

## The efficient generation of phase sets by the random approach

K HENDERSON and M M WOOLFSON

Department of Physics, University of York, Heslington, York YO1 5DD, UK

**Abstract.** It is shown that a large number of plausible sets of phases satisfying triple-phase relationships may be found very efficiently and cheaply. Sets of phases satisfying the original set of phase relationships ( $\phi$ s) may be combined with 'solutions' ( $\psi$ s) for the set of relationships with any translationally-derived constant angles removed. An application to a known structure demonstrates the validity of the technique and a theoretical interpretation is given for a case where a true solution results from a  $\phi$  and  $\psi$  combination.

**Keywords.** Direct methods; phase sets; random approach

### 1. Introduction

An individual triple-phase relationship, with phases expressed in cycles, may be written as

$$\phi_p \pm \phi_q \pm \phi_r + b \approx n, \quad (1)$$

where the constant angle  $b$  may arise from translational symmetry,  $\approx$  means 'is probably close to' and  $n$  is some integer. A complete set of such relationships may be written in matrix notation as

$$\kappa(A\phi + \mathbf{b}) \approx \kappa\mathbf{n}, \quad (2)$$

where the elements of the diagonal matrix  $\kappa$  are the usual  $\kappa$  values,  $2N^{-\frac{1}{2}} |E_p E_q E_r|$  for an equal-atom structure.

In the YZARC method, developed from ideas given by Baggio *et al* (1978), initial random values are assigned to about 100 phases which are subsequently refined by an iterative process to satisfy equation (2). Some 50–100 sets of refined phases are then extended by the tangent formula and thereafter treated and evaluated in the same way as sets of phases generated by MULTAN 80 (Main *et al* 1980).

The YZARC method has been extremely successful in solving structures with up to 100 atoms or so in the asymmetric unit. However it transpires that, in general, the larger the structure the greater is the number of phase sets which need to be examined in order for a correct one to appear. For a small protein it may be necessary for many thousands of phase sets to be generated in order to obtain just one which may reveal the structure. This would be very expensive in computer time if done in a straightforward way but, as will be demonstrated, an economical process is available.

### 2. $\psi$ sets

Consider a set of equations

$$\kappa A\psi = \kappa\mathbf{m} \quad (3)$$

where  $\mathbf{A}$  is the same matrix as in (2) and  $\mathbf{m}$  is a vector whose elements are close to integers. The YZARC process may be used to find  $\psi$ 's which satisfy (3).

A  $\phi$  solution of (2),  $\phi_i$ , and a  $\psi$  solution of (3),  $\psi_j$ , are combined to give

$$\kappa\{\mathbf{A}(\phi_i \pm \psi_j) + \mathbf{b}\} \approx \kappa(\mathbf{n}_i \pm \mathbf{m}_j). \quad (4)$$

Since the elements of  $\mathbf{n}_i \pm \mathbf{m}_j$  will be close to integers then  $\phi_i \pm \psi_j$  will be near solutions of (2). If there are  $P$   $\phi$  solutions and  $Q$   $\psi$  solutions then there will be  $P(2Q + 1)$  available solutions or near-solutions of (2). The advantage of this is apparent when we consider a numerical case. With  $P = Q = 50$  then  $P(2Q + 1) = 5050$  and this number of possible phase sets is obtained with little more computer time than is required to generate 100 sets of phases with YZARC.

A question which arises is the following: 'For each  $\phi$  solution is there a  $\psi$  solution which can be combined with it to give an almost-correct set of phases?'

If  $\phi_i$  is the set of true phases then for a particular solution of (2),  $\phi_i$ , a correction vector would be

$$\Delta\phi_i = \phi_i - \phi_i.$$

If  $\Delta\phi_i$  is substituted into (3) and refined and if the refined elements are close to their initial values ( $\leq 40^\circ$  different, say) then we may assume that there is an obtainable  $\psi$  which will, combined with  $\phi_i$ , give a substantially correct solution.

Tests with several structures indicate that for about 20% of  $\phi$  solutions there is a potentially-available  $\psi$  which will lead to a structure solution.

