which is the fifth major result of this paper.

It is noteworthy that in the important special case that the f structure is a native protein and the g structure a heavy-atom isomorphous derivative then, on the basis of some preliminary calculations, it appears that

$$\beta_0 < 0, \quad \beta_1 > 0, \quad \beta_2 < 0, \quad \beta_3 > 0, \quad (3.25)$$

and that $|\beta_0|$, $|\beta_1|$, $|\beta_2|$, $|\beta_3|$ form an approximate arithmetic progression.

4. Concluding remarks

Recent advances in direct methods have been here integrated with the method of isomorphous replacement, and the probabilistic theory of the three-phase structure invariants for an isomorphous pair of structures has been worked out in some detail. The analysis includes the special case that one member of the pair is a native protein and the other member is a heavy-atom isomorphous derivative. A great deal of additional work remains to be done, *e.g.* the theory of the two-phase structure invariant and the higher-order structure invariants and seminvariants, the theory of structurally isomorphous triples, quartets, *etc.*, and the role of anomalous dispersion.

The initial applications of the work described here, using error-free data from a native protein and a single heavy-atom derivative, have been made, and these are presented in the following paper (Hauptman, Potter & Weeks, 1982). Although these initial results are very encouraging, attempts to apply these methods to the solution of unknown macromolecular structures must first overcome the obstacle presented by structures in which the number and occupancy factors of the heavy atoms may be unknown *a priori*. For such structures methods must be devised for estimating the parameters, β, β_0, β_1, β_2, and β_3 [equations (3.5)–(3.9)], which appear in the distributions (3.12), (3.16), (3.19), and (3.22), in terms of observed intensities only. Already (3.5) expresses β in terms of α, the square root of the correlation coefficient of the pair $|E_H|^2$, $|G_H|^2$. In a similar way it may be shown that β_0, β_1, β_2, and β_3, are likewise expressible in terms of the $|E_H|^2$ and the $|G_H|^2$ alone, but this work is outside the scope of the present paper and will be published separately.

Finally, the effect of errors inherent in experimental data as well as imperfect isomorphism has not been considered in this paper, the major purpose of which has been to formulate the basic theory. It is intended to present a study of the effect of errors in real data at a later date.

The research described here was supported by DHHS Grant No. GM26195 and Grant No. CHE-7911282 from the National Science Foundation, as well as a grant from the James H. Cummings Foundation, Inc. The author is indebted to Dr Suzanne Fortier who verified the lengthy calculations briefly summarized in this paper.

References

HAUPTMAN, H. (1975*a*). *Acta Cryst.* A**31**, 671–679.
HAUPTMAN, H. (1975*b*). *Acta Cryst.* A**31**, 680–687.
HAUPTMAN, H., POTTER, S. & WEEKS, C. (1982). *Acta Cryst.* A**38**, 294–300.
KARLE, J. & HAUPTMAN, H. (1958). *Acta Cryst.* **11**, 264–269.
SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography*. New York: Pergamon Press.
WATSON, G. N. (1958) *A Treatise on the Theory of Bessel Functions*. Cambridge Univ. Press.

Acta Cryst. (1982). A**38**, 294–300

On Integrating the Techniques of Direct Methods and Isomorphous Replacement. II. The First Applications

BY HERBERT HAUPTMAN, STEPHEN POTTER AND CHARLES M. WEEKS

Medical Foundation of Buffalo, Inc., 73 High Street, Buffalo, NY 14203, USA

(*Received* 18 *June* 1981; *accepted* 27 *October* 1981)

Abstract

With error-free diffraction data from the protein cytochrome c_{550} from *Paracoccus denitrificans*, having molecular weight $M_r \simeq 14\,500$, space group $P2_12_12_1$, and a single $PtCl_4^{2-}$ derivative, estimates (0 or π) of the three-phase structure invariants are obtained by recently secured direct methods employing the six-magnitude first neighborhood [Hauptman (1982). *Acta Cryst.* A**38**, 289–294] and compared with the known values.

0567-7394/82/030294-07$01.00

The comparison shows that these methods are capable of estimating reliably several tens of thousands of those invariants having the extreme values 0 or π, approximately. It is therefore anticipated that direct methods will play an increasingly important role in the determination of macromolecular structures.

