Acta Cryst. (1975). A**31**, 401

On the Determination of Structure Invariants. II. A Method Using the Distribution of Triple-Phase Invariants

By T. Debaerdemaeker

Universität Marburg, Fachbereich Geowissenschaften, Marburg/Lahn, Germany (BRD)

and M. M. Woolfson

Department of Physics, University of York, Heslington, York, England

(*Received* 1 *November* 1974; *accepted* 10 *December* 1974)

The number of triple-phase relationships interrelating a set of phases, Q, may greatly exceed the number of unknown phases, M. Consequently it is usually possible to find a linearly independent set of M triple-phase invariants and the remaining Q–M may be expressed as linear combinations of these. From the expected distribution in the values of all the invariants and the interrelationships between them it is possible to find values for the independent set of relationships which are closer to their true values than zero, the *ab initio* expection value of each of them. Ways of using the calculated values of the invariants are described and results are given for various trial structures. Possible improvements in the invariant-determining process are discussed.

Introduction

The symbolic-addition and multiple-solution methods, which are widely used to solve crystal structures at the present time, make use of the triple-phase relationship

$$S_{\mathbf{h}_1,\mathbf{h}_2}=\varphi_{\mathbf{h}_1}+\varphi_{\mathbf{h}_2}+\varphi_{\mathbf{h}_3}\approx 0 \bmod (2\pi) \qquad (1)$$

where $\mathbf{h}_1+\mathbf{h}_2+\mathbf{h}_3=0$ and $\approx$ implies 'is distributed about'. The distribution of S has been given by Cochran (1955) and the variance of the distribution was given in analytical and graphical form by Karle & Karle (1966).

The usual methods of using the S's, the triple-phase invariants, assume initially that each of them has the value zero. This assumption eventually introduces inconsistencies, where two relationships would give different indications for a new phase, and an averaging formula, the tangent formula,

$$\tan\varphi_{\mathbf{h}}=\frac{\sum\limits_{\mathbf{h}'}|E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}|\sin(\varphi_{\mathbf{h}'}+\varphi_{\mathbf{h}-\mathbf{h}'})}{\sum\limits_{\mathbf{h}'}|E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}|\cos(\varphi_{\mathbf{h}'}+\varphi_{\mathbf{h}-\mathbf{h}'})}, \qquad (2)$$

is then brought into play to combine the differing indications.

However, the values of M unknown φ's could be determined from a properly selected set of M S's if the values of these S's were known precisely. Starting with some known phases, which could include those used to fix the origin and enantiomorph and also $\sum_1$ phases, it requires one phase relationship containing two known and one unknown phase, to find each unknown phase from

$$\varphi_{\mathbf{h}_1}=S_{\mathbf{h}_1,\mathbf{h}_2}-\varphi_{\mathbf{h}_2}-\varphi_{\mathbf{h}_3}\,. \qquad (3)$$

In this way M unknown phases may be found in terms of M S's. Of course this would not normally give a solution of sufficient accuracy if each S was put equal to zero, its most likely value *ab initio*.

The general form of the final equation giving the values of φ is

$$\varphi_i=\sum_{r=1}^{M}a_{i,r}S_r+b_i \qquad (4)$$

where i may take values from 1 to M, the a's are integer coefficients and the b's arise because of the known phases and any translational symmetry associated with the space group.

Interdependence of the triple-phase relationships

If a sufficiently large number of unknown phases, M, is considered then the number of triple-phase relationships linking them, Q, will be much greater than M. Typically, with $M=250$ one may have Q of order 4000 or so. From what has been said previously it is possible to express the φ's as linear combinations of a linearly independent set of M of the S's. Clearly, by substitution of expressions such as (4) for the φ's in the remaining $(Q-M)$ triple-phase relationships the values of these S's can be expressed in terms of the linearly independent set of M. These latter clearly form a basis in terms of which all S's derived from the specified set of φ's may be expressed.

We may write

$$S_t=\sum_{r=1}^{M}a_{t,r}S_r+b_t \qquad (5)$$

where t takes the values from 1 to Q. For $t\le M$, $b_t=0$ and $a_{t,r}=\delta_{t,r}$ the Kronecker delta.

It should be noted that equations (5) are exact equations and the 'quadrupole' described by Viterbo & Woolfson (1973) is a special type of such relationship involving a total of only four S's.

