Acta Cryst. (1990). A**46**, 409-413

On the Application of Phase Relationships to Complex Structures. XXX. *Ab Initio* Solution of a Small Protein by *SAYTAN*

By M. M. Woolfson and Yao Jia-xing

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

(*Received* 23 *March* 1989; *accepted* 11 *December* 1989)

Abstract

An account is given of experiments with the program *SAYTAN* to determine directly the structure of a small protein, avian pancreatic polypeptide. It was found that by making a large number of trials with sets of initially random phases a straightforward run of *SAYTAN* would yield phases for about 800 reflexions with a mean phase error less than 40°. From this point it was possible to phase extend to 2000 reflexions using *SAYTAN* and then, by recycling the information from *E* maps, eventually to obtain 2000 or even more reflexions with a mean phase error of order 28°. Final cycles of weighted Fourier syntheses led to maps which could easily be interpreted with models so leading to a complete elucidation of the structure. It was found that standard figures of merit, such as those used in *MULTAN* and *SAYTAN* and one based on negative quartets, were not effective in recognizing good sets of phases for large structures. The full exploitation of direct methods to solve protein structures awaits the discovery of a new figure of merit more effective in ranking trial phase sets.

Introduction

In pursuit of the goal of applying direct methods to the solution of protein structures, Woolfson & Yao Jia-xing (1988) gave an example of the successful application of the Sayre-equation tangent formula, incorporated in the computer package *SAYTAN*, to phase extension for a small protein. The structure used for the demonstration was that of the known 36 amino acid globular hormone, avian pancreatic polypeptide (App), space group *C*2 with $a = 34{\cdot}18$, $b = 32{\cdot}92$, $c = 28{\cdot}44$ Å and $\beta = 105{\cdot}30°$ (Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell, 1983). Phases within 3 Å resolution had been found by a combination of single isomorphous replacement and anomalous scattering. We took as our starting point 129 of these phases for reflexions with large *E* values; they had a mean phase error, $\langle|\Delta\varphi|\rangle$, of 26·5°. By the use of a multisolution approach with *SAYTAN*, phase extension to a total of 1500 phases within 1 Å resolution was carried out and 11 out of 20 trials gave $\langle|\Delta\varphi|\rangle$ less than 34°. The best set had $\langle|\Delta\varphi|\rangle = 31{\cdot}9°$.

This success made us wonder whether, in fact, it was necessary to have the starting phases at all and whether the structure could be solved *ab initio* from the native protein data. Our exploratory efforts in that direction are reported here.

Initial phasing

Our first approach was simply to run the structure on *SAYTAN*, treating it as we would any small structure problem. The program was instructed to select the 800 largest *E*'s and 200 smallest ones, the latter playing a vital role in the *SAYTAN* procedure. There were 11 060 triplet relationships (t.r.'s) generated which linked the large *E*'s, excluding those which had a value of κ less than 0·1. Thus each t.r. was of the form

$$\varphi_3(\mathbf{h},\mathbf{k}) = \varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k})$$

with

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

where

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

In addition there were 7575 contributors to the small *E*'s, that is to say pairs of large *E*'s contributing to the summation

$$\sum E(\mathbf{k})E(\mathbf{l}-\mathbf{k})$$

where $\mathbf{l}$ is the index of a small *E*. After processing this information with the *CONVERGENCE* routine in *SAYTAN* a number of the initially chosen *E*'s were rejected by the program on the grounds that they were insufficiently well connected to the rest of the system. This left 727 large *E*'s interconnected by 9726 t.r.'s and 183 small *E*'s with 6434 contributors.

Starting with random phases we made 1000 trials refining the phases to convergence with *SAYTAN*. The program was run in a mode where large quartet terms are excluded, which means that the Sayre-equation condition for the large *E*'s is imposed less rigorously, but more economically, by the weighting scheme suggested by Hull & Irwin (1978). Since the structure is known we were able to check the value of $\langle|\Delta\varphi|\rangle$ for each set of phases and six of them were

0108-7673/90/050409-05$03.00

less than 45° with the lowest value 41·8°. We also applied another test which was to calculate the value of

$$c = \langle \min(|\varphi_{3,i}|, |180 - \varphi_{3,i}|)\rangle$$

where the average is over all the triple invariants, the ith one of which is $\varphi_{3,i}$ expressed in degrees. If c is small, say less than 15°, then this indicates that the corresponding E map will have a pseudo centre of symmetry but none of the six sets was of this type. Hence each of the six sets could be regarded as a very promising starting point for the eventual solution of the complete structure.