1. Introduction

If f_j and g_j denote atomic structure factors for a corresponding pair of isomorphous structures, the respective normalized structure factors $E_{\mathbf{H}}$ and $G_{\mathbf{H}}$ are defined by

$$E_{\mathbf{H}} = |E_{\mathbf{H}}| \exp(i\varphi_{\mathbf{H}}) = \frac{1}{\alpha_{20}^{1/2}} \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{H}.\mathbf{r}_j) \tag{1.1}$$

$$G_{\mathbf{H}} = |G_{\mathbf{H}}| \exp(i\psi_{\mathbf{H}}) = \frac{1}{\alpha_{02}^{1/2}} \sum_{j=1}^{N} g_j \exp(2\pi i \mathbf{H}.\mathbf{r}_j), \tag{1.2}$$

where

$$\alpha_{mn} = \sum_{j=1}^{N} f_j^m g_j^n, \tag{1.3}$$

some of the f_j's (or g_j's) may be zero (or negative in the neutron diffraction case), and $\mathbf{r}_j$ is the position vector of the atom labeled j.

If

$$\mathbf{H} + \mathbf{K} + \mathbf{L} = 0, \tag{1.4}$$

then there are four kinds of three-phase structure invariant:

$$\omega_0 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}, \tag{1.5}$$

$$\omega_1 = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \psi_{\mathbf{L}}, \tag{1.6}$$

$$\omega_2 = \varphi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}}, \tag{1.7}$$

$$\omega_3 = \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}}. \tag{1.8}$$

The 'pure' invariants ω_0 and ω_3 correspond to the f and g structures, respectively, but the 'mixed' invariants ω_1 and ω_2 have no analog in the non-isomorphous case. Thus, for a fixed number of phases, the number of three-phase structure invariants for an isomorphous pair of structures far exceeds the number available in the non-isomorphous case. The first neighborhood of each of the structure invariants (1.5)–(1.8) is defined to consist of the six magnitudes:

$$|E_{\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{\mathbf{L}}|, \tag{1.9}$$

$$|G_{\mathbf{H}}|, |G_{\mathbf{K}}|, |G_{\mathbf{L}}|. \tag{1.10}$$

The conditional probability distributions,

$$P_\mu = P_\mu(\Omega_\mu | R_1, R_2, R_3, S_1, S_2, S_3), \mu = 0, 1, 2, 3, \tag{1.11}$$

of the structure invariants ω_μ, $\mu = 0, 1, 2, 3$ [equations (1.5)–(1.8)], assuming as known the six magnitudes (1.9) and (1.10) in their first neighborhood:

$$|E_{\mathbf{H}}| = R_1, \quad |E_{\mathbf{K}}| = R_2, \quad |E_{\mathbf{L}}| = R_3; \tag{1.12}$$

$$|G_{\mathbf{H}}| = S_1, \quad |G_{\mathbf{K}}| = S_2, \quad |G_{\mathbf{L}}| = S_3; \tag{1.13}$$

have been found [see the preceding paper (Hauptman, 1982, equations (3.12), (3.16), (3.19), and (3.22)] and are expressible in pure exponential form:

$$P_\mu = \frac{1}{K_\mu} \exp(A_\mu \cos \Omega_\mu), \mu = 0, 1, 2, 3, \tag{1.14}$$

where the parameters A_μ and K_μ are known functions of the α_{mn}'s [equation (1.3)] and $R_1, R_2, R_3, S_1, S_2, S_3$ [equations (1.12) and (1.13)]. [See equations (3.12)–(3.24) of Hauptman (1982).] Clearly P_μ [equation (1.14)] has a unique maximum at $\Omega_\mu = 0$ or π according as $A_\mu > 0$ or $A_\mu < 0$, respectively, and the larger the value of $|A_\mu|$ the smaller the variance of the distribution. Thus one identifies those three-phase structure invariants ω_μ, $\mu = 0, 1, 2, 3$, having the values 0 or π, by calculating the parameters A_μ and selecting those having the largest magnitudes. Then

$$\omega_\mu \simeq 0 \text{ or } \pi \tag{1.15}$$

according as the corresponding A_μ satisfies

$$A_\mu \gg 0 \text{ or } A_\mu \ll 0, \tag{1.16}$$

respectively (the neighborhood principle).

2. The applications

With the presumed known coordinates of the protein cytochrome c_{550} from *Paracoccus denitrificans*, molecular weight $M_r \simeq 14500$, space group $P2_12_12_1$, and a single $PtCl_4^{2-}$ isomorphous derivative, respective normalized structure factors $E_{\mathbf{H}}$ and $G_{\mathbf{H}}$ [equations (1.1) and (1.2)] were calculated to a resolution of 2·5 Å (4159 E's and 4159 G's). The parameters A_μ, $\mu = 0, 1, 2, 3$, which define the conditional probability distributions (1.14) of the three-phase structure invariants (1.5)–(1.8), were calculated in accordance with three different protocols (Table 1):

Protocol 1. With the full set of reflections at 2·5 Å resolution, the 1000 phases φ corresponding to the 1000 largest $|E|$'s of the native protein and the 1000 phases

Table 1. *The three protocols*

Protocol	Number of phases in the basis set: Phases φ	Phases ψ	Resolution
1	1000	1000	2·5 Å
2	2000	2000	2·5 Å
3	1000	1000	4·0 Å

ψ corresponding to the 1000 largest $|G|$'s of the derivative were used to generate the three-phase structure invariants (1.5)–(1.8), and the parameters A_μ were calculated.

