Acta Cryst. (1985). A41, 571-577

A Direct-Methods Solution to the Phase Problem in the Single Isomorphous Replacement Case: Theoretical Basis and Initial Applications

BY SUZANNE FORTIER, NANCY J. MOORE AND MARIE E. FRASER

Department of Chemistry, Queen's University, Kingston, Canada K7L 3N6

(*Received* 25 *March* 1985; *accepted* 31 *May* 1985)

Abstract

The probabilistic theory of the three-phase structure invariants for a pair of isomorphous structures [Hauptman (1982). *Acta Cryst.* A38, 289-294] is re-examined. The analysis leads to distributions capable of estimating cosine invariants in the full range of -1 to $+1$. In particular, it is shown that heavy-atom substructure information can be incorporated easily into the distributions. The initial applications, using calculated diffraction data from the protein cytochrome c_{550}, $M_R \simeq 14\,500$, and its $PtCl_4^{2-}$ derivative show that a remarkable increase in accuracy results from the use of the revised distributions, particularly after the incorporation of heavy-atom substructure information. Finally, it is shown that in the individual phase determinations the redundant cosine invariants play a role identical to that of the multiple isomorphous derivatives and thus provide the basis for the solution of the phase problem in the single isomorphous replacement case.

1. Introduction

Because of their obvious benefits, macromolecular phasing methods based on the single isomorphous replacement experiment (hereinafter referred to as SIR) have been investigated by several researchers. In particular, approaches based on the combination of direct methods and SIR have been investigated ever since the birth of direct methods as described by Fan Hai-fu, Han Fu-son, Qian Jin-zi and Yao Jia-xing (1984). More recently, the probabilistic (Hauptman, 1982) and algebraic (Karle, 1983) bases for the estimation of the three-phase structure invariants for a pair of isomorphous structures were introduced. The theoretical work and the initial applications (Hauptman, Potter & Weeks, 1982; Karle, 1983) clearly showed the promising potential of the fused direct-methods–SIR approach. In particular, Hauptman's formula, when applied to error-free diffraction data from the protein cytochrome c_{550} and a single heavy-atom derivative, proved capable of yielding several tens of thousands of reliably estimated three-phase invariants having the extreme values of 0 or π. It seemed natural to anticipate that the standard and well tested machinery of direct methods - convergence mapping, multisolution approach, tangent refinement - could then be used in an automated fashion to determine the individual phases, thus solving the phase problem in the SIR case.

In the applications, however, a few troublesome anomalies were detected. In particular, in the tangent refinement, it was observed that large numbers of invariants used in the individual phase determinations did not show a random error distribution. Surprisingly, the errors were strongly biased both in signs and magnitudes. While this problem appeared at first to impose severe limitations on the technique, it also suggested that the power of the fused direct-methods–SIR approach had not yet been fully exploited. Specifically, it suggested the presence of systematic errors in the procedure used. Once these errors were characterized and corrected, it could then be anticipated that extremely accurate estimates of the invariants, and the individual phases, would be possible.

In the present paper, the three-phase invariant conditional probability distribution of Hauptman (1982) is re-examined. The analysis not only identifies the source of the systematic errors but also shows that the errors can be eliminated by a simple alteration of the original distribution. Furthermore, the analysis shows that heavy-atom substructure information can be incorporated easily into the distributions. As a result of the alteration, cosine invariants in the full -1 to $+1$ range can be obtained with unprecedented accuracy, especially after the incorporation of heavy-atom substructure information into the distribution.

A two-step phasing procedure is proposed. In the first step, Hauptman's formula is used for the determination of the heavy-atom substructure. In the second step, the additional information from the heavy-atom substructure is actively used to estimate the cosine invariants in the full range of -1 to $+1$. While the cosine invariants normally yield a twofold ambiguity in the phases, it is shown that the ambiguity is resolved by the redundancy of the invariants used in individual phase determinations. In fact, redundant cosine invariants play a role identical to that of the multiple isomorphous derivatives.

0108-7673/85/060571-07$01.50

2. Theoretical basis

2.1. *The three-phase invariant conditional probability distributions for a pair of isomorphous structures*

For each reciprocal-lattice vector **H**, there exist two normalized structure factors $E_{\mathbf{H}}$ and $G_{\mathbf{H}}$. For a triplet of reciprocal-lattice vectors **H**, **K**, **L** satisfying **H** + **K** + **L** = 0, there exist eight structure invariants

$$\begin{aligned}
\omega_1 &= \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}},\\
\omega_2 &= \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}},\\
\omega_3 &= \varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \varphi_{\mathbf{L}},\\
\omega_4 &= \psi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}},\\
\omega_5 &= \varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}},\\
\omega_6 &= \psi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}},\\
\omega_7 &= \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \varphi_{\mathbf{L}},\\
\omega_8 &= \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}},
\end{aligned} \tag{1}$$

where the φ's and the ψ's are the phases associated with the isomorphous pair of structures.