Although this first attempt was very encouraging it was a fairly marginal situation for App and suggested that there was not much power in hand to tackle even larger structures. We contrasted the six satisfactory solutions from 1000 trials in the *ab initio* case with the 14 out of 20 successes when phases were extended from a small number of known ones. The next experiment incorporated an element of the phase-extension process, which is to keep a number of phases fixed until the final one or two cycles of refinement. Some 50 to 100 of the original random phases were kept fixed until the penultimate cycle, in the hope that this would stop the tendency for perpetual drifting of the determined phase values and give much greater stability to the system. Despite this procedure being intuitive, rather than rigorously derived, it is, in fact, very successful. With the random phases of the 50 largest E's kept fixed until the last cycle of refinement (when they were allowed to relax) and with the same conditions otherwise, there were now 11 solutions with mean phase error less than 45° with the best set having a mean phase error of 38·0°. This result seemed quite stable to moderate changes in the parameters associated with the *SAYTAN* run, for example the weight associated with the small quartets through which the small E's influence the process. At this stage we decided that it should be possible to carry out phase extension and proceed to a full solution of the structure.

Phase extension

For the phase-extension trials the 727 large E's for which phases had been found were first ranked in order of the associated values of α (magnitudes of tangent-formula indications) which, at least theoretically, give a measure of the reliability of the determination of each individual phase. Then a new *SAYTAN* run was started, in which 2000 large E's and 200 small ones were included, and with the same minimum value of κ as previously: 0·1. 155 462 connecting t.r.'s were found for the large E's and 35 256 contributors to the small ones. After processing through the *CONVERGENCE* routine 1818 large E's and 199 small E's remained with 131 961 t.r.'s and 31 966 contributors respectively.

Table 1. *Effects of phase extension to* 1818 *reflexions starting with m 'known' phases selected from* 727 *phases found with SAYTAN*

Ten trials were made for each value of m.

m	Range of mean errors
300	40·0–48·0
400	40·8–42·8
500	41·8–44·6
600	42·3–43·0
700	42·7–43·0

The extension procedure involved taking the m top ranked of the 727 known phases and giving them the values found in the original *SAYTAN* run. The remaining $1818 - m$ reflexions were then given random values and *SAYTAN* refinement was carried out keeping the m reflexions fixed throughout the whole process. Based on our previous experience (Woolfson & Yao Jia-xing, 1988) we made only ten trials for each value of m but, as it turned out, a single run would probably have sufficed since all trials for each value of m gave little variation in the values of $\langle|\Delta\varphi|\rangle$ for the complete set of 1818 reflexions. The results from the phase extension are given in Table 1. It can be seen that the minimum value of $\langle|\Delta\varphi|\rangle$ was only weakly correlated with m, at least for the values of m in the range 300–700. Our conclusion is that the phase-extension procedure is very robust and that once a secure base of a few hundred phases is available then phase extension to 2000 phases may be carried out with confidence by *SAYTAN*.

We arbitrarily took the results of the tenth phase-extension trial for $m = 300$, which gave a mean phase error of 41·5°, to see how far we could go towards a complete and automatic structure determination. In Fig. 1 we show sections of the E map with the 1818 reflections for these phases, compared with the corresponding sections of the E map calculated with correct phases.

The processing of E maps

Using the determined phases for 1818 reflexions an E map was calculated and interpreted by the peak search program in *SAYTAN* to give 453 peaks. From our knowledge of the structure we were able to investigate the extent of the correspondence between the E-map peaks and actual atomic positions. We found that for the largest peaks 78 out of the top 100 were within 0·5 Å of a true atomic position and for greater numbers of peaks the ratios were 105/150 and 121/200. The progress in developing the structure was as follows.