Protocol 2. The same as protocol 1 except that the 2000 phases φ and 2000 phases ψ were used to generate the three-phase structure invariants.

Protocol 3. The same as protocol 1 except that the resolution is reduced to 4 Å. (At this resolution only 1076 E's and 1076 G's were available.)

Table 2. *Two thousand three-phase cosine invariants estimated to have the extreme values* ± 1

A representative sample of 50 three-phase structure invariants for cytochrome c_{550} and the $PtCl_4^{2-}$ derivative are chosen from the first 2000 arranged in descending order of $|A_\mu|$ in accordance with protocol 2. According as $A_\mu > 0$ or $A_\mu < 0$ the value of the cosine invariant is estimated to be $+1$ or -1, respectively. In the column headed 'Kind of invariant', the entry $\psi\varphi\varphi$, for example, means $\psi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$, *etc.*, where $\mathbf{H} + \mathbf{K} + \mathbf{L} = 0$.

Serial no.	Kind of invariant	$\|E_{\mathbf{H}}\|$	$\|E_{\mathbf{K}}\|$	$\|E_{\mathbf{L}}\|$	$\|G_{\mathbf{H}}\|$	$\|G_{\mathbf{K}}\|$	$\|G_{\mathbf{L}}\|$	A_μ	Calc. cos	True cos
1	$\psi\varphi\varphi$	1·2	1·0	1·0	1·8	0·2	0·2	8·656	+1	+1·00
2	$\psi\psi\psi$	1·2	0·4	0·1	1·8	1·1	0·9	8·648	+1	+1·00
3	$\varphi\psi\psi$	1·0	0·8	0·8	0·2	1·5	1·5	−8·565	−1	−1·00
4	$\varphi\psi\psi$	1·0	0·8	0·4	0·2	1·5	1·1	−8·527	−1	−1·00
5	$\psi\psi\psi$	1·2	0·8	0·5	1·8	1·5	1·2	8·518	+1	+1·00
6	$\varphi\varphi\psi$	1·7	1·0	0·5	0·9	0·2	1·2	8·467	+1	+1·00
7	$\psi\psi\psi$	1·2	0·5	0·4	1·8	1·2	1·1	8·461	+1	+1·00
8	$\psi\psi\varphi$	1·2	1·2	1·0	1·8	1·8	0·2	−8·401	−1	−1·00
9	$\varphi\psi\psi$	1·0	0·8	0·5	0·2	1·5	1·1	−8·222	−1	−1·00
10	$\varphi\psi\psi$	1·7	1·2	0·8	0·9	1·8	1·5	−8·132	−1	−1·00
491	$\varphi\varphi\psi$	1·4	1·2	0·5	0·7	0·4	1·1	6·592	+1	+1·00
492	$\varphi\varphi\psi$	1·8	0·9	0·4	2·3	0·2	1·1	−6·591	−1	−0·81
493	$\psi\psi\psi$	0·9	0·8	0·5	1·4	1·4	1·1	6·591	+1	+0·98
494	$\psi\varphi\psi$	1·9	0·9	0·3	2·4	0·2	0·8	−6·586	−1	−0·97
495	$\varphi\varphi\psi$	1·9	0·9	0·3	2·4	0·2	0·8	−6·586	−1	−0·97
496	$\varphi\varphi\psi$	2·2	1·9	0·5	2·7	2·5	1·2	6·585	+1	+0·93
497	$\varphi\psi\psi$	2·2	1·9	0·5	2·7	2·5	1·2	6·585	+1	+0·95
498	$\psi\varphi\psi$	2·2	1·9	0·5	2·7	2·5	1·2	6·585	+1	+0·95
499	$\psi\psi\psi$	2·2	1·9	0·5	2·7	2·5	1·2	6·585	+1	+0·96
500	$\varphi\varphi\varphi$	2·7	1·9	1·1	1·9	2·5	0·4	6·584	+1	+1·00
991	$\psi\varphi\varphi$	4·4	2·1	1·2	3·4	1·3	1·8	6·172	+1	+1·00
992	$\varphi\psi\varphi$	1·8	1·5	1·2	1·1	2·0	1·8	−6·172	−1	−0·98
993	$\varphi\psi\psi$	1·4	0·4	0·5	0·6	1·0	1·0	−6·171	−1	−0·99