It is interesting to insert into equations (5) the values zero for the basis set of S's. This gives the constants, b, as the indicated values for the complete set of S's and in Fig. 1(a) is shown a typical distribution of 2000 values of S which might result in such a case. Such a distribution is unrealistic and if the values of S for the complete data set were known they might be expected to give a distribution pattern as shown in Fig. 1(b).

Probable values of the basis invariants

We have used a knowledge of the expected distribution of values of the complete set of S's as a means of finding improved values for the basis set. We define a function

$$\eta(S_1,S_2,\ldots S_M)=\sum_{r=1}^{Q}|E^3|_r\cos(S_r) \qquad (6)$$

where $|E^3|_r$ is the product of the magnitudes of the three E's associated with the rth phase relationship and each S on the right-hand side may be expressed in terms of the basis as in expression (5). We attempt to refine the values of the basic set of S's by maximizing the value of η starting with each value of S equal to zero. This maximization will tend to change the distribution of the total set of S's away from the starting point, as illustrated in Fig. 1(a), towards the distribution shown in Fig. 1(b).

The refinement procedure

Various methods of maximizing the value of η have been tried. It turns out that the simplest method, the parameter-shift method, is also the most efficient. In this method the values of S are considered singly and the value of η is computed with the S under investigation changed by $k\delta$ where δ is a selected quantity and k is varied in steps of unity from -5 to $+5$ (Bhuiya & Stanley, 1963). The value of S is then changed to $S+k_m\delta$ where k_m is the value of k which gave a maximum value of η. This is done for all the S's in turn and the whole process is repeated in cyclic fashion. As the refinement proceeds it is necessary from time to time to reduce the value of δ until the point is reached where the shifts are too small to be significant. For example in the first cycle the step may be 18° and reduce to 1° by the ninth or tenth cycle.

Tests have been made of this idea with a number of known trial structures. In every case the finally determined values of the structure invariants were better, in a statistical sense, than the usual assumption that each of them was equal to zero.

For example we took as a test case the photolysis product of Karle, Karle & Estlin (1967). The strongest 2000 $\sum_2$ relationships were selected linking 179 of the largest E's. Seven of the phases were taken as known – four which fixed the origin and enantiomorph and three others which the *CONVERGENCE* routine in *MULTAN* identified as necessary to include in the starting set. A basis of 172 invariants was selected whose mean deviation from zero was 37·9°. After parameter-shift refinement the mean deviation of the calculated values of these invariants from their true values was 31·9°. The corresponding quantities for the complete set of 2000 relationships were 43·0 and 34·3°. It must be stressed however that this improvement is an overall one – some individual invariants may actually get worse as a result of the refinement procedure.

This result is typical of the general run of results using this technique. There is a reduction of between 15 and 20% in the mean error of invariants in the basis set and a somewhat smaller percentage reduction for the total set of invariants.

This reduction should not be underestimated. If the mean error is changed by a factor α then the variance is changed by a factor α^2. From the work of Karle & Karle (1966) it can be seen that over the range of interest the variance of the phase relationships varies approximately as κ^{-1} where

$$\kappa=2N^{-1/2}|E^3|\,. \qquad (7)$$

Thus to change the average variance by a factor α^2 requires N, the number of atoms per unit-cell, to change by a factor α^4. Hence the mean error of invariants is reduced by 15% by reducing the number of atoms in the cell to about 0·5 of their original number. A reduction in the mean error of invariants by 15% may thus be regarded, as far as direct methods are concerned, as equivalent to handling data corresponding to a much simpler structure.

Using the invariants

From the determined values of the basis invariants one may, by inverting the set of equations of type (1), find corresponding values for the phases. It could be argued that these phases have been determined from all the invariants, albeit indirectly, and since the invariants have been analysed as a complete set the pitfall of con-

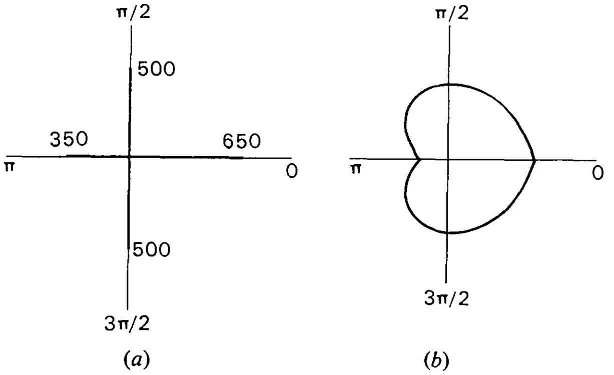

Fig. 1. (a) The distribution of values of a complete set of triple-phase invariants when each of the basic set is set at zero. (b) The expected distribution of values of the triple-phase invariants.

sidering the invariants in a chain process, where one step badly in error may invalidate all subsequent phase determination, will have been circumvented. Our experience has been rather mixed up to the present time and it appears that generally one does get better phases from the calculated invariants but occasionally better phases are obtained by a chain process using the tangent formula. In Table 1 some results are summarized.

