Acta Cryst. (1993). D**49**, 3–8

Phasing Macromolecular Structures *via* Structure-Invariant Algebra

By Herbert Hauptman

Medical Foundation of Buffalo, Inc., 73 *High Street*, *Buffalo*, *New York* 14203, *USA*

and Fusen Han

Upjohn Laboratories, *The Upjohn Company*, *Kalamazoo*, *MI* 49001, *USA*

(*Received* 2 *June* 1992; *accepted* 10 *August* 1992)

Abstract

Owing to the breakdown of Friedel's law when anomalous scatterers are present, unique values of the three-phase structure invariants in the whole range from 0 to 2π are determined by measured values of diffraction intensities alone. Two methods are described for going from presumed known values of these invariants to the values of the individual phases. The first, dependent on a scheme for resolving the 2π ambiguity in the estimate $\omega_{\mathbf{HK}}$ of the triplet $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}$, solves by least squares the resulting redundant system of linear equations $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} = \omega_{\mathbf{HK}}$. The second attempts to minimize the weighted sum of squares of differences between the true values of the cosine and sine invariants and their estimates. The latter method is closely related to one based on the 'minimal principle' which determines the values of a large set of phases as the constrained global minimum of a function of all the phases in the set. Both methods work in the sense that they yield values of the individual phases substantially better than the values of the initial estimates of the triplets. However, the second method proves to be superior to the first but requires, in addition to estimates of the triplets, initial estimates of the values of the individual phases.

1. Introduction

The presence of one or more anomalous scatterers in a crystal has important implications for X-ray crystallography which stem from the breakdown of Friedel's law. Bijvoet (1949) appears to have been the first to recognize the importance of this effect when he exploited it to determine absolute configuration.

It is also known that anomalous dispersion supplements, in an important way, the method of isomorphous replacement for the determination of macromolecular structure. Recent developments, by using anomalous intensity data collected at two or more different wavelengths [see *e.g.* Karle (1984) and Hendrickson, Smith, Phizackerley & Merritt (1988)] or by integrating anomalous dispersion with the traditional techniques of direct methods, have led to new phasing methods for macromolecular structures which promise to strengthen the standard techniques.

It has been known for some time that the availability of single-wavelength anomalous-dispersion data determines unique values of the three-phase structure invariants; and methods for estimating their values are now known. In this paper two methods are described for obtaining the values of the individual phases from the values of the triplets, which are presumed to be known. In the first, estimates of the triplets themselves are used to construct a redundant system of linear equations which is solved by least squares and which incorporates a scheme for resolving the 2π ambiguity inherent in such a formulation. This method determines the values of the phases *ab initio*, which is to say that it does not require any prior structural knowledge or initial estimates for the values of the phases.

The second method employs the estimates of the cosines and sines of the structure invariants (rather than the invariants themselves) to define a function of the phases, the global minimum of which is reached when the phases are set equal to their true values. In contrast to the first method, this one does require some prior structural knowledge or initial estimates of the values of the individual phases and is, therefore, to be regarded as a technique for phase refinement rather than an *ab initio* method of phase determination. It has the advantage, however, that it yields better values for the individual phases.

The first attempt, by algebraic means, to combine the traditional direct-methods techniques with anomalous dispersion appears to have been made by Kroon, Spek & Krabbendam (1977) who showed how to derive phase information (but not explicit formulas for the triplets themselves) in terms of the Bijvoet inequalities. [See also Karle (1984) for related work.] This work was soon followed by a probabilistic version (Heinerman, Krabbendam, Kroon & Spek, 1978) which employed the triplet amplitudes rather than the magnitudes of the individual