994	$\psi\psi\psi$	0·8	0·7	0·4	1·4	1·3	1·0	6·171	+1	+0·99
995	$\varphi\psi\psi$	4·4	2·1	1·2	3·4	1·3	1·8	6·170	+1	+1·00
996	$\psi\psi\psi$	4·4	2·1	1·2	3·4	1·3	1·8	6·170	+1	+1·00
997	$\psi\varphi\psi$	2·1	1·7	0·8	2·6	0·9	1·4	−6·169	−1	−1·00
998	$\varphi\varphi\psi$	2·1	1·7	0·8	2·6	0·9	1·4	−6·169	−1	−1·00
999	$\varphi\psi\psi$	1·2	1·1	0·5	0·4	1·6	1·1	−6·168	−1	−1·00
1000	$\psi\varphi\psi$	2·7	1·9	1·9	3·1	2·5	2·5	6·167	+1	+1·00
1491	$\varphi\varphi\varphi$	4·4	2·2	1·9	3·4	2·7	2·4	−5·882	−1	−0·98
1492	$\varphi\psi\psi$	1·5	1·3	0·4	2·0	1·8	1·1	5·881	+1	+0·96
1493	$\psi\varphi\psi$	1·2	1·2	0·4	1·7	0·5	1·1	−5·881	−1	−0·83
1494	$\varphi\psi\psi$	1·7	0·5	0·5	0·9	1·1	1·0	−5·881	−1	−0·87
1495	$\psi\varphi\psi$	1·9	1·6	1·3	2·4	0·8	1·8	−5·880	−1	−1·00
1496	$\varphi\varphi\psi$	1·9	1·6	1·3	2·4	0·8	1·8	−5·880	−1	−1·00
1497	$\varphi\psi\psi$	1·8	1·1	0·4	2·3	1·6	1·0	5·879	+1	+0·57
1498	$\psi\psi\psi$	1·2	0·6	0·4	1·8	1·2	1·0	5·878	+1	+1·00
1499	$\varphi\varphi\psi$	2·4	1·0	0·5	2·9	0·2	1·1	−5·878	−1	−0·99
1500	$\psi\varphi\psi$	2·4	1·0	0·5	2·9	0·2	1·1	−5·878	−1	−0·99
1991	$\psi\varphi\psi$	1·6	1·4	0·8	2·1	0·7	1·4	−5·681	−1	−0·99
1992	$\varphi\psi\psi$	0·9	0·8	0·5	1·5	1·4	1·1	5·681	+1	+0·98
1993	$\varphi\psi\psi$	1·5	1·2	0·4	2·0	1·7	1·1	5·680	+1	+0·98
1994	$\varphi\psi\psi$	1·8	1·0	0·4	1·1	1·6	1·0	−5·680	−1	−1·00
1995	$\varphi\varphi\varphi$	1·6	1·4	0·9	2·1	0·6	0·2	5·679	+1	+1·00
1996	$\varphi\psi\psi$	2·4	1·2	0·4	2·9	1·8	1·0	5·679	+1	+0·89
1997	$\psi\psi\psi$	2·4	1·2	0·4	2·9	1·8	1·0	5·679	+1	+0·92
1998	$\varphi\psi\psi$	1·8	0·3	0·2	2·3	0·9	0·8	5·677	+1	+0·91
1999	$\psi\psi\psi$	2·1	1·7	0·8	2·6	0·9	1·4	−5·676	−1	−1·00
2000	$\varphi\psi\psi$	2·1	1·7	0·8	2·6	0·9	1·4	−5·676	−1	−1·00

Employing protocol 2, the 2000 structure invariants having the largest magnitudes $|A_\mu|$ were used in the construction of Table 2, which identifies most reliably those structure invariants having the extreme values 0 or π or, equivalently, those cosine invariants having the extreme values ± 1. Column 2 of Table 2 identifies the kind of invariant, ω_0, ω_1, ω_2 or ω_3 [equations (1.5)–(1.8)], columns 3–8 give the values of the six magnitudes [equations (1.12) and (1.13)] in the first neighborhood of the invariant, and column 9 the values of A_μ. According as $A_\mu \gg 0$ or $A_\mu \ll 0$ the cosine invariant is assigned the value $+1$ or -1, respectively (Calc. cos, column 10). For the purpose of comparison the true value of the cosine invariant is shown in the final column.