### 3. Figures of merit for small sets of phases

In the process which is to be described it is necessary to associate a figure of merit (FOM), a measure of plausibility, to sets of 100 or less phases. It is well-known that the FOMs used in MULTAN are very effective but these are applied to phase sets which define the structure well, in the sense that a Fourier synthesis would show the correct structure, or most of it, quite clearly.

We have found a group of FOMs, suitable for a small data set, which can be expressed in terms of the quantities

$$\alpha_{\mathbf{h}} = \left| \sum_{\mathbf{h}'} \kappa_{\mathbf{h}, \mathbf{h}'} \exp \{i(\phi_{\mathbf{h}} - \phi_{\mathbf{h}'} - \phi_{\mathbf{h}-\mathbf{h}'})\} \right|. \quad (5)$$

These are:

$$(i) \text{ ABSFOM} = \sum_{\mathbf{h}} |\alpha_{\mathbf{h}}| / \sum_{\mathbf{h}} \alpha_{\mathbf{h}, e}, \quad (6)$$

$$(ii) R_1 = \sum_{\mathbf{h}} |\alpha_{\mathbf{h}, e} - k\alpha_{\mathbf{h}}| / \sum_{\mathbf{h}} \alpha_{\mathbf{h}, e}, \quad (7)$$

$$(iii) R_2 = \left\{ \sum_{\mathbf{h}} (S\alpha_{\mathbf{h}, e} - \alpha_{\mathbf{h}})^2 / \sum_{\mathbf{h}} \alpha_{\mathbf{h}, e}^2 \right\}^{\frac{1}{2}}, \quad (8)$$

$$\text{and (iv) MR} = \min_{\mathbf{h}} (\alpha_{\mathbf{h}} / \alpha_{\mathbf{h}, e}) \quad (9)$$

where an estimate of the theoretically expected value of  $\alpha_h$  is

$$\alpha_{h,e} = \sum_{h'} \kappa_{h,h'} I_1(\kappa_{h,h'}) / I_0(\kappa_{h,h'}), \quad (10)$$

$$k = \sum_h \alpha_{h,e} / \sum_h \alpha_h, \quad (11)$$

$$\text{and } S = \sum_h \alpha_{h,e} \alpha_h / \sum_h \alpha_{h,e}^2 \quad (12)$$

which makes  $R_2$  a minimum.

These FOMS are not good as absolute figures of merit but they do seem to work quite well as ranking figures. In a large number of experiments it has been found that combining a  $\psi$  with a  $\phi$  usually lowers ABSFOM. For a few combinations there is a small increase in ABSFOM.

Since a promising  $\phi \pm \psi$  should have a reasonably large ABSFOM ( $\geq 1$ , say) then only those  $\phi$  solutions for which ABSFOM  $\geq 0.95$  have been used for combining with  $\psi$ s. This rule of thumb may need to be modified in the light of further experience.

#### 4. A trial

The idea of combining  $\phi$  and  $\psi$  sets was tested with the known structure of cortisone,  $C_{21}H_{28}O_5$ ;  $P2_12_12_1$ ;  $a = 10.040$ ,  $b = 23.649$ ,  $c = 7.784$  Å;  $Z = 4$  (figure 1).

For a normal YZARC run 100 sets of 100 phases were refined to satisfy equation (2), expanded to 204 phases and then analysed by the usual MULTAN procedure. Four correct solutions were found—sets 46, 57, 59 and 98; these revealed all the non-hydrogen atoms of the structure.

For each of the 100  $\phi$ s (each containing 100 phases) from the YZARC run FOMS were calculated. There were then derived 51  $\psi$  sets by running YZARC with (3). Each  $\phi$  for which ABSFOM  $\geq 0.95$  was combined with each  $\psi$  solution and a  $\phi \pm \psi$  combination was investigated further if its ABSFOM  $\geq 0.95$  and if its  $R_1$  was smaller than that for  $\phi$  alone.