Let

$$\begin{aligned}
|E_{\mathbf{H}}| &= R_1, \quad |E_{\mathbf{K}}| = R_2, \quad |E_{\mathbf{L}}| = R_3;\\
|G_{\mathbf{H}}| &= S_1, \quad |G_{\mathbf{K}}| = S_2, \quad |G_{\mathbf{L}}| = S_3.
\end{aligned} \tag{2}$$

The conditional probability distributions of the three-phase structure invariants ω_i given the six magnitudes $|E_{\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{\mathbf{L}}|, |G_{\mathbf{H}}|, |G_{\mathbf{K}}|, |G_{\mathbf{L}}|$ in their first neighborhood are given by

$$P_i(\Omega_i|R_1, R_2, R_3, S_1, S_2, S_3) \simeq (1/K_i)\exp(A_i\cos\Omega_i), \quad i = 1, 2, \dots, 8, \tag{3}$$

where

$$K_i = 2\pi I_0(A_i) \tag{4}$$

and I_0 is the modified Bessel function (Hauptman, 1982). The A_i values are given by

$$\begin{aligned}
A_i = 2\{&\beta_1\tau_1R_1R_2R_3\\
&+\beta_2[\tau_{21}R_1R_2S_3 + \tau_{22}R_1S_2R_3 + \tau_{23}S_1R_2R_3]\\
&+\beta_3[\tau_{31}R_1S_2S_3 + \tau_{32}S_1R_2S_3 + \tau_{33}S_1S_2R_3]\\
&+\beta_4\tau_4S_1S_2S_3\},
\end{aligned} \tag{5}$$

where the β's are functions of the atomic scattering factors, and $\tau = C_1C_2C_3$ is obtained by comparing the ith structure factor associated with the coefficient of τ with the ith structure factor associated with the invariant. If they are of the same type, *i.e.* both R or both S, then $C_i = 1{\cdot}0$, $i = 1, 2, 3$. If one is of type R and the other of type S, then

$$C_i = I_1(2\gamma R_iS_i)/I_0(2\gamma R_iS_i), \quad i = 1, 2, 3, \tag{6}$$

where I_1 and I_0 are the modified Bessel functions and

$$\gamma \simeq 1/2[1 + 4/(\text{diffraction ratio})^2] \tag{7}$$

for the special case of a native protein and a heavy-atom derivative (Fortier, Weeks & Hauptman, 1984).

The Bessel-function ratio $I_1(2\gamma R_iS_i)/I_0(2\gamma R_iS_i)$ is the expected value of the cosine of the phase difference $(\psi_i - \varphi_i)$ associated with the two magnitudes R_i and S_i (Sim, 1960). The equivalent cosine functions can therefore be substituted for the τ functions in the distribution. Let

$$\psi_i - \varphi_i = \pm\alpha_i, \tag{8}$$

then, for example, the conditional probability distribution of the ω_1 invariant can be written as

$$P_1(\Omega_1) \simeq (1/K_1)\exp(A_1\cos\Omega_1), \tag{9}$$

where

$$\begin{aligned}
A_1\cos\Omega_1 = 2\cos\Omega_1\{&\beta_1R_1R_2R_3 + \beta_2[R_1R_2S_3\cos\alpha_3\\
&+ R_1S_2R_3\cos\alpha_2 + S_1R_2R_3\cos\alpha_1]\\
&+\beta_3[R_1S_2S_3\cos\alpha_2\cos\alpha_3\\
&+S_1R_2S_3\cos\alpha_1\cos\alpha_3\\
&+S_1S_2R_3\cos\alpha_1\cos\alpha_2]\\
&+\beta_4S_1S_2S_3\cos\alpha_1\cos\alpha_2\cos\alpha_3\}.
\end{aligned} \tag{10}$$

Similar expressions are obtained for each of the ω_i invariants. For the sake of simplicity, let us assume that $\alpha_2 = \alpha_3 = 0$ and $|\alpha_1| \neq 0$. Equation (10) then becomes

$$\begin{aligned}
A_1\cos\Omega_1 = 2\cos\Omega_1\{&\beta_1R_1R_2R_3\\
&+\beta_2[R_1R_2S_3 + R_1S_2R_3]\\
&+\beta_3R_1S_2S_3\} + \{\cos(\Omega_1 + \alpha_1)\\
&+\cos(\Omega_1 - \alpha_1)\} \times \{\beta_2S_1R_2R_3\\
&+\beta_3[S_1R_2S_3 + S_1S_2R_3] + \beta_4S_1S_2S_3\}.
\end{aligned} \tag{11}$$