Employing all three protocols, the summary Table 3, based on the 32000 structure invariants having the largest values of $|A_\mu|$, was constructed. Table 4 was constructed from Table 3 by accumulation into the five groups shown. The entries in the columns headed 'Average mag. of error' are defined by

$$\langle |\text{calc. cos} - \text{true cos}| \rangle. \qquad (2.1)$$

As reference to columns 4–9 of Table 3 and columns 3–8 of Table 4 shows, it is better, for fixed resolution, to have 4000 phases (2000 φ's plus 2000 ψ's) in the basis set (protocol 2) than 2000 phases (1000 φ's plus 1000 ψ's, protocol 1) because the larger basis set generates a far greater number of invariants with large values of $|A_\mu|$. Comparison of columns 4–6 with columns 10–12 of Table 3 and of columns 3–5 with columns 9–11 of Table 4 shows that, for a fixed number of phases in the basis set, more reliable estimates are obtained at the lower resolution, for reasons unknown at this time. In all cases the ability of our formulas to estimate reliably large numbers of invariants, a consequence of the unexpectedly large number of invariants with large values of $|A_\mu|$, is noteworthy. In fact, Tables 3 and 4 show that, with protocol 2, no error in sign occurs among the first 12000 invariants, only 43 errors in sign occur among the top 32000 invariants, and the average magnitude of the error in estimated cosines for these 32000 invariants is only 0·108. Tables 3 and 4 clearly show, as expected, that the larger the value of $|A_\mu|$ the more reliable is the estimate of the cosine. It should be

Table 3. *Average magnitude of the error in estimated values* (± 1) *of* 32000 *three-phase cosine invariants for cytochrome* c_{550} *and the* $PtCl_4^{2-}$ *derivative arranged in groups in descending order of* $|A_\mu|$

The columns headed 'Number with wrong sign' show the number of cosines estimated to be $+1$ (-1) whose true values are in fact negative (positive).

			Protocol 1			Protocol 2			Protocol 3		
Group no.	Serial nos.	Number in group	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign
1	1–200	200	6·796	0·018	0	7·527	0·017	0	7·177	0·021	0
2	201–500	300	5·931	0·043	0	6·795	0·024	0	6·410	0·024	0
3	501–1000	500	5·434	0·059	0	6·348	0·026	0	5·881	0·034	0
4	1001–1500	500	5·062	0·065	0	6·018	0·034	0	5·480	0·032	0
5	1501–2000	500	4·774	0·074	0	5·770	0·040	0	5·197	0·043	0
6	2001–2500	500	4·550	0·099	0	5·593	0·042	0	4·988	0·052	0
7	2501–3000	500	4·366	0·092	0	5·439	0·042	0	4·822	0·057	0
8	3001–3500	500	4·209	0·147	8	5·316	0·059	0	4·673	0·061	0
9	3501–4000	500	4·074	0·142	8	5·209	0·046	0	4·551	0·067	0
10	4001–4500	500	3·954	0·126	0	5·111	0·060	0	4·447	0·065	0
11	4501–5000	500	3·847	0·132	1	5·027	0·060	0	4·346	0·067	0
12	5001–6000	1000	3·703	0·131	1	4·911	0·064	0	4·211	0·106	8
13	6001–7000	1000	3·527	0·163	8	4·774	0·075	0	4·056	0·115	9
14	7001–8000	1000	3·381	0·154	2	4·660	0·085	0	3·925	0·095	0
15	8001–9000	1000	3·246	0·179	6	4·556	0·082	0	3·802	0·112	1
16	9001–10000	1000	3·136	0·185	8	4·461	0·097	0	3·693	0·124	3
17	10001–12000	2000	2·985	0·191	11	4·336	0·097	0	3·550	0·136	11
18	12001–14000	2000	2·814	0·222	38	4·189	0·118	8	3·382	0·143	11
19	14001–16000	2000	2·661	0·209	37	4·064	0·118	8	3·235	0·153	10
20	16001–18000	2000	2·530	0·224	50	3·951	0·117	0	3·105	0·170	19
21	18001–20000	2000	2·417	0·240	54	3·858	0·126	2	2·992	0·184	14
22	20001–22000	2000	2·316	0·236	42	3·764	0·124	1	2·891	0·183	13
23	22001–24000	2000	2·226	0·251	51	3·677	0·128	4	2·800	0·182	19
24	24001–26000	2000	2·142	0·245	51	3·598	0·139	2	2·714	0·196	24
25	26001–28000	2000	2·067	0·275	88	3·525	0·138	0	2·637	0·213	43
26	28001–32000	4000	1·962	0·289	177	3·429	0·155	18	2·534	0·221	94

Table 4. *Average magnitude of the error in estimated values* (± 1) *of* 32 000 *three-phase cosine invariants for cytochrome* c_{550} *and the* $PtCl_4^{2-}$ *derivative cumulated in groups from Table* 3 *and arranged in descending order of* $|A_\mu|$

The columns headed 'Number with wrong sign' show the number of cosines estimated to be +1 (−1) whose true values are in fact negative (positive).