Stage 1

The top 100 peaks were chosen as input to the Karle recycling procedure (Karle, 1968) which is a

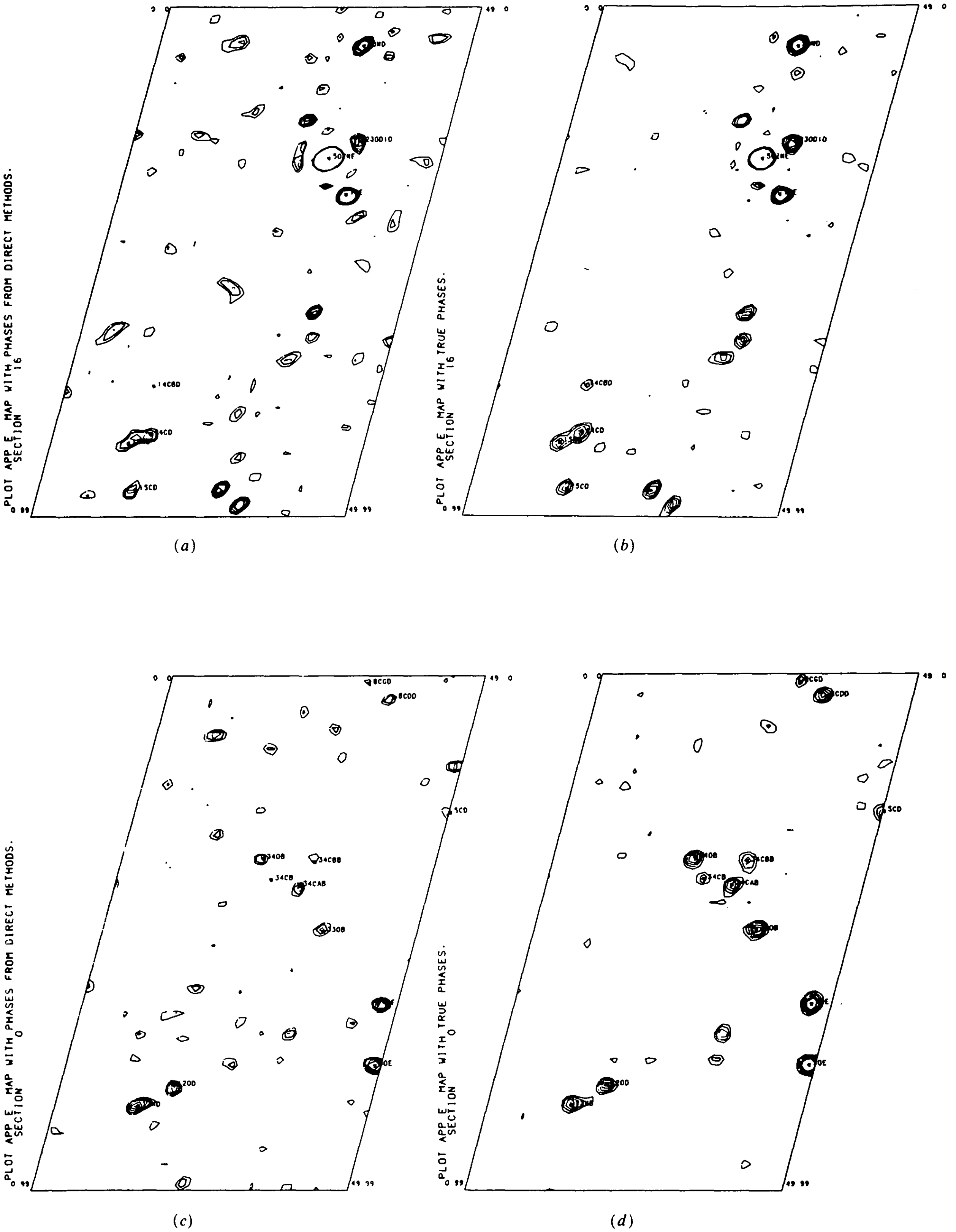

Fig. 1. A comparison of two sections of the E map calculated after the first stage of phase extension (PE) compared with that obtained with true phases. (a) PE for $y = 16/100$; (b) Correct-phase map for $y = 16/100$; (c) PE for $y = 0/100$; (d) Correct-phase map for $y = 0/100$.

component of *SAYTAN*. These 100 peaks, representing 0·32 of the complete structure, gave 1477 acceptable phase indications, with an R factor between the calculated partial structure factors and the observed E magnitudes of 40·9%. *SAYTAN* phase refinement and extension to 1893 phases then gave a mean phase error of 31·94° and led to an E map in which 87/100 and 142/200 top peaks could be associated with true atomic positions.