Each of the 384 combinations satisfying these criteria was extended to 204 phases by the tangent formula and then subjected to the usual MULTAN figures of merit. Of the 384 combinations 26 gave the entire structure.

For each of the original  $\phi \pm \psi$  combinations (with 100 phases) the FOMS (6) to (9) were calculated and a combined figure of merit, CFOM, produced from them (as is done in

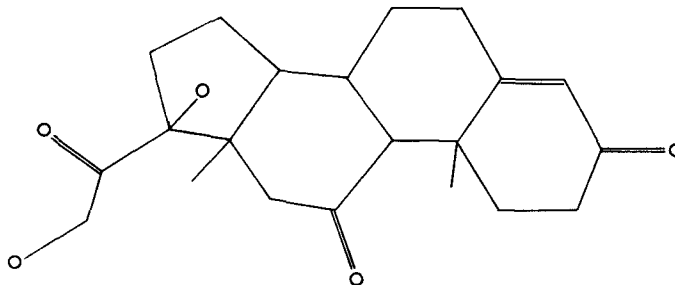


Figure 1. A molecule of cortisone.

**Table 1.** The 26 combinations giving a solution for cortisone

Set	$\phi$	$\psi$	CFOM	Set	$\phi$	$\psi$	CFOM
71	14	21	3.30	378*	98	-3	2.67
290*	59	-32	3.12	212*	46	3	2.58
315	71	31	2.94	365	92†	-50√	2.55
215*	46	-19	2.92	188	43†	-50√	2.55
220*	46	24	2.90	189	43†	50√	2.55
219*	46	-24	2.81	211*	46	-3	2.54
223*	46	-32	2.80	153	31†	50√	2.54
216*	46	19	2.78	366	92†	50√	2.49
270*	57	19	2.77	278*	57	51	2.46
269*	57	-19	2.77	152	31†	-50√	2.46
273*	57	-24	2.73	264*	57	-3	2.45
274*	57	24	2.70	279*	57	32	2.45
381*	98	-32	2.68	265*	57	3	2.34

\*  $\psi$  trivial with all elements nearly equal zero.

† These  $\phi$ s are identical.

√  $\psi_{50}$  has all elements close to 0 or  $\pi$ .

MULTAN). Table 1 shows the 26 solutions in ranking order of CFOM.

Set 71 had the highest CFOM of all the 384 combinations produced and led to an  $E$  map of exceptionally high quality. It is also worth noting that set 265, the correct solution with the lowest CFOM was 161st in ranking order for the 384 combinations so that all the correct solutions were comfortably contained in the top half of the ranking order.

## 5. An interpretation

Set 71, that with the highest CFOM, corresponded to the combination  $\phi_{14} + \psi_{21}$ . Figures of merit were:

$$\begin{array}{l} \text{for } \phi_{14} \\ \text{ABSFOM} = 1.000 \quad R_1 = 0.220 \\ R_2 = 0.231 \quad MR = 0.059 \end{array}$$

$$\begin{array}{l} \text{for } \phi_{14} + \psi_{21} \\ \text{ABSFOM} = 1.060 \quad R_1 = 0.201 \\ R_2 = 0.198 \quad MR = 0.290. \end{array}$$

It can be seen that all the figures of merit are better for the combination than for  $\phi_{14}$  alone. This type of behaviour is rare and in our limited experience seems nearly always to indicate that a correct solution has been found by combination.

In figure 2(a) there is shown a peak-interpreted  $E$  map for  $\phi_{14}$  and in figure 2(b) the corresponding map for  $\phi_{14} + \psi_{21}$ . It is clear that  $\phi_{14}$  is showing a partially correct fragment of the molecule but displaced. In figure 2(b) the peaks are ringed corresponding to part of the configuration shown in figure 2(a). The coordinates of two corresponding points are shown and it is clear that the displacement of the fragment in figure 2(a) is approximately (0.44, 0, 0).