While the expected value of the phase difference $\psi_1 - \varphi_1$ can be estimated, its sign is not known. In the form of the distribution shown in (11), both signs are considered equally probable and their contributions are averaged, as is done in the standard SIR technique (Blow & Rossmann, 1961). The consequence of averaging the two possible values is not only the restriction of cosine estimates to one or other of the extreme values, +1 or −1, but also the introduction of the systematic errors mentioned in the *Introduction.*

2.2. *Estimating cosine invariants in the full range of* −1 *to* +1

We continue to consider the special case that $\alpha_2 = \alpha_3 = 0$ and $|\alpha_1| \neq 0$. Again the two possible signs of α_1 are considered equally probable. However, instead of averaging out their contributions as in (11), the $A_1\cos\Omega_1$ expression is calculated twice, assuming

Table 1. *Examples of estimated three-phase cosine invariants with values different from* -1 *or* $+1$ *for cytochrome* c_{550} *and its* PtCl_4^{2-} *derivative*

$\lvert E_{\mathbf{H}}\rvert$	$\lvert E_{\mathbf{K}}\rvert$	$\lvert E_{\mathbf{L}}\rvert$	$\lvert G_{\mathbf{H}}\rvert$	$\lvert G_{\mathbf{K}}\rvert$	$\lvert G_{\mathbf{L}}\rvert$	$\lvert\alpha_{\mathbf{H}}\rvert$	$\lvert\alpha_{\mathbf{K}}\rvert$	$\lvert\alpha_{\mathbf{L}}\rvert$	Invariant type*	Calc. cos	True cos
2·2	1·3	1·2	2·7	1·8	1·8	3·9	16·2	2·8	$\psi\varphi\psi$	0·69	0·68
1·9	1·9	1·6	2·4	2·4	1·1	0·9	1·5	24·9	$\varphi\varphi\varphi$	−0·71	−0·69
4·8	1·5	1·3	5·2	2·0	1·4	0·8	4·6	27·3	$\psi\psi\varphi$	0·15	0·15
1·9	1·9	1·7	1·3	2·4	1·0	17·6	0·9	3·7	$\psi\varphi\varphi$	0·59	0·61
2·6	1·5	1·2	1·8	2·0	1·5	0·8	4·6	27·0	$\varphi\psi\varphi$	−0·24	−0·26

* The entry $\psi\varphi\psi$, for example, means $\psi_{\mathbf{H}}+\varphi_{\mathbf{K}}+\psi_{\mathbf{L}}$, *etc.*

first the one sign and then the other, *i.e.*

$$\begin{aligned}A_1\cos\Omega_1=2\cos\Omega_1\{&\beta_1R_1R_2R_3\\&+\beta_2[R_1R_2S_3+R_1S_2R_3]\\&+\beta_3R_1S_2S_3\}+2\cos(\Omega_1+\alpha_1)\{\beta_2S_1R_2R_3\\&+\beta_3[S_1R_2S_3+S_1S_2R_3]+\beta_4S_1S_2S_3\}\end{aligned}\quad(12)$$

and

$$\begin{aligned}A_1\cos\Omega_1=2\cos\Omega_1\{&\beta_1R_1R_2R_3\\&+\beta_2[R_1R_2S_3+R_1S_2R_3]\\&+\beta_3R_1S_2S_3\}+2\cos(\Omega_1-\alpha_1)\{\beta_2S_1R_2R_3\\&+\beta_3[S_1R_2S_3+S_1S_2R_3]+\beta_4S_1S_2S_3\}.\end{aligned}\quad(13)$$

It is easily seen that (12) and (13) can yield estimates of Ω_1 ranging over the full 0 to 360° interval. Furthermore, it is clear that the two sign possibilities yield enantiomorphic estimates of Ω_1, and thus determine uniquely the cosine of the invariant.

In Table 1 a few examples are used to demonstrate that the mode of the distribution can be significantly different from 0 or 180°, even when the α magnitude is relatively small. Comparison of (11) with (12) and (13) shows that (11) does not yield the mode of the distribution, with its associated A value, but rather the A value at the 0° (or 180°) angle, as depicted in Fig. 1. This explains the extremely good correlation between averaged A magnitudes and averaged error magnitudes obtained in the extensive calculations of Hauptman, Potter & Weeks (1982). Systematic deviations from the 0° (or 180°) estimates are indeed accounted for in the original distribution. They result in a lowering of the A value or an increase in the variance.

The advantages of using (12) or (13) over (11) are considerable. With (12) or (13) one can estimate the magnitude of the deviation from 0° (or 180°) or, more specifically, one can estimate the value of the cosine invariant; with (11), the magnitudes of the systematic deviations cannot be estimated, rather they are buried in the A values.

