We thank the National Research Council of Canada for financial support, the University of British Columbia Computing Centre for assistance, and Professor A. Rosenthal and Dr A. Brink for the crystals.

References

BOEYENS, J. C. A. & KRUGER, G. J. (1970). *Acta Cryst.* B**26**, 668–672.
CROMER, D. T. & LIBERMAN, D. (1970). *J. Chem. Phys.* **53**, 1891–1898.
CROMER, D. T. & MANN, J. B. (1968). *Acta Cryst.* A**24**, 321–324.
DREW, M. G. B., TEMPLETON, D. H. & ZALKIN, A. (1969). *Acta Cryst.* B**25**, 261–267.
DUPONT, L., DIDEBERG, O. & WELTER, A. (1975). *Acta Cryst.* B**31**, 1018–1022.
FANFANI, L., NUNZI, A., ZANAZZI, P. F. & ZANZARI, A. R. (1974). *Acta Cryst.* B**30**, 127–132.
HAMILTON, W. C. (1965). *Acta Cryst.* **18**, 502–510.
KARLE, J. & HAUPTMAN, H. (1956). *Acta Cryst.* **9**, 635–651.
KARLE, J. & KARLE, I. L. (1966). *Acta Cryst.* **21**, 849–859.
KIM, S. H. & JEFFREY, G. A. (1967). *Acta Cryst.* **22**, 537–545.
ROSENTHAL, A. & BRINK, A. J. (1975). *J. Carbohydr. Nucleosides Nucleotides*, **2**, 343–356.
STEWART, R. F., DAVIDSON, E. R. & SIMPSON, W. T. (1965). *J. Chem. Phys.* **42**, 3175–3187.

Acta Cryst. (1978). B**34**, 1599–1608

The Application of *MULTAN* to the Analysis of Isomorphous Derivatives in Protein Crystallography

BY KEITH S. WILSON

Laboratory of Molecular Biophysics, Department of Zoology, South Parks Road, Oxford, England

(*Received* 21 *October* 1977; *accepted* 30 *November* 1977)

The application of *MULTAN* to the analysis of isomorphous derivatives of four proteins (phosphorylase b, phosphoglycerate kinase, hagfish insulin and triose phosphate isomerase) is described. The method leads to the correct structure for almost all the derivatives studied. The best phase set can be selected on the basis of the 'ABSFOM' figure of merit which represents the internal consistency of the phase set. The method is limited by the experimental accuracy of the isomorphous difference.

1. Introduction

The preparation of isomorphous derivatives can be the most time-consuming step in the determination of the crystal structure of a protein. After diffusion of a heavy-atom reagent there should be measurable changes in the diffracted intensities. These changes must be analysed to give an atomic model and the parameters of the model refined. For the first derivative the analysis is conventionally carried out by inspection of the difference Patterson synthesis.

For proteins of high molecular weight multi-site binding may be required to produce intensity changes which are statistically significant. Multi-site binding will necessarily occur when non-crystallographic symmetry elements are present. Unfortunately the difference Patterson map becomes increasingly complex as the number of equivalent positions in the space group and the number of binding sites per asymmetric unit rise.

Table 1. *Abbreviations used in the text*

EMP	Ethyl mercury phosphate
AUCN	Gold cyanide
PCMB	Mercury *p*-chlorobenzoate
BAKERS	The Bakers' dimercurial reagent
ABSFOM, RESID, PSIZERO	Figures of merit for the phase sets generated by *MULTAN* defined in § 2
F_H	The structure factor moduli of the heavy-atom structure
F_P	The structure factor moduli of the native protein
F_{PH}	The structure factor moduli of the isomorphous derivative
ΔF	The isomorphous difference $\|\|F_{PH}\| - \|F_P\|\|$
E_h	The normalized structure factor
R_c	Conventional R factor for least-squares refinement of the centric terms: $R_c = \sum(F_o - F_c)/\sum F_o$, where F_o is the observed and F_c the calculated structure factor amplitude. F_o is the isomorphous difference, ΔF, in this paper
B	Atomic temperature factor

The direct methods of phase determination offer an alternative to the Patterson synthesis. They were first successfully applied to the problem by Steitz (1968) and have been used by, among others, Schevitz *et al.* (1972) and Navia & Sigler (1974). The method is intended primarily for the first derivative when the difference Fourier synthesis cannot be used. For subsequent derivatives it may be useful when there are sites common to more than one isomorphous compound.