Let us assume that for each phase

$$\phi_{ih} + \psi_{jh} = \phi_{ih}, \quad (13)$$

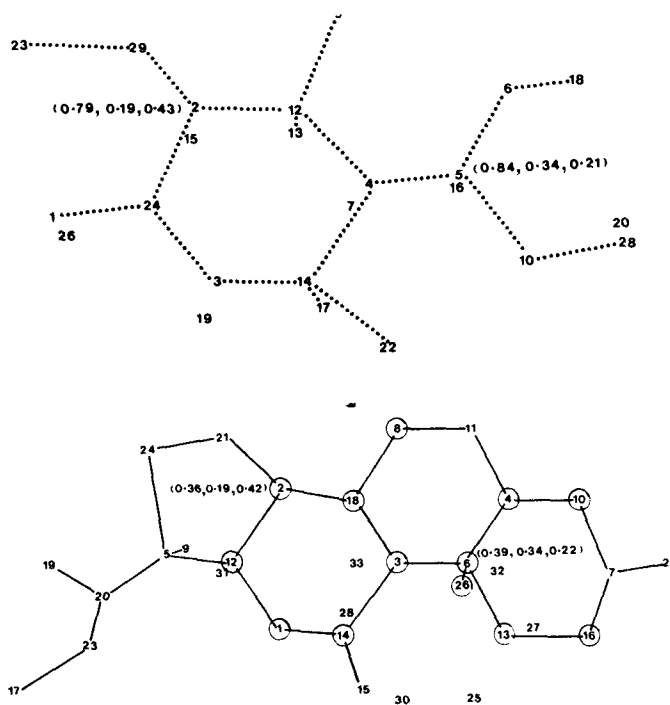


Figure 2. a. The fragment from  $\phi_{14}$ . b. The molecule from  $\phi_{14} + \psi_{21}$ .

where  $\phi_{ih}$  is the true phase. Then since

$$E_h = |E_h| \exp(i\phi_{ih}), \quad (14)$$

we may write

$$E_h = |E_h| \exp(i\phi_{ih}) \exp(i\psi_{jh}). \quad (15)$$

A 'true'  $E$  map is the Fourier transform of the set of  $E_h$ . Hence it is the convolution of the  $E$  map obtained using the phases  $\phi_i$  with the map obtained using phases  $\psi_j$  and unit amplitudes. This latter map we may call the error map (Silva and Viterbo 1980).

The error map calculated from  $\psi_{21}$  has as its largest peaks:

Height	$x$	$y$	$z$
3931	0.4326	0.0000	0.0000
2108	0.4312	0.0941	0.8673
1927	0.4358	0.0901	0.1737
1875	0.4330	0.0036	0.6953
1594	0.0346	0.0276	0.3380

We can recognise the highest peak as the shift in the fragments displayed in the two parts of figure 2. This way of thinking about errors in sets of phases generated by direct methods can be most productive and may lead to even more efficient processes for generating large numbers of plausible phase sets.

## 6. Conclusion

The tests carried out with cortisone show that the idea of combining  $\phi$  and  $\psi$  solutions is basically sound. Some questions to which answers are needed are:

- (i) Can we recognise and eliminate duplicate  $\phi$  and  $\psi$  sets?
- (ii) Is the number of different  $\psi$  sets fairly small and if so how do we find all of them?
- (iii) Can we find even better figures of merit for small sets of phases?

For really complicated structures (even small proteins) many thousands or tens of thousands of sets of phases must be investigated. While it seems that we can easily generate many sets we must find good and fast ways to evaluate their plausibility.

## Acknowledgements

One of us ( $\kappa\text{H}$ ) is grateful to the Wellcome Trust for support during the period in which this work was being done. The generous support of the Science and Engineering Research Council for this work is also acknowledged.

## References

- Baggio R, Woolfson M M, Declercq J P and Germain G 1978 *Acta Crystallogr.* **A34** 883  
Main P, Fiske S J, Hull S E, Lessinger L, Germain G, Declercq J P and Woolfson M M 1980 MULTAN 80, a system of computer programs for the automatic solution of crystal structures from x-ray diffraction data. Univ. of York, England.  
Silva A and Viterbo D 1980 *Acta Crystallogr.* **A36** 1065