The Bessel-function ratio (6) can be used to estimate the magnitude of the phase difference, α_i, appearing in (12) and (13) but the variance of the estimate may be large and thus the estimate must be used with caution. However, once the heavy-atom substructure has been determined, the cos α_i value can be calculated from the structure-factor magnitudes according to

$$\cos\alpha_i=(F^2_{PH_i}+F^2_{P_i}-F^2_{H_i})/2F_{PH_i}F_{P_i},\quad(14)$$

where F_{PH_i}, F_{P_i} and F_{H_i} are the ith-reflection structure-factor magnitudes of the heavy-atom derivative, native and heavy-atom substructure, respectively.

Equations (12) and (13) can be extended to the general case where $|\alpha_1|$, $|\alpha_2|$, and $|\alpha_3|$ are all non zero. Again the magnitudes of the phase differences can be calculated while the signs are unknown. There exist, therefore, eight possible sign combinations. Calculation of the distribution for each of the eight sign combinations yields four enantiomorphic pairs of Ω_1 estimates or four cosine-invariant estimates. The cosine invariant is thus estimated as the weighted average

$$\cos\Omega_{\mathrm{Av}}=\sum_i A_i\cos\Omega_i/\sum_i A_i,\quad i=1,2,3,4,\quad(15)$$

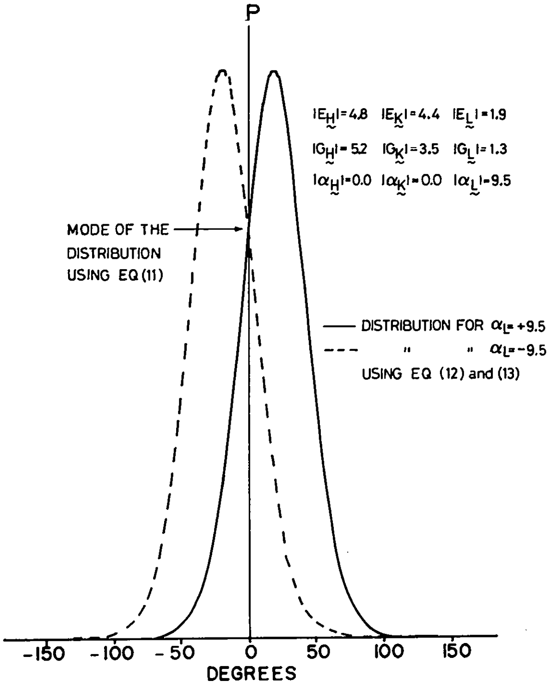

Fig. 1. Probability distributions of a three-phase structure invariant calculated using (11), (12) and (13).

Table 2. *Average magnitude of the error in estimated values of* 25 000 *three-phase cosine invariants for cytochrome* c_{550} *and its* $\mathrm{PtCl_4^{2-}}$ *derivative*

				Number of invariants with errors					
Top	Protocol	Average $\lvert A\rvert$	Average $\lvert$error$\rvert$	GE* 15°	GE 30°	GE 45°	GE 60°	GE 75°	GE 90°
1000	1	6·3	9·1	198	30	1	0	0	0
	2	6·1	16·5	543	66	0	0	0	0
	3	5·8	15·1	398	164	36	12	0	0
5000	1	5·2	10·5	1222	246	17	0	0	0
	2	5·0	18·0	2821	734	20	8	8	8
	3	4·7	19·4	2449	1232	466	135	36	18
10 000	1	4·6	12·0	2745	770	115	20	16	16
	2	4·4	18·9	5719	1900	111	31	24	24
	3	4·1	22·9	5598	3219	1409	496	145	49
15 000	1	4·2	12·8	4576	1368	246	31	24	24
	2	4·0	19·5	8715	3274	262	44	25	24
	3	3·7	25·1	8900	5463	2649	1051	357	102
20 000	1	3·9	13·4	6600	2069	397	40	24	24
	2	3·7	19·9	11745	4736	499	74	30	25
	3	3·4	27·0	12416	7949	4123	1772	635	205
25 000	1	3·6	14·0	8706	2886	571	72	41	40
	2	3·5	20·5	14938	6283	810	131	53	42
	3	3·2	28·6	16030	10544	5813	2706	1013	328

Protocol 1: Cosine estimates in the full −1 to +1 range (§ 2.2) were calculated using the calculated magnitudes of the phase differences (equation 14).
Protocol 2: Cosine estimates in the full −1 to +1 range (§ 2.2) were calculated using the estimated magnitudes of the phase differences (equation 6).
Protocol 3: Cosine estimates limited to the two extreme values +1 and −1 were calculated (equation 3-5).