0907-4449/93/010003-06$06.00

normalized structure factors $|E|$. Although Heinerman *et al.* did suggest that the joint probability distribution of the six structure factors $E_{\mathbf{H}}$, $E_{\mathbf{K}}$, $E_{-\mathbf{H}-\mathbf{K}}$, $E_{-\mathbf{H}}$, $E_{-\mathbf{K}}$, $E_{\mathbf{H}+\mathbf{K}}$ might lead to improved results, no one appears to have anticipated that the six magnitudes $|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, $|E_{-\mathbf{H}-\mathbf{K}}|$, $|E_{-\mathbf{H}}|$, $|E_{-\mathbf{K}}|$ and $|E_{\mathbf{H}+\mathbf{K}}|$ would actually yield unique estimates for the triplets $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}$ until this work was actually done (Hauptman, 1982; Giacovazzo, 1983). The reader is referred to Fortier (1991) and Krabbendam (1991) for additional details of the historical background.

2. Evaluation of the estimated structure invariants

The latest developments of direct methods have provided several new probabilistic formulas to estimate the structure-invariant values more accurately (Hauptman, 1982; Giacovazzo, 1983; Peschar & Schenk, 1987). Some of these new formulas, for example the formulas dependent on anomalous scattering data, yield unique estimates for the structure invariants instead of their cosines alone. The quality of these estimates can be evaluated in three aspects: (i) they should not be strongly biased compared to the true invariant values; (ii) the weight function, called A values in Hauptman's formula [not to be confused with the $A_{\mathbf{HK}}$ of equation (10)] should reflect the reliability of each individual estimate; (iii) any set of invariants, linear combinations of three phases, should satisfy the quadrupole identities at least approximately (Viterbo & Woolfson, 1973). In short, a set of well estimated invariants must have small quadrupole residuals.

A complete evaluation has been carried out for the known 134-amino-acid protein structure, PtCl_4^{2-} derivative of cytochrome c_{550}, space group $P2_12_12_1$ with $a = 42.70$, $b = 82.17$, $c = 31.56$ Å (Timkovich & Dickerson, 1973, 1976). The coordinates and the experimental structure factors were obtained from the Protein Data Bank (Bernstein *et al.*, 1977). There are 1018 atoms in the coordinate listing and 2996 unique measured structure factors to 2.5 Å resolution. There are no anomalous-dispersion data available for this structure. A simulated anomalous-dispersion data set was obtained as follows.

Table 1. 25 244 *estimated structure invariant values from the* 503 *Friedel pairs of anomalous data, using* (4)

To show their unbiased features the values are sorted into two groups; one contains all estimated values smaller than the true values ($0 < \omega_{\rm true} - \omega_{\rm est} < 180°$, the two left-hand columns), the other contains estimates larger than the true values ($0 < \omega_{\rm est} - \omega_{\rm true} < 180°$, the two right-hand columns).

A	No.	$\langle\lvert\omega_{\rm error}\rvert\rangle$ (°)	No.	$\langle\lvert\omega_{\rm error}\rvert\rangle$ (°)
2.5	2	72.0	0	
2.4	3	117.6	0	
2.3	1	66.1	1	80.0
2.2	1	124.3	3	23.3
2.1	5	33.5	3	46.1
2.0	10	31.2	4	91.5
1.9	9	47.5	6	36.9
1.8	10	56.0	3	10.3
1.7	14	78.2	6	46.2
1.6	33	48.7	21	48.4
1.5	30	69.0	31	49.3
1.4	41	60.1	38	58.5
1.3	80	58.8	56	55.6
1.2	121	57.5	99	48.4
1.1	190	61.2	178	55.2
1.0	295	64.7	284	53.4
0.9	476	60.8	456	59.4
0.8	758	58.6	731	63.0
0.7	1249	63.2	1186	62.1
0.6	1819	65.9	1851	66.5
0.5	2694	70.7	2660	67.9
0.4	2992	73.1	2929	72.2
0.3	1770	75.9	1763	73.1
0.2	183	78.5	149	77.1