		Protocol 1			Protocol 2			Protocol 3		
Group no.	Number in group	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign
1	1000	5·855	0·046	0	6·718	0·024	0	6·299	0·028	0
2	5000	4·655	0·097	17	5·692	0·043	0	5·110	0·050	0
3	10000	4·027	0·130	42	5·182	0·062	0	4·524	0·080	21
4	20000	3·354	0·173	232	4·630	0·088	18	3·888	0·119	86
5	32000	2·888	0·207	641	4·233	0·108	43	3·437	0·150	279

emphasized finally that the most reliably estimated invariants are not necessarily those corresponding to the most intense reflections (in sharp contrast to the non-isomorphous case) but correspond instead to reflections of only moderate intensity, as inspection of Table 2, in particular columns 2–8, shows.

3. Comparison with the non-isomorphous case

In the case of no isomorphism the only available three-phase structure invariants are the 'pure' invariants ω_0 or ω_3, equations (1.5) and (1.8), where (1.4) holds. In this case the first neighborhood of ω_0, for example, consists of only the three magnitudes (1.9) and the conditional probability distribution of ω_0, given the three magnitudes (1.9), is well known to be

$$P = \frac{1}{K} \exp(A \cos \Omega), \tag{3.1}$$

where

$$A = \frac{2\alpha_{30}}{\alpha_{20}^{3/2}} |E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{L}}|, \tag{3.2}$$

$$K = \frac{1}{2\pi} I_0(A), \tag{3.3}$$

I_0 is the modified Bessel function and α_{mn} is defined by (1.3). Clearly (3.1) has the same form as (1.14) but since A, equation (3.2), is always positive, only the estimate +1 for the three-phase cosine invariant, cos ω_0, is now possible, and this estimate is reliable only when A is large, *i.e.* when the variance of the distribution (3.1) is small.

With the 2000 phases φ for the native protein corresponding to the largest $|E|$ values, pure three-phase cosine invariants cos ω_0 were generated and their values (always +1) were estimated *via* the traditional non-isomorphous A values, equation (3.2). These were arranged in descending order of A, and estimates of a representative sample of 20 chosen from the top 100 are shown in Table 5. Summary Tables 6 and 7 list the average magnitude of the errors and numbers of wrong signs in groups arranged in descending order of A for the first 10000. Calculations for the non-isomorphous case for the native protein and for the derivative separately have also been carried out using the analogs of protocols 1, 2, and 3. The results are so similar to those summarized in Tables 5 and 6 that there is no need to describe them in further detail here. As anticipated there are in all these cases far fewer invariants having large values of A than there were in the isomorphous case (compare Tables 5, 6 and 7 with

Table 5. *One hundred three-phase cosine invariants* cos $(\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}})$ *estimated to have the extreme value* +1 *(the only possible estimate) via the traditional (non-isomorphous)* A *value*

A representative sample of 20 three-phase structure invariants for cytochrome c_{550} is chosen from the first 100 arranged in descending order of A. The 2000 phases φ for the native protein corresponding to the 2000 largest $|E|$ values were used to generate the three-phase structure invariants. The resolution is 2·5 Å.

Serial no.	$\|E_{\mathbf{H}}\|$	$\|E_{\mathbf{K}}\|$	$\|E_{\mathbf{L}}\|$	$\|G_{\mathbf{H}}\|$	$\|G_{\mathbf{K}}\|$	$\|G_{\mathbf{L}}\|$	A	Calc. cos	True cos
1	8·6	8·6	3·3	8·0	8·0	3·6	8·282	+1	−1·00
2	8·6	4·1	3·8	8·0	3·5	3·5	4·506	+1	+1·00
3	8·6	4·0	3·6	8·0	3·8	3·2	4·250	+1	−1·00
4	8·6	4·4	3·1	8·0	3·4	3·0	4·024	+1	−1·00
5	8·6	3·7	3·3	8·0	3·6	3·6	3·557	+1	−1·00
6	8·6	4·0	2·8	8·0	3·8	2·3	3·296	+1	+1·00
7	8·6	5·3	1·9	8·0	4·9	2·4	2·911	+1	0·00
8	8·6	3·7	2·5	8·0	3·4	1·9	2·675	+1	+0·85
9	5·3	3·8	3·7	4·9	3·5	3·4	2·512	+1	−0·85
10	8·6	2·9	2·9	8·0	2·6	2·0	2·444	+1	−0·59
91	8·6	3·5	1·2	8·0	3·4	1·0	1·175	+1	+0·61
92	5·3	2·6	2·5	4·9	2·7	2·1	1·172	+1	+0·99
93	4·4	3·3	2·4	3·4	3·6	2·8	1·172	+1	−1·00
94	8·6	2·2	1·8	8·0	2·1	1·3	1·152	+1	+0·98
95	8·6	2·5	1·6	8·0	1·8	1·4	1·150	+1	+1·00
96	4·1	3·4	2·4	3·5	3·1	2·9	1·150	+1	−1·00
97	8·6	2·5	1·5	8·0	2·3	1·9	1·149	+1	+0·44
98	4·4	3·5	2·2	3·4	3·4	1·9	1·147	+1	+0·92
99	4·4	3·4	2·3	3·4	3·3	2·1	1·145	+1	−0·08
100	8·6	2·4	1·6	8·0	2·4	1·8	1·144	+1	−0·17