* GE = greater than or equal to.

and the associated A magnitude is taken as the average of the four A magnitudes times a weighting function whose value depends on the maximum difference between the average cosine and the four individual estimates. *i.e.*

$$A_{\mathrm{Av\ weighted}} = \tfrac{1}{4}\sum A_i \times \cos(\max|\Omega_{\mathrm{Av}} - \Omega_i|), \qquad i = 1, 2, 3, 4. \quad (16)$$

In general, not all three α's are significantly different from zero. Therefore, although four possible cosine invariants are estimated, they do not in most cases differ much from one another. Finally, we note that the present method is applicable, whether the heavy atoms form a non-centrosymmetric array or not.

3. The applications

The procedure described in the previous section for the estimation of cosine invariants was tested on the protein cytochrome c_{550} from *Parococcus Denitrificans*, molecular weight $M_r \simeq 14\,500$, space group $P2_12_12_1$, and a single $\mathrm{PtCl_4^{2-}}$ isomorphous derivative (Timkovich & Dickerson, 1973, 1976). This protein and its $\mathrm{PtCl_4^{2-}}$ derivative have been used by other workers for tests on integrated direct-methods - SIR phasing techniques (Hauptman, Potter & Weeks, 1982; Karle, 1983) and thus provide a good basis for comparisons. Coordinates were obtained from the Protein Data Bank (Bernstein *et al.*, 1977) and used to calculate structure factors and normalized structure factors to a resolution of 2·5 Å (4159 E's and 4159 G's). The phases φ corresponding to the 1000 largest $|E|$'s of the native protein and the phases ψ corresponding to the 1000 largest $|G|$'s of the derivative were used to generate the three-phase structure invariants. The cosine invariants were estimated according to three protocols:*

Protocol 1: Cosine estimates in the full −1 to +1 range (§ 2.2) were calculated using the calculated magnitudes of the phase differences (equation 14).

Protocol 2: Cosine estimates in the full −1 to +1 range (§ 2.2) were calculated using the estimated magnitudes of the phase differences (equation 6).

Protocol 3: Cosine estimates limited to the two extreme values +1 and −1 were calculated (equations 3-5).

The invariants were ranked in descending order of A, and the 25 000 largest A invariants obtained in protocol 3 (which are not necessarily the largest obtained in protocols 1 and 2) were used for the comparison of the three protocols. In Table 2 are summarized the results obtained for these 25 000 invariants. These results clearly confirm the validity of the theoretical basis described in § 2.2. They show very convincingly that taking advantage of the information that becomes available, once the heavy-atom substructure has been determined, results in a considerable gain in the accuracy of the cosine-invariant estimates. The average error over the 25 000

* The calculations were done on a 16-bit PDP11/23 computer. The programs used were written by S. A. Potter, C. M. Weeks and G. D. Smith of the Medical Foundation of Buffalo, Inc., and adapted by N. J. Moore.

Table 3. *A representative sample of* 20 *three-phase cosine invariants for cytochrome* c_{550} *and its* $PtCl_4^{2-}$ *derivative*

Serial no. of invariant	A	Calc. cos	True cos	|Error| (°)
100	7·02	0·96	0·98	4·8
200	6·61	0·76	0·79	2·7
300	6·42	0·96	0·98	4·8
400	6·26	0·23	0·32	5·4
500	6·14	0·43	0·34	5·6
600	6·07	0·90	0·93	4·3
700	5·99	0·95	0·83	15·7
800	5·93	0·99	1·00	8·1
900	5·86	0·88	0·84	4·5
1000	5·78	0·71	0·88	16·4
1100	5·71	0·92	0·81	12·8
1200	5·65	0·54	0·70	11·7
1300	5·59	0·89	0·80	9·7
1400	5·54	−0·98	−1·00	11·5
1500	5·50	0·77	0·60	13·5
1600	5·44	0·77	0·84	6·8
1700	5·39	−0·95	−1·00	18·2
1800	5·34	0·98	0·94	8·5
1900	5·30	0·89	0·85	4·7
2000	5·26	0·98	0·96	4·8

Table 4. *A representative family of invariants with cosines estimated according to protocols* 1 *and* 3 *for cytochrome* c_{550} *and its* $PtCl_4^{2-}$ *derivative*

$|E_H|=1{\cdot}5$, $|E_K|=1{\cdot}2$, $|E_L|=1{\cdot}1$, $|G_H|=1{\cdot}9$, $|G_K|=1{\cdot}8$, $|G_L|=1{\cdot}8$.