First, the magnitudes $|F_{\mathbf{H}}|$ and $|F_{\bar{\mathbf{H}}}|$ were calculated from the known structure for all 2996 reflections with measured intensities. Of these, 503 pairs of the more intense reflections had a calculated difference $||F_{\mathbf{H}}| - |F_{\bar{\mathbf{H}}}||$ which was at least 2% of $|F_{\mathbf{H}}|$. (Current X-ray technology permits a 2% difference to be measured.) Next, to each experimentally measured magnitude $|F_{\mathbf{H}}|_{\rm exp}$, an amount equal to 1% of the calculated difference $||F_{\mathbf{H}}| - |F_{\bar{\mathbf{H}}}||$ was added and subtracted in order to get the simulated Friedel pair $|F_{\mathbf{H}}|$, $|F_{\bar{\mathbf{H}}}|$. Then, 25 244 structure invariants were estimated by means of Hauptman's probabilistic formula [Hauptman, 1982, equation (3.65)], the conditional probability distribution of the triplet Ω (2), given the six magnitudes (3):

$$P(\Omega \,|\, |E_{\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{-\mathbf{H}-\mathbf{K}}|, |E_{-\mathbf{H}}|, |E_{-\mathbf{K}}|, |E_{\mathbf{H}+\mathbf{K}}|) = (1/C)\exp[A\cos(\Omega - \omega)] \qquad (1)$$

where Ω represents the triplet:

$$\Omega = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}. \qquad (2)$$

The six magnitudes (in general distinct because of the breakdown of Friedel's law),

$$|E_{\mathbf{H}}|,\ |E_{\mathbf{K}}|,\ |E_{-\mathbf{H}-\mathbf{K}}|,\ |E_{-\mathbf{H}}|,\ |E_{-\mathbf{K}}|,\ |E_{\mathbf{H}+\mathbf{K}}|, \qquad (3)$$

constitute the first neighbourhood of the triplet Ω. The parameters C, A and ω of the distribution (1) are explicitly expressed in terms of the six magnitudes (3) and the atomic scattering factors, in general complex because of the anomalous dispersion, of the atoms in the crystal. Since the distribution (1) has a unique maximum at $\Omega = \omega$ in the whole interval $(0, 2\pi)$, (1) yields the estimate

$$\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} = \omega. \qquad (4)$$

These 25 244 structure invariants were used to generate 221 057 quadrupole identities.

The results of these estimates are listed in Table 1. The average value of the absolute error for all 25 244 invariants is $\langle|\omega_{\rm error}|\rangle = 68.2°$; the average value of the signed error is $\langle\omega_{\rm error}\rangle = 2.1°$; if the A values are used as a weight function when averaging,

$\langle|\omega_{\text{error}}|\rangle$ is reduced to 65.9°. Among the 25 244 invariants, 999 with $A > 1.0$ give $\langle|\omega_{\text{error}}|\rangle = 56.2°$; 24 245 with $A < 1.0$ give $\langle|\omega_{\text{error}}|\rangle = 68.7°$. Column three of Table 1, with 12 786 invariants, has $\langle|\omega_{\text{error}}|\rangle = 68.9°$ and column five, with 12 458 invariants has $\langle|\omega_{\text{error}}|\rangle = 67.45°$. The calculated quadrupole residual is 8.2°. All of these statistics show that (i) the estimated values, ω, are unbiased, (ii) A values are meaningful and (iii) the quadrupole residual is small.

3. The 2π ambiguity

In view of (1)–(4), formulas having probabilistic validity are now available which yield unique estimates (and measures of reliability) for the three-phase structure invariants (triplets), $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} = \omega_{\mathbf{HK}}$, provided that anomalous-dispersion data are available (Hauptman, 1982; Giacovazzo, 1983). The existence of these formulas raises the question: Is it possible to go from estimates of the structure invariants to the values of the individual phases or, to put it another way, can one solve the redundant system of linear equations

$$\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} = \omega_{\mathbf{HK}} + 2\pi N_i \qquad (5)$$