We also tried the procedure with other parameters. For example, if the 200 highest peaks were originally used for Karle recycling then 1567 acceptable phase indications were found; the R factor was 38·4% and the resultant E map after refinement and extension gave 94/100 and 141/200 top peaks associated with atomic positions. One further stage of Karle recycling, this time using only the top 100 peaks as the partial structure, led to 1900 phases with a mean error of 27·7° and an E map giving 95/100 and 160/200 top peaks near atomic positions.

Again, instead of carrying out the phase extension to approximately 2000 reflexions we extended to approximately 3000 and also to 4000 reflexions. The mean phase errors were a few degrees higher for these greater phase extensions but they may have had a higher information content. However, the advantage that these greater extensions might have offered did not seem to justify the extra effort of handling more than 500 000 t.r.'s.

Stage 2

From our latest E map the top 200 peaks were selected, of which 160 were approximately in atomic positions, and a weighted Fourier map was calculated with all 16 538 independent reflexions within the sphere of observation. This map showed 100/100, 180/200, 214/250, 230/270 and 244/370 correctly positioned peaks. If the top 270 of these peaks were inserted into the weighted Fourier map procedure then the outcome, in the same terms, was 100/100, 193/200, 230/250, 236/270 and 254/370. At this stage an electron-density map calculated with the available phases looked very similar to the completely correct map, and a relatively simple stage of model fitting is enough to complete the structure determination.

Concluding remarks

We have shown that, in principle, it should be possible to solve directly the structure of a small protein, for which there are good data, by means of *SAYTAN*. The program is used in four different processes:

(i) Derive a basis set of a few hundred phases by a routine run of the program making many trials. This process is made far more effective and stable by keeping a number of the initially random phases fixed for all but the last cycle or two of phase refinement.

(ii) For each plausible phase set select a few hundred of the most reliably indicated phases, use these as a base to extend to two thousand phases or so.

(iii) From the resultant E maps select 100–200 of the highest peaks as a means of improving phases by the Karle recycling procedure. This stage can be repeated until it is judged that convergence has been achieved.

(iv) The top 200 or so highest peaks from the final E map can be used to estimate phases for the complete set of observed structure factors and used to compute a weighted Fourier map. This stage can be repeated until there is no apparent improvement.

An important element leading to the fairly short timescale for the development work reported here has been the computing environment chosen. We have used a standard personal computer (PC) incorporating a 4 Mbyte transputer, a combination which is inexpensive and offers one third of the speed of a mainframe VAX 8650 cluster. Each of the experiments involving 1000 trials took about 15 h (an overnight run) on the transputer system, but the equivalent 5 h would not have been so easily available on a mainframe shared by many hundreds of users. It seems to us that the way ahead for most crystallographic computing is *via* dedicated systems used intensively and replaced at intervals as the technology advances.

There is, however, a negative side to this report. We were only able to recognize phase sets with small phase errors because we were dealing with a known structure. It turns out that all the *MULTAN*/*SAYTAN* standard figures of merit are just not discriminating enough for structures of this size. We also tried a figure of merit exclusive to *SAYTAN*, the Sayre figure of merit (SFOM), which measures how well Sayre's equation is satisfied for the complete set of E's (both large and small), and *NQEST* (De Titta, Edmonds, Langs & Hauptman, 1975), but neither was effective. Our next, and most important, objective will be to find a figure of merit which is reliable for large structures. It may be too much to hope that we should be able to pick out the best phase set for such structures by means of our figure of merit, but at least we might expect it to be close to the top of the plausibility ranking order. Modern graphics workstations should then enable a visual examination, with model fitting, to pick out the correct electron-density map and so successfully complete the structure solution.

On the assumption that the figure-of-merit problem can be solved we are hoping to extend this work towards somewhat larger structures at high resolution and perhaps even to very large structures at low resolution where Sayre's equation still holds quite well.