Tables 2, 3 and 4, respectively, in particular protocol 2 of Tables 3 and 4) resulting in the relatively poor estimates shown in Tables 5–7. In fact, comparison of Tables 6 and 7 with Tables 3 and 4, protocol 2, respectively, shows that the ability to integrate the techniques of direct methods and isomorphous replacement leads to an enormous increase in the number of invariants whose values may be estimated reliably.

Table 6. *Average magnitude of the error in estimated values (always +1) of 10 000 three-phase cosine invariants* $\cos(\varphi_H + \varphi_K + \varphi_L)$ *for cytochrome* c_{550}, *via the traditional (non-isomorphous) A value, arranged in descending order of A*

The 2000 phases φ for the native protein corresponding to the 2000 largest $|E|$ values were used to generate the three-phase structure invariants. The resolution is 2·5 Å. The column headed 'Number with wrong sign' shows the number of cosines estimated to be +1 whose true values are in fact negative.

Group no.	Serial nos.	Number in group	Average value of A	Average mag. of error	Number with wrong sign
1	1–200	200	1·364	0·802	74
2	201–500	300	0·802	0·760	112
3	501–1000	500	0·631	0·808	191
4	1001–1500	500	0·539	0·807	188
5	1501–2000	500	0·487	0·835	193
6	2001–2500	500	0·451	0·849	203
7	2501–3000	500	0·425	0·899	215
8	3001–3500	500	0·406	0·884	218
9	3501–4000	500	0·389	0·830	195
10	4001–4500	500	0·374	0·888	216
11	4501–5000	500	0·361	0·882	216
12	5001–6000	1000	0·346	0·878	428
13	6001–7000	1000	0·329	0·854	417
14	7001–8000	1000	0·315	0·833	393
15	8001–9000	1000	0·303	0·876	424
16	9001–10000	1000	0·292	0·902	434

Table 7. *Average magnitude of the error in estimated values (always +1) of 10 000 three-phase cosine invariants* $\cos(\varphi_H + \varphi_K + \varphi_L)$ *for cytochrome* c_{550} *cumulated in groups from Table 6 and arranged in descending order of A*

The column headed 'Number with wrong sign' shows the number of cosines estimated to be +1 whose true values are in fact negative.

Group no.	Number in group	Average value of A	Average mag. of error	Number with wrong sign
1	1000	0·829	0·792	377
2	5000	0·509	0·846	2021
3	10000	0·413	0·857	4117

4. Effect of the sign of A_μ on the reliability of the estimate

Since the variance of the distribution (1.14) depends only on the magnitude of A_μ, it is natural to anticipate that the reliability of the estimate of the cosine invariant depends only on $|A_\mu|$ and is independent of the sign of A_μ. However, extensive calculations show that there is a small dependence on the sign of A_μ; for fixed $|A_\mu|$ estimates of -1 are slightly more reliable than estimates of $+1$. Tables 8 and 9, summarizing the calculations for protocol 2 when $A_\mu > 0$ and when $A_\mu < 0$, respectively, show a small but significant bias; when $A_\mu < 0$ the estimates tend to be somewhat more reliable than when $A_\mu > 0$, for fixed $|A_\mu|$. These tables also show, however, that, for each fixed positive number p, there are more invariants with positive values of A_μ than with negative values of A_μ satisfying $|A_\mu| > p$; *i.e.* there are more invariants having large positive values of A_μ than invariants having large negative (in magnitude) values of A_μ. The two effects roughly compensate each other so that for fixed average magnitude of error, ε, the number of invariants with positive A_μ and average error ε is roughly equal to the number of invariants with negative A_μ and average error ε. (Compare columns 5 and 8 of Table 8, and columns 4 and 9 of Table 9.) Calculations for protocols 1 and 3 (not detailed here) confirm this effect.