Table 1. *Comparison of phase determination from calculated values of structure invariants, S_{calc}, with those from use of the tangent formula (T.F.)*

Structure	Number of unknown phases	Total number of $\sum_2$	Mean error from S_{calc}	Mean error from T.F.
Photolysis product	201	2000	56·6	83·3
Estrone	279	2000	45·2	19·5
Sulphur 18	179	2000	15·4	65·8
Cortisone	300	2000	22·4	87·7

We feel convinced that the technique we are using in processing the total set of interdependent invariants is giving potentially useful information. What we are not convinced about, however, is that we are using the best technique to find the invariants or that we are using this information in the best possible way to solve structures and investigation in this area is continuing.

Possible extensions of the method

One drawback to the procedure we are using is that the correct values of the invariants do not give a maximum of η – neither an absolute maximum nor, indeed, even a local maximum. This we have checked by starting the refinement process with all correct values of the invariants and they have moved well away from the starting point. This is shown in Table 2 which corresponds to a situation with 92 phases and 410 invariants. This trial was made with unit weight for each invariant, rather than $|E^3|$ as in equation (6), but the general quality of the result is not influenced greatly by the weighting scheme.

Table 2. *The refinement of phases starting from* (I) *all invariants initially equal to their correct values and* (II) *all invariants initially equal to zero*

	(I)		(II)	
Cycle	η	$\lvert\Delta S\rvert$	η	$\lvert\Delta S\rvert$
Start	219·4	0·0°	110·8	30·9°
1	309·2	10·8	292·6	30·0
2	313·3	15·0	297·5	29·3
3	315·1	16·1	307·6	28·3
4	317·5	17·3	312·6	27·9
5	319·9	18·5	316·9	27·2
6	321·7	19·5	319·7	27·0
7	323·2	20·6	321·8	26·8
8	324·1	21·3	323·2	26·6
9	324·9	21·9	324·2	26·5
10	325·5	22·0	324·2	26·5

Starting with each S equal to zero the value of η was 110·8, *i.e.* the average value of a cosine invariant was 0·27 (110·8/410). The mean error in S at this stage was 30·9°. After ten cycles of refinement the average cosine invariant equals 0·79 and the mean error in S has fallen to 26·5°.

With each S given its true value initially the average cosine is 0·53 and ten stages of refinement change this to 0·79 with a mean error in S of 22·0°. The sets of final phases from the two starting points differ by an average of 18·5°, *i.e.* they are almost as different from each other as from the correct set of phases.

Clearly what is required is a function for maximization (or minimization) for which the correct set of phases is an extremum and preferably an extremum of greatest magnitude. Some ideas for better functions follow.

Any relationships between phases may be used in the same way as, and added to, the triple-phase relationship to strengthen the determination of the basis set. A relationship offering this possibility is the 'negative quartet' described by Hauptman (1974). This involves a set of four phases, the vector sum of whose indices is zero and the relationship may be expressed as

$$\varphi_{\mathbf{h}_1}+\varphi_{\mathbf{h}_2}+\varphi_{\mathbf{h}_3}+\varphi_{\mathbf{h}_4}\approx\pi \tag{8}$$

when $|E_{\mathbf{h}_1}|,|E_{\mathbf{h}_2}|,|E_{\mathbf{h}_3}|,|E_{\mathbf{h}_4}|$ are all large and $|E_{\mathbf{h}_1+\mathbf{h}_2}|$, $|E_{\mathbf{h}_1+\mathbf{h}_3}|$ and $|E_{\mathbf{h}_1+\mathbf{h}_4}|$ are all small. As expressed by Hauptman the relationship is that the cosine of the sum of the four phases is probably negative, which explains Hauptman's terminology.

The distribution of well-chosen negative quartets about π can be fairly tight and several hundred quartets may be available in a favourable case. It has been found by Hauptman that negative quartets can be used to establish a sensitive figure of merit for multisolution methods to distinguish the correct set of phases. This suggests that if they are actually used in the phase-determining process they may well be very discriminating and favour the correct set of phases.