	Protocol 1		Protocol 3		
Invariant type	A	Calc. cos	A	Calc. cos	True cos
$\varphi\varphi\varphi$	6·85	0·54	4·40	1·00	0·59
$\varphi\varphi\psi$	6·88	0·54	4·33	1·00	0·56
$\varphi\psi\varphi$	6·88	0·54	4·34	1·00	0·56
$\varphi\psi\psi$	6·91	0·54	4·66	1·00	0·54
$\psi\varphi\varphi$	6·84	0·77	4·42	1·00	0·81
$\psi\varphi\psi$	6·88	0·77	4·75	1·00	0·79
$\psi\psi\varphi$	6·88	0·77	4·76	1·00	0·79
$\psi\psi\psi$	6·91	0·77	5·10	1·00	0·77

invariants, when estimated using protocol 1, has decreased by a factor of two, as compared to the average error of the same set of invariants estimated using protocol 3. More importantly, though, the number of invariants for which calculated and true cosines differ by more than 45° decreased considerably, both when the true and when the estimated differences were used. Even when the phase differences are estimated using protocol 2, a substantial gain in accuracy is obtained.

In this test calculation, protocol 1 yields errors smaller than 45° for 97·7% of the invariants and errors smaller than 30° for 88·5% of the invariants; the accuracy of these results is unprecedented. As expected, the average A magnitude decreases from protocol 1 to 3. As was explained in § 2.2, this is a result of the fact that protocol 3 determines the A value at the 0° (or 180°) angle rather than the A value at the mode of the distribution. Table 3 shows a representative sample of 20 three-phase structure invariants taken from the 2000 largest A-value invariants estimated using protocol 1. It is seen that accurate estimates can be made in the full range of −1 to +1. The ability to estimate reliably cosine invariants with values significantly different from −1 or +1 obviates problems of enantiomorph discrimination. In addition, it is evident from Table 2 that the number of invariants determined with sufficient accuracy to be used in phasing procedures has by no means been exhausted. Therefore, we anticipate that the use of the heavy-atom substructure information will not only yield an increase in accuracy but also, and equally important, an increase in the number of accessible phases. It should be noted that heavy-atom substructure information can also be incorporated into the algebraic approach formulae (Karle, 1983, equation 12), again allowing cosine-invariant estimates different from −1 to +1. Test calculations based on this approach have not been reported yet and thus cannot be compared with the present results.

4. Proposed phasing procedure

4.1. *The role of the cosine invariants*

The role of the cosine invariants in the phasing of SIR data can be understood easily when their outcomes are compared to those obtained from a distribution capable of yielding only estimates of cosines having the extreme values of +1 or −1. For a triplet of reciprocal-lattice vectors **H**, **K**, **L** satisfying **H** + **K** + **L** = 0, there exist eight structure invariants formed between the φ and ψ phases associated with the isomorphous pair of structures, and defined in (1). When the distribution limited to estimates of +1 or −1 is used, the invariants belonging to a common **H**, **K**, **L** family generally all have the same cosine estimate (+1 or −1), although their A values may be different in magnitude as shown in Table 4. In fact the only cases for which members of a common family can have different estimates are those for which the A values are extremely small, and for that reason are of limited use in the determination of the individual phases. It can therefore be predicted that tangent refinement, when applied to these families of invariants, will tend to force native protein and derivative phases to be of equal value. This would not be a serious problem if the output phases were an average of their corresponding protein and derivative phases. However, there is no guarantee that this will be the case. In particular, for a fixed enantiomorph, if a given invariant has an associated heavy-atom substructure invariant with a value of 0 or 180°, then for every solution in the individual phases there exists a second solution whose phases are enantiomorphic to the first triplet of phases, with respect to the heavy-atom substructure invariant.

Let φ_i and ζ_i be the ith phases of the native and heavy-atom substructure, respectively, and let

$$\begin{aligned}\varphi_{\mathbf{H}} &= \zeta_{\mathbf{H}} + \beta_{\mathbf{H}},\\ \varphi_{\mathbf{K}} &= \zeta_{\mathbf{K}} + \beta_{\mathbf{K}},\\ \varphi_{\mathbf{L}} &= \zeta_{\mathbf{L}} + \beta_{\mathbf{L}},\end{aligned} \tag{17}$$

then

$$\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}} = \zeta_{\mathbf{H}} + \zeta_{\mathbf{K}} + \zeta_{\mathbf{L}} + \beta_{\mathbf{H}} + \beta_{\mathbf{K}} + \beta_{\mathbf{L}}. \tag{18}$$

If

$$\zeta_{\mathbf{H}} + \zeta_{\mathbf{K}} + \zeta_{\mathbf{L}} = 0 \text{ or } 180^\circ,$$

then

$$\beta_{\mathbf{H}} + \beta_{\mathbf{K}} + \beta_{\mathbf{L}} = 0 \text{ or } 180^\circ$$

and

$$\beta_{\mathbf{H}} + \beta_{\mathbf{K}} + \beta_{\mathbf{L}} = -\beta_{\mathbf{H}} - \beta_{\mathbf{K}} - \beta_{\mathbf{L}}.$$

