences, National Institutes of Health, Bethesda, Maryland 20892, USA

Abstract

The effect of scatterers, located in variable special positions, on the probability density function of the magnitude of the normalized structure factor has been investigated. Exact characteristic functions have been obtained for all the statistically different variable special positions in triclinic, monoclinic and orthorhombic space groups except in *Fdd*2 and in the space groups based on the point group 222, and the probability density functions have been evaluated from their Fourier or Fourier–Bessel series expansions. It is seen that the effect of heavy scatterers, located in the special positions investigated, is very marked and should be accounted for in cases of space-group ambiguities.

Introduction

The effects of the presence of atoms in special positions have been investigated in earlier work on intensity statistics (Karle & Hauptman, 1953; Hauptman & Karle, 1953; Collin, 1955; Hargreaves, 1956; Sim, 1958; Foster & Hargreaves, 1963; Ilyukhin & Nikitin, 1963), but no exact studies of these effects on the probability density function (p.d.f.) of the structure factor have so far been attempted. One of the reasons for avoiding the study of special positions is their very large number in all the 230 space groups,

and consequent apparent difficulties in arriving at tractable and reasonably concise formulae. Since, however, the qualitative effects of (heavy) scatterers in special positions on intensity statistics may well be of considerable significance in the determination of space-group symmetry in cases of ambiguities, a study of such effects was thought to be of interest.

The techniques used in this paper are similar to those we employed in our previous studies of intensity statistics, based on exact solutions of random-walk models (*e.g.* Shmueli, Weiss, Kiefer & Wilson, 1984; Shmueli & Weiss, 1987). Only low-symmetry space groups are treated, and it is seen that the number of different expressions that need to be developed is much smaller than the formal number of crystallographically different Wyckoff positions in the space groups investigated. The present treatment is confined to the variable special positions (*i.e.* lines and planes), since the contributions of scatterers located in fixed special positions can be calculated and subsequently subtracted from the (scaled) intensity; examples of the latter process can be found in the works of Collin (1955), Sim (1958), Srinivasan & Parthasarathy (1976) and Pradhan, Ghosh & Nigam (1985). The results presented in this paper encompass all the variable special positions in monoclinic and orthorhombic space groups, except those in space groups based on the point group 222 and in the space group *Fdd*2.