Another possibility is the use of Sayre's equation directly to find phases.

Sayre's equation for normalized structure factors may be written as

$$E_{\mathbf{h}}=K_{\mathbf{h}}\sum_{\mathbf{h}'}E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}\ , \tag{9}$$

where $K_{\mathbf{h}}$ is a constant which may be readily determined.

Multiplying each side by the complex conjugate of $E_{\mathbf{h}}, E_{-\mathbf{h}}$, gives

$$|E_{\mathbf{h}}|^2=K_{\mathbf{h}}\sum_{\mathbf{h}'}E_{-\mathbf{h}}E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}\ , \tag{10}$$

an equation involving only structure-invariant quantities.

At this point it may be as well to correct often-held misconceptions concerning Sayre's equation. One of these is that Sayre's equation involves the condition of non-negativity of electron density and the second is that Sayre's equation can only be applied to E's if there is an infinity of data. In fact if the data considered are E's terminated by, say, the Cu $K\alpha$ limiting sphere then these correspond to 'atoms' which are reasonably sharp but which have negative diffraction ripples. However peaks in an E map with all the data will be well resolved and there will be little overlap of neighbouring atoms. If the atoms are equal then the square of this density will also contain equal resolved peaks and this is the necessary and sufficient condition for Sayre's equation to be valid. The constant $K_{\mathbf{h}}$ is given by

$$K_{\mathbf{h}} = g_{\mathbf{h}}/f_{\mathbf{h}}$$

where $g_{\mathbf{h}}$ is the 'scattering-factor' for the squared 'atom' and $f_{\mathbf{h}}$, the scattering factor for the original equal 'atoms' will equal $N^{-1/2}$. The value of $g_{\mathbf{h}}$ is found as a self-convolution of $f_{\mathbf{h}}$ and is proportional to the over-lapped volume of two spheres, of radius equal to the limiting sphere, whose centres are at a distance apart equal to the distance of the point $\mathbf{h}$ from the origin of reciprocal space.

Tests of Sayre's equation have shown that it applies reasonably well when restricted to a subset of the largest E's as long as a constant scaling factor is applied which makes the aveiage magnitude of the right-hand sides of the equations equal to the average of the left-hand sides.

Table 3. *Residual of scaled Sayre's equation for various $|E|$ cut-off values*

$\lvert E\rvert_{\min}$	Number of reflexions	Residual
0·00	940	0·060
1·00	341	0·125
1·25	209	0·169
1·50	113	0·270

The 'residual' for the two sides of the equations, properly scaled, for various values of the E cut-off are given in Table 3 for a trial 40-atom structure.

The individual equations such as (10) can be written as a pair of equations

$$|E_{\mathbf{h}}|^2 = K_{\mathbf{h}} \sum_{\mathbf{h}'} |E_{\mathbf{h}}E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}| \cos(\varphi_{\mathbf{h}'} + \varphi_{\mathbf{h}-\mathbf{h}'} - \varphi_{\mathbf{h}}) \quad (12a)$$

and

$$0 = K_{\mathbf{h}} \sum_{\mathbf{h}} |E_{\mathbf{h}}E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}| \sin(\varphi_{\mathbf{h}'} + \varphi_{\mathbf{h}-\mathbf{h}'} - \varphi_{\mathbf{h}}) \,. \quad (12b)$$

The φ's in the equation may be expressed as linear combinations of a basis set of S's and a refinement of these S's can be based on obtaining equality on the two sides of equations such as (12*a*) and (12*b*). It is also possible to include Sayre's equations for which $|E_{\mathbf{h}}| = 0$, a type of information which it is rarely possible to use in phase determination.

An investigation of the use of negative quartets and of Sayre's equation is currently being pursued.

We wish to express our gratitude to the Computing Centres of the Universities of Marburg and York for the services they have provided. Our thanks are also due to the Science Research Council and to the Deutshe Forschungsgemeinschaft (SFB 127) for support of this work.

References

BHUIYA, A. K. & STANLEY, E. (1963). *Acta Cryst.* **16**, 981–984.

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

HAUPTMAN, H. (1974). *Acta Cryst.* A**30**, 472–476.

KARLE, I. L., KARLE, J. & ESTLIN, J. A. (1967). *Acta Cryst.* **23**, 494–500.

KARLE, J. & KARLE, I. L. (1966). *Acta Cryst.* **21**, 849–859.

VITERBO, D. & WOOLFSON, M. M. (1973). *Acta Cryst.* A**29**, 205–208.