Thus

$$\begin{aligned}\varphi_{\mathbf{H}} &= \zeta_{\mathbf{H}} - \beta_{\mathbf{H}},\\ \varphi_{\mathbf{K}} &= \zeta_{\mathbf{K}} - \beta_{\mathbf{K}},\\ \varphi_{\mathbf{L}} &= \zeta_{\mathbf{L}} - \beta_{\mathbf{L}}\end{aligned} \tag{19}$$

is also a solution. Therefore, unless the basis set contains phases with values significantly different from those of the heavy-atom substructure, the tangent refinement will tend to converge to the heavy-atom substructure phases. This problem is well known in small-molecule applications and has been observed recently in macromolecule applications by Xu *et al.* (1984). It is in every way similar to the problem of loss of enantiomorph discrimination. It can be quite serious, since the distribution tends to yield large A-value-invariant estimates when the product $(S_{\mathbf{H}} - R_{\mathbf{H}})(S_{\mathbf{K}} - R_{\mathbf{K}})(S_{\mathbf{L}} - R_{\mathbf{L}})$ is large (Fortier, Weeks & Hauptman, 1984). These are precisely the invariants for which the associated heavy-atom substructure invariants are most likely to be 0 or 180°. Clearly, this problem disappears when cosine estimates in the full range of -1 to $+1$ are used.

4.2. *Individual phase determination: analogy between the role of the cosine invariants and the role of the multiple isomorphous derivatives.*

The cosine invariant yields enantiomorphic estimates of the invariant angle, Ω and $-\Omega$. Therefore, when the invariant is used to determine the value of an individual phase, two possible values for the phase, differing by 2Ω, will be obtained. The ambiguity is easily resolved, however, by the redundancy of invariants used in the individual phase determinations, in a manner that is analogous to that of the multiple isomorphous replacement method.

As an example let us consider the following invariants taken from the 25 000 invariant set described in § 3. We have

	calc. cos	true cos
$\cos(\varphi_{179} - \varphi_{444} - \varphi_{641} + 180^\circ)$	−0·39	−0·38
$\cos(\varphi_{153} - \varphi_{179} - \varphi_{247})$	0·80	0·82
$\cos(\varphi_{153} - \varphi_{179} - \varphi_{641} + 180^\circ)$	0·80	0·60.

We seek to determine the value of φ_{179}, assuming as known the values of the remaining phases in these three invariants:

$$\varphi_{153} = 44^\circ \qquad \varphi_{247} = -90^\circ$$
$$\varphi_{444} = -58^\circ \qquad \varphi_{641} = -90^\circ.$$

Using the first invariant, we have

$$\cos(\varphi_{179} - \varphi_{444} - \varphi_{641} + 180^\circ) = -0{\cdot}39$$
$$\varphi_{179} = \varphi_{444} + \varphi_{641} + 180^\circ \pm 113^\circ$$

and

$$\varphi_{179} = 145 \text{ or } 279^\circ.$$

Using the second invariant gives

$$\cos(\varphi_{153} - \varphi_{179} + \varphi_{247}) = 0{\cdot}80$$
$$\varphi_{179} = \varphi_{153} + \varphi_{247} \pm 37^\circ$$

and

$$\varphi_{179} = 351 \text{ or } 277^\circ.$$

Thus the ambiguity in the phase φ_{179} is resolved and $\varphi_{179} \simeq 278^\circ$. The third invariant serves to confirm the previous choice, as a third derivative would:

$$\cos(\varphi_{153} + \varphi_{179} + \varphi_{641} + 180^\circ) = 0{\cdot}80$$
$$\varphi_{179} = -\varphi_{153} - \varphi_{641} + 180^\circ \pm 37^\circ$$

and

$$\varphi_{179} = 263 \text{ or } 189^\circ.$$

Since the cosine invariants mimic the role played by the multiple isomorphous derivatives, the determination of a phase by n invariant contributors is equivalent to the determination of a phase by n derivatives. As n normally tends to be large, very accurate estimates of the phases can be obtained, provided of course that the cosine invariants are determined with sufficient accuracy.

4.3. *From the cosine invariants to the individual phases*

With cosine invariants in hand, it is clear that tangent refinement is no longer the best tool available for the determination of the individual phases. Rather, a least-squares analysis of cosine invariants (Karle & Hauptman, 1957; Hauptman, 1972) can be used to evaluate the individual phases. In particular, the value of an individual phase $\varphi_{\mathbf{h}}$ is estimated as the value that minimizes the function

$$\varphi = \sum_{\mathbf{k}} w_{\mathbf{k}}[\cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}) - c_{\mathbf{k}}]^2 / \sum_{\mathbf{k}} w_{\mathbf{k}},$$

where the $c_{\mathbf{k}}$'s are the cosine estimates and the weights, $w_{\mathbf{k}}$'s, are a function of the A values. Details of this procedure can be found in Hauptman (1972).