for the unknown phases φ when the estimates $\omega_{\mathbf{HK}}$ are presumed to be known? The problem is complicated by the fact that, although unique estimates for the $\omega_{\mathbf{HK}}$ in the interval 0 to 2π are known, an integer multiple of 2π [$2\pi N_i$ in (5)], unknown *a priori*, must be added to these estimates if the system of equations (5) is to be consistent (Woolfson, 1977). The resolution of this 2π ambiguity has been described in a recent paper (Han, DeTitta & Hauptman, 1991). Briefly, the values of three origin-fixing phases are arbitrarily specified, an additional p (say ten) unknown phases are selected, $q > p$ triplets are generated from these $p + 3$ phases, in this way leading to a system of q equations (5). From this system of equations a suitably chosen linearly independent subset consisting of p equations in the p unknown phases is chosen. The solution of this system of equations yields initial estimates for the values of the p unknown phases. The 2π ambiguity in the estimate of the $\omega_{\mathbf{HK}}$ is no problem at this stage since the addition of an arbitrary integer multiple of 2π to the right-hand sides of the system (5) merely adds a like multiple of 2π to an individual phase φ because the number of equations is equal to the number of unknowns. Once initial estimates of the p unknown phases φ are determined, they may be improved by solving the full redundant system of q ($> p$) equations (5) involving the initial p phases. Since initial estimates of these p phases are now available the 2π ambiguity is readily resolved by calculating the left-hand side of each equation (5) and adding a suitable integer multiple of 2π to the estimate on the right-hand side in such a way as to insure approximate equality. A least-squares adjustment in the initial estimates of the p unknown phases then suffices to obtain better estimates of their values.

With refined values of the starting set of $p + 3$ phases available, additional phases are obtained, one at a time, as follows. A new phase $\varphi_{\mathbf{H}}$ is selected and used to generate all triplets involving $\varphi_{\mathbf{H}}$ and the starting set of $p + 3$ phases, in this way adjoining an additional q_1 equations (5) to the q equations (5) already used. If $\varphi_{\mathbf{H}}$ is restricted to lie in the interval 0 to 2π, and the known values of the initial $p + 3$ phases φ and the known estimates on the right-hand side of (5) are employed, then the number of unknown integer multiples of 2π to be added to the right-hand side of each equation (5) is greatly restricted. In fact, a simple argument shows that for the additional q_1 equations (5) involving the single unknown phase $\varphi_{\mathbf{H}}$ there are precisely $q_1 + 1$ different sets of additive integer multiples of 2π (Han *et al.*, 1991). The least-squares solution of each of these $q_1 + 1$ systems of equations (5), each consisting of $q + q_1$ equations in $p + 1$ unknowns, yields $q_1 + 1$ sets of values for the $p + 1$ unknown phases. The best solution is selected by choosing the one with the smallest residual. The process is repeated and, in the later stages, combined with the tangent formula to yield values of individual phases with average errors of approximately 30° when estimates of the structure invariants having average errors of 65° are available. Details of the applications are described in the next section.

4. Least-squares solution of the structure-invariant equations

The structure invariants are the linear combinations of phase angles (5). When we directly use these equations, the integer N's have to be taken into account. However, the linear-equation method uses a set of these equations which are sufficient in number to solve for values of an equal number of phases. In the situation where the number of unknowns and the number of linear equations are exactly equal it is possible to arbitrarily assign the integer portions. This method will allow us to have a small starting set of phases without consideration of the integer problem. This starting equation set must (i) be linearly independent; (ii) have a number of equations equal to the number of unknowns; (iii) have good estimated invariant values; and (iv) interact strongly with other invariants. The best place to find the required set of these equations is at the bottom of a special convergence map. For our test data set, the first linear equation set contains seven invariants, ten phases are involved and among them three are origin-fixing reflections. The average A value of these seven invariants is 1.03, which is somewhat larger than the average of the total. The origin-fixing reflections are treated as discussed in a previous publication (Han *et al.*, 1991).