5. Concluding remarks

Recently secured formulas for estimating the values (0 or π) of the three-phase structure invariants, employing a combination of direct methods and isomorphous replacement, have been applied to the protein cytochrome c_{550} and a single heavy-atom derivative, using error-free diffraction data. The number of invariants reliably estimated to be 0 or π in this way runs into tens of thousands, far greater than is possible with the traditional techniques of direct methods. The same calculations have been carried out on the calcium-binding protein parvalbumin, molecular weight about 11500, space group $C2$, and a single heavy-atom derivative, with results which confirm in every particular those described in detail in this paper. It is therefore anticipated that these methods will play an increasingly important role in macromolecular structure determination.

The major purpose of the present paper has been to show that the formulas derived in the preceding paper are valid when error-free diffraction data are assumed to be available. A major unanswered question is concerned with the effects of error in real data and of imperfect isomorphism, and a study of this problem will be presented at a later date. However, it is not unreasonable to expect that the adverse effect resulting from these errors will be reduced during the process of

Table 8. *Average magnitude of the error in estimated values* (+1) [(−1)] *of the* 18 814 [13 186] *three-phase cosine invariants, among the* 32 000 *of Table* 3, *protocol* 2, *for which* $A_\mu > 0$ [$A_\mu < 0$]

When $A_\mu < 0$ the number of invariants in the last group is 1186, not 2000.

			$A_\mu > 0$			$A_\mu < 0$		
Group no.	Serial nos.	Number in group	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign
1	1–200	200	7·237	0·024	0	7·156	0·107	0
2	201–500	300	6·474	0·026	0	6·298	0·027	0
3	501–1000	500	6·007	0·034	0	5·751	0·037	0
4	1001–1500	500	5·639	0·045	0	5·365	0·043	0
5	1501–2000	500	5·385	0·051	0	5·094	0·056	0
6	2001–2500	500	5·206	0·054	0	4·891	0·056	0
7	2501–3000	500	5·053	0·061	0	4·724	0·075	0
8	3001–3500	500	4·922	0·070	0	4·587	0·073	0
9	3501–4000	500	4·808	0·079	0	4·468	0·084	0
10	4001–4500	500	4·709	0·084	0	4·369	0·084	0
11	4501–5000	500	4·621	0·088	0	4·278	0·096	0
12	5001–6000	1000	4·495	0·097	0	4·149	0·100	0
13	6001–7000	1000	4·346	0·102	0	4·002	0·109	0
14	7001–8000	1000	4·217	0·132	8	3·881	0·113	1
15	8001–9000	1000	4·108	0·127	8	3·771	0·111	1
16	9001–10000	1000	4·009	0·120	0	3·669	0·125	2
17	10001–12000	2000	3·873	0·131	1	3·536	0·138	0
18	12001–14000	2000	3·720	0·130	2	3·411	0·143	4
19	14001–16000	2000	3·581	0·139	2			
20	16001–18814	2814	3·443	0·161	14			

Table 9. *Average magnitude of the error in estimated values* (+1) [(−1)] *of the* 18 814 [13 186] *three-phase cosine invariants cumulated in groups from Table* 8

	$A_\mu > 0$				$A_\mu < 0$				
Group no.	Number in group	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign	Group no.	Number in group	Average value of $\|A_\mu\|$	Average mag. of error	Number with wrong sign
1	1000	6·393	0·029	0	1	1000	6·196	0·030	0
2	4000	5·475	0·052	0	2	4000	5·190	0·056	0
3	7000	5·058	0·071	0	3	7000	4·748	0·075	0
4	10000	4·774	0·087	16	4	10000	4·456	0·087	4
5	14000	4·495	0·100	19	5	13186	4·222	0·100	8
6	18814	4·240	0·113	35					

phase determination because of the large number of available structure invariants relative to the number of unknown phases. In applying these formulas to the solution of real structures the obvious approach would be to employ a modified tangent procedure in which the estimated values of the structure invariants (0 or π) would be used; in short to employ the existing machinery of direct methods, suitably modified. Finally, it should be pointed out that, in view of the available evidence, the present work may well make possible unique macromolecular structure determination *via* a combination of direct methods and single isomorphous replacement, thus making unnecessary the need for the several isomorphous derivatives which are presently required by the conventional isomorphous replacement technique.

This work was supported by DHHS Grant Nos. GM26195 and AM25192 and by NSF Grant No. CHE-7911282, as well as a grant from the James H. Cummings Foundation, Inc.

Reference

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