4.4. *A two-step phasing procedure*

The initial application of Hauptman's distribution was carried to tangent refinement and computation of density maps, and showed that the technique yields accurate heavy-atom substructure information (Weeks, Potter, Smith, Hauptman & Fortier, 1984). Furthermore, the technique is applicable at fairly low resolution (4 Å) and to multiple-site as well as single-site derivatives. On the other hand, the present work shows that the introduction of heavy-atom substructure information in the distribution results in accurate cosine-invariant estimates in the full range of -1 to $+1$. A two-step phasing procedure naturally comes to mind. In the first step, Hauptman's distribution is used to determine the heavy-atom substructure. As was noted in § 4.1, tangent refinement, when applied to the invariants restricted to cosine estimates of $+1$ or -1, tends to converge to heavy-atom substructure phases. This can now be used to advantage, since the only information sought is that concerning the heavy-atom substructure. Once the heavy-atom substructure has been determined and refined, the magnitudes of the phase differences between the native and the derivative are calculated, and the cosine invariants are estimated according to the procedure described in § 2.2. The individual phases are then obtained by least-squares analysis of the cosine invariants. One advantage of the proposed two-step phasing procedure is due to the cyclic nature of the calculations. In the estimation of the cosine invariants, the generation of the three-phase invariants is by far the most lengthy section of the computation. In the scheme proposed, the invariants are generated in the first step only, and are then, in the second step, simply re-estimated.

5. Concluding remarks

The theoretical basis for the integration of direct methods and the SIR technique was introduced by Hauptman in 1982. Although the initial applications of this theory appeared very promising, several problems were detected when the technique was carried to the determination of the individual phases, and these limitations proved serious. In the present work, it has been shown that rather simple alterations of the original distribution, combined with the use of heavy-atom substructure information, eliminate systematic errors and, furthermore, yield estimates of the cosine invariants in the full -1 to $+1$ range. Of particular importance and interest is the analogy between multiple cosine invariants and multiple isomorphous derivatives in the determination of the individual phases.

In their 1982 publication, Hauptman, Potter & Weeks profferred the view that a combination of direct methods and single isomorphous replacement might well make possible unique macromolecular structure determination. The extension of Hauptman's theory presented in this paper not only supports this view but, in fact, provides the basis for the solution of the phase problem in the SIR case. Naturally, questions arise concerning effects of errors in the data and problems associated with structure-factor normalization. Any views on these matters are purely speculative at this point. However, it is certainly not unrealistic to predict that these problems will not prove insurmountable. The history of the traditional direct-methods applications serves as an example. Although most of the theory presently used was developed by the mid 1960's, the methods did not reach their full power until some ten or fifteen years later. While the power of present-day direct methods is largely a result of the substantial amount of work and expertise that provided the link between theory and applications, it is also a consequence of the tremendous increase in the accuracy of diffraction data. Owing to recent technological advances, more and more accurate macromolecular diffraction data can now be expected. Perhaps this time, theoretical and experimental advances are better synchronized.

We would like to thank S. A. Potter, C. M. Weeks and G. D. Smith for making their computer programs available to us and R. H. Blessing for his critical reading of the manuscript. Financial assistance from the Natural Sciences and Engineering Research Council of Canada and from Queen's University is gratefully acknowledged.

References

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

Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195–1202.

Fan Hai-fu, Han Fu-son, Qian Jin-zi & Yao Jia-xing (1984). *Acta Cryst.* A**40**, 489–495.

Fortier, S., Weeks, C. M. & Hauptman, H. (1984). *Acta Cryst.* A**40**, 544–548.

Hauptman, H. (1972). *Crystal Structure Determination: The Role of the Cosine Seminvariants.* New York and London: Plenum Press.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.

Hauptman, H., Potter, S. & Weeks, C. M. (1982). *Acta Cryst.* A**38**, 294–300.

Karle, J. (1983). *Acta Cryst.* A**39**, 800–805.

Karle, J. & Hauptman, H. (1957). *Acta Cryst.* **10**, 515–524.

Sim, G. A. (1960). *Acta Cryst.* **13**, 511–512.

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

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

Weeks, C. M., Potter, S. A., Smith, G. D., Hauptman, H. & Fortier, S. (1984). Proc. Am. Crystallogr. Assoc. Meet., 20–25 May 1984, Lexington, Kentucky, USA. Abstr. Q2.

Xu, Z. B., Yang, D. S. C., Furey, W. Jr, Sax, M., Rose, J. & Wang, B. C. (1984). Proc. Am. Crystallogr. Assoc. Meet., 20–25 May 1984, Lexington, Kentucky, USA. Abstr. PC2.