The second step resolves the integer problem in a different way. It uses the results of the starting equations to calculate the integer portion of all the invariants (11 in this case) which involve only the first ten phases (only seven of which are to be determined since three serve to fix the origin). By combining the estimated values of the triplets and the calculated integers, a set of overdetermined linear equations (11 equations with seven unknowns) is obtained and solved by a full-matrix least-squares method.

When any invariant involves other than the first ten phases, the integer problem has to be considered. To simplify the procedure of finding correct integers, we add only one phase at a time. As described above for one 'new phase' and *m* more 'new invariants' only a few possible integer sets need be tested. Based on the fact that the correct integers should normally give smaller residuals for the equations, we only need to keep one or a few sets of the results which have the smallest residuals for later consideration. Our experience shows that the average phase error quickly diminishes as the phase set grows, a consequence of the increasing overdetermination of the system (5) as more and more phases become involved. In Table 2, the early steps always give large errors, but these fall off quickly as the set of phases develops. In our test, the average error of phases rapidly diminished to 21.7° when 16 phases and 31 invariants were involved and leveled off at 21.0° for 50 phases and 200 invariants. Then the base of phases steadily expanded to 100 and, finally, 250 (Table 2). Note that by the time estimates of 250 phases are found, the average phase error has increased to 51.9° despite the great overdetermination. This is a consequence of the need to use poorly estimated triplets with small *A* values resulting in increasing numbers of incorrect integer assignments on the right-hand side of (1). A significant improvement employing the minimal function approach is described next.

Table 2. 250 *phases determined by the trial-and-error least-squares method* (5)

No. of phases	No. of invariants	$\langle\lvert\varphi_{\mathrm{error}}\rvert\rangle$ (°)
10	7	116.3
10	11	102.5
16	31	21.7
50	200	21.0
100	679	30.3
250	3249	51.9

5. The minimal principle (a digression)

Identities among the phases

The normalized structure factors are defined by

$$E_{\mathbf{H}} = |E_{\mathbf{H}}|\exp(i\varphi_{\mathbf{H}}) = \frac{1}{N^{1/2}} \sum_{j=1}^{N} \exp(2\pi \mathbf{H}\cdot\mathbf{r}_j). \quad (6)$$

N is the number of atoms in the unit cell, for simplicity assumed here to be identical, and $\mathbf{r}_j$ is the position vector of the atom labeled *j*. Since the number of magnitudes $|E_{\mathbf{H}}|$ obtained from the X-ray diffraction experiment usually greatly exceeds the number of parameters (3*N*) needed to define the crystal structure, elimination of the atomic position vectors $\mathbf{r}_j$ from the system of equations (6) leads to a system of identities among the phases $\varphi_{\mathbf{H}}$ only, dependent on the presumed known magnitudes $|E|$, which must of necessity be satisfied:

$$G(\varphi_{\mathbf{H}}\,||E_{\mathbf{H}}|) = 0. \quad (7)$$

The minimal principle

The minimal function $R(\varphi)$, a function of the phases, is defined by:

$$R(\varphi) = \frac{\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \left[\cos T_{\mathbf{HK}} - \frac{I_1(A_{\mathbf{HK}})}{I_0(A_{\mathbf{HK}})} \right]^2}{\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}} \quad (8)$$

where

$$T_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} \quad (9)$$

$$A_{\mathbf{HK}} = (2/N^{1/2})|E_{\mathbf{H}}E_{\mathbf{K}}E_{\mathbf{H}+\mathbf{K}}| \quad (10)$$

I_0 and I_1 are the modified Bessel functions, and the sums are taken over all reciprocal lattice vectors **H**, **K** associated with the triplets $T_{\mathbf{HK}}$ which are generated by a specified basis set of phases $\{\varphi\}$ corresponding to the largest values of $|E|$. Note that, in view of (9), (8) defines $R(\varphi)$ as a function of the phases φ. Then, provided that the basis set of phases $\{\varphi\}$ is chosen sufficiently large and the phases are constrained by the system of identities (7), the minimal function R(φ) has a constrained global minimum at the point that all the phases are equal to their true values for some choice of origin and enantiomorph (the minimal principle). In this way the problem of phase determination is replaced by the problem of finding the constrained global minimum of the minimal function $R(\varphi)$, equation (8), a known function of the phases. Methods have been devised for finding this minimum and they are described elsewhere in this issue (Weeks, DeTitta, Miller & Hauptman, 1993).