We wish to express our gratitude to the Wellcome Trust for the provision to one of us (YJ-x) of a fellowship and to the SERC for its continued support of direct methods development at York.

References

De Titta, G. T., Edmonds, J. W., Langs, D. A. & Hauptman, H. (1975). *Acta Cryst.* **A31**, 472-479.

Glover, I., Haneef, I., Pitts, J.-E., Wood; S. P., Moss, D., Tickle, I. J. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293-304.

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

Karle, J. (1968). *Acta Cryst.* **B24**, 182-186.

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

Acta Cryst. (1990). **A46**, 413-422

A Study of the Antisymmetric and Symmetric Parts of the Anharmonic Vibration in Zinc using Synchrotron Radiation

By Gabriela Kumpat and Elisabeth Rossmanith

Mineralogisch-Petrographisches Institut der Universität Hamburg, D-2000 Hamburg, Grindelallee 48, Federal Republic of Germany

(*Received* 30 *October* 1989; *accepted* 11 *December* 1989)

Abstract

The 'almost forbidden' Bragg reflection 061 and the very weak Bragg reflection 0,1,12 of a Zn single-crystal sphere have been carefully analysed to study the antisymmetric and symmetric features of vibrational anharmonicity. The intensity measurements were carried out at room temperature using synchrotron radiation with $\lambda = 0{\cdot}7100\,(3)$ Å taking into account multiple-beam effects. Data recorded on the Hamburger Synchrotronstrahlungslabor (HASYLAB) are discussed in terms of the anharmonic atomic vibrations using the effective one-particle-potential formalism. The outcome concerning third- and fourth-order anharmonicity is in accordance with previous results of the authors derived by least-squares fitting of measured Bragg intensities and disprove results given by Merisalo & Larsen [*Acta Cryst.* (1979) **A35**, 325-327] and Merisalo, Järvinen & Kurittu [*Phys. Scr.* (1978), **17**, 23-25]. The measured very weak intensity of the almost forbidden 061 reflection can be well interpreted in terms of a small but significant antisymmetric anharmonic thermal motion of the Zn atoms characterized by the third-order anharmonic temperature parameter $\alpha_{33} = -0{\cdot}16\,(2) \times 10^{-19}$ J Å^{-3}.

Introduction

The conditions limiting possible reflections for special atomic positions in the unit cell, as given in *International Tables for X-ray Crystallography* (1974), are correct only in the case of centrosymmetric scattering centres. In conventional structure analysis the assumption is made that the electron distribution of an atom has spherical symmetry. The Bragg reflections not matching the conditions limiting possible reflections are supposed therefore to have zero intensity. In general, however, crystal atoms are not expected to be spherically symmetric, and therefore such reflections are not strictly forbidden.

The deviations from spherical symmetry of the atomic electron cloud can be due to static directional distortions of the electronic charge distribution associated with chemical bonding in the structure as well as with the dynamic asphericity associated with anharmonic temperature vibration. Because asphericity due to chemical bonding mainly affects the Bragg intensity of the low-order reflections, the effect of anharmonic motion can usually be separated by a measurement at high values of the scattering vector **h** where the effects of bonding can be ignored.

In Dawson's structure-factor formalism (Dawson, 1967) the deviations from spherical symmetry of the atomic electron cloud is taken into consideration by replacing the spherical atomic electron densities of the atoms by vibration-modified aspherical densities ρ'_j given as the convolution of the aspherical at-rest distribution ρ_j and the aspherical nuclear thermal smearing function t_j:

$$\rho'_j = (\rho_c + \rho_a)_j * (t_c + t_a)_j, \qquad (1)$$

where the subscripts c and a denote the parts of ρ and t which possess centrosymmetry and antisymmetry, respectively, about the nuclear positions r_j of the atoms j in the unit cell.

The Fourier-transform relation between the structure factor and the electron density of the crystal is then given by

$$F(\mathbf{h}) = \sum f_j(\mathbf{h})\, T_j(\mathbf{h}) \exp(2\pi i \mathbf{h}\mathbf{r}_j) = A(\mathbf{h}) + iB(\mathbf{h}), \qquad (2)$$

where the transforms of ρ_j and t_j, the atomic scattering factor f_j and the temperature factor T_j, are both

0108-7673/90/050413-10$03.00