An important aspect in the implementation of the minimal principle is the ability to calculate *ab initio*, *i.e.* without prior knowledge of the values of the phases, the value, R_T, of this constrained global minimum:

$$R_T = \tfrac{1}{2} + \frac{1}{\sum\limits_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}} \left[\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} (\tfrac{1}{2} t'_{\mathbf{HK}} - t^2_{\mathbf{HK}}) \right] < \tfrac{1}{2} \quad (11)$$

where

$$t_{\mathbf{HK}} = I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}}) \quad (12)$$

$$t'_{\mathbf{HK}} = I_2(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}}) \quad (13)$$

and the I's are the modified Bessel functions.

6. The minimal function approach employing estimated values of the triplets

Instead of using the estimate $\omega_{\mathbf{HK}}$ in equation (5), one may bypass entirely the problem posed by the 2π ambiguity by employing the estimates $c_{\mathbf{HK}}$ and $s_{\mathbf{HK}}$ for $\cos\omega_{\mathbf{HK}}$ and $\sin\omega_{\mathbf{HK}}$, respectively, together with known weights $A_{\mathbf{HK}}$. Furthermore, in sharp contrast to the use of the expected values (I_1/I_0 alone) used in (8), the following minimal function uses the known probabilistic estimates of both cosines and sines with associated (known) weights $A_{\mathbf{HK}}$ (Hauptman, 1982; Giacovazzo, 1983):

$$R(\varphi) = \{ \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} [(\sin T_{\mathbf{HK}} - s_{\mathbf{HK}})^2 + (\cos T_{\mathbf{HK}} - c_{\mathbf{HK}})^2] \} / \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \quad (14)$$

where $T_{\mathbf{HK}}$ is defined by (9). One anticipates that those phases are correct which minimize $R(\varphi)$. In contrast to the 'pure' direct-methods approach (not requiring anomalous intensity data) of Weeks *et al.*, which seeks the global minimum of (8) subject to the equations of constraint (7), no attempt is made here to impose these constraints. In view of the work of Weeks *et al.*, it is presumed that the ability to impose these constraints would improve the phase determination, but how to do this for a macromolecule when only a limited low-resolution data set is available, is an open question. This method requires approximate initial values of the individual phases and these can be obtained from the atomic position(s) of a presumed known partial structure or of the anomalous scatterer(s). The tests of this method are based on the same data for cytochrome c_{550} as used previously. Initial estimates of the phases were simply taken to be equal to the phases calculated from the Pt position which was assumed to be known. Three test results were generated as shown in Table 3. All the iterative full-matrix least squares were carried out for 20 cycles. As seen from Table 3, the large redundancy (50–360) makes this method so powerful that it reduces the average triplet error $\langle T_{\text{error}} \rangle$ dramatically and, at the same time, the average phase error was reduced by 25–30°. Comparison of Tables 2 and 3, in particular the final average phase error of 32.8° for the 503 phases of Table 3 compared with the average error of 51.9° for the 250 phases of Table 2, clearly shows the superior performance of the minimal function approach (14).

Table 3. *Average initial and final triplet and phase errors for three different sets of phases based on the presumed known* Pt *position and using the minimal function* (14)

Data No.	No. of reflections	No. of triplets	$\langle \lvert T_{\text{error}} \rvert \rangle$ (°) Initial	$\langle \lvert T_{\text{error}} \rvert \rangle$ (°) Final	$\langle \lvert \varphi_{\text{error}} \rvert \rangle$ (°) Initial	$\langle \lvert \varphi_{\text{error}} \rvert \rangle$ (°) Final
1	503	25244	79.0	44.1	64.1	32.8
2	994	185264	76.5	49.7	66.3	37.2
3	1398	504918	78.6	54.2	68.8	42.5

Table 4. *Average initial and final triplet and phase errors for* 503 *phases and* 25 244 *triplets based on presumed known light-atom fragments of cytochrome* c_{550} *having the number of atoms shown in column* 1 *and employing the minimal function* (14)

No. of known atoms	$\langle \lvert T_{\text{error}} \rvert \rangle$ (°) Initial	$\langle \lvert T_{\text{error}} \rvert \rangle$ (°) Final	$\langle \lvert \varphi_{\text{error}} \rvert \rangle$ (°) Initial	$\langle \lvert \varphi_{\text{error}} \rvert \rangle$ (°) Final
400	70.6	57.8	39.8	32.5
300	75.8	57.9	46.4	32.6
200	81.6	58.2	53.1	32.8
100	85.5	58.2	64.1	32.8
75	87.7	57.8	68.5	32.6
60	87.8	71.1	70.8	46.0
50	88.6	75.4	73.4	54.3

To investigate the potential use of this method for phase refinement, several tests with different starting sets of phases were also carried out for the 503 reflections and 25 244 triplets data set. The results are listed in Table 4. All refinements used 20 cycles of full-matrix least squares. The results show the very strong convergence feature of this method even for presumed known all-light-atom fragments consisting of as few as 75 atoms.

7. Discussion

The number of structure invariants increases dramatically as the number of reflections becomes large. Thus 503 reflections generate 25 244 structure invariants while 1398 phases generate 504 918. The largest data set tested consisted of 1979 phases, and the resulting 1979 × 1979 full-matrix least-squares calculation involved 1 488 228 structure invariants. Clearly, enormous computing power is required and a huge memory is essential for the successful application of the methods described here. Another important factor is the requirement of an accurate anomalous-scattering data set which is needed to generate a sufficient number of well estimated structure invariants.

Phase refinement is the most time consuming part of a protein structure determination procedure. Our test results show that the methods described here have the potential to refine the phases from a set of initial values.

This research was supported in part by the National Science Foundation Grant No. CHE-8822296 and the National Institutes of Health Grant No. 1P01 GM46733-01.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bijvoet, J. M. (1949). *Proc. K. Ned. Acad. Wet. Ser. B*, **52**, 313–314.

Fortier, S. (1991). *Direct Methods of Solving Crystal Structures*. In *Proceedings of the NATO ASI* (*Erice*, 1990), pp. 311–333. New York: Plenum Press.

Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 585–592.

Han, F., DeTitta, G. & Hauptman, H. (1991). *Acta Cryst.* A**47**, 484–490.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 632–641.

Heinerman, J. J. L., Krabbendam, H., Kroon, J. & Spek, A. L. (1978). *Acta Cryst.* A**34**, 447–450.

Hendrickson, W. A., Smith, J. L., Phizackerley, R. P. & Merritt, E. A. (1988). *Proteins*, **4**, 77–78.

Karle, J. (1984). *Acta Cryst.* A**40**, 374–379.

Krabbendam, H. (1991). *Direct Methods of Solving Crystal Structures*. In *Proceedings of the NATO ASI* (*Erice*, 1990), pp. 335–351. New York: Plenum Press.

Kroon, J., Spek, A. L. & Krabbendam, H. (1977). *Acta Cryst.* A**33**, 382–385.

Peschar, R. & Schenk, H. (1987). *Acta Cryst.* A**43**, 513–522.

Timkovich, R. & Dickerson, R. E. (1973). *J. Mol. Biol.* **79**, 39–56.

Timkovich, R. & Dickerson, R. E. (1976). *J. Biol. Chem.* **251**, 4033–4046.

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

Weeks, C., DeTitta, G., Miller, R. & Hauptman, H. (1993). *Acta Cryst.* D**49**, 179–181.

Woolfson, M. M. (1977). *Acta Cryst.* A**33**, 219–225.