The experiments described below were undertaken in the hope that a simple procedure for a direct-methods analysis would be found. This study differs from those of other workers in that the method requires a minimum of intervention by the user and has been tested on an appreciable number of isomorphous derivatives.

Abbreviations used in the paper are detailed in Table 1.

2. Method

The derivative structure factor moduli, F_{PH}, were scaled to those of the native protein, F_P. A value of unity was assumed to be a satisfactory approximation to the scale factor, K:

$$K = \frac{\sum F_{PH}^2}{\sum F_P^2}. \tag{2.1}$$

The isomorphous difference, ΔF, was taken as the experimental estimate of the heavy-atom structure factor modulus F_H:

$$\Delta F = ||F_{PH}| - |F_P||. \tag{2.2}$$

The ΔF were used in the computation of normalized structure factors, $E_\mathbf{h}$:

$$E_\mathbf{h} = \frac{\Delta F_\mathbf{h}}{\left(\varepsilon \sum_{j=1}^{N} Z_j^2\right)^{1/2}}, \tag{2.3}$$

where N is the number of atoms in the unit cell, Z_j is the atomic sacttering factor of atom j, and ε takes into account the effect of space-group symmetry on diffracted intensity. Overall scale and temperature factors were calculated in the usual way.

ΔF is an imperfect estimate of F_H, firstly because it is the difference in modulus whereas F_H is the vector difference and secondly owing to the observational errors in F_P and F_{PH}. The former error leads to a systematic underestimation of F_H (except when F_P and F_{PH} are collinear) and to a statistical distribution for the $E_\mathbf{h}$ which approximates to that for a centrosymmetric structure. The latter error is more serious. Large random fractional errors in ΔF lead to spuriously large F_H and hence $E_\mathbf{h}$. If these are involved in the earliest stages of phase development the errors introduced may be propagated throughout the phase set.

The program *MULTAN* was used for the phase determination. *MULTAN* is a multisolution program which is based upon a weighted tangent formula (Germain & Woolfson, 1968; Germain, Main & Woolfson, 1970, 1971). The phase indication is given by

$$\tan \varphi_\mathbf{h} = \frac{\sum_\mathbf{k} W_\mathbf{k} W_{\mathbf{h}-\mathbf{k}} |E_\mathbf{k} E_{\mathbf{h}-\mathbf{k}}| \sin(\varphi_\mathbf{k} + \varphi_{\mathbf{h}-\mathbf{k}})}{\sum_\mathbf{k} W_\mathbf{k} W_{\mathbf{h}-\mathbf{k}} |E_\mathbf{k} E_{\mathbf{h}-\mathbf{k}}| \cos(\varphi_\mathbf{k} + \varphi_{\mathbf{h}-\mathbf{k}})} = \frac{T_\mathbf{h}}{B_\mathbf{h}}, \tag{2.4}$$

where $W_\mathbf{h}$ is the weight associated with the phase $\varphi_\mathbf{h}$:

$$W_\mathbf{h} = \alpha_\mathbf{h}/5 \quad \text{for } \alpha_\mathbf{h} \le 5 \tag{2.5a}$$

$$= 1 \quad \text{for } \alpha_\mathbf{h} \ge 5 \tag{2.5b}$$

$$\alpha_\mathbf{h} = 2\sigma_3 \sigma_2^{-3/2} |E_\mathbf{h}| (T_\mathbf{h}^2 + B_\mathbf{h}^2)^{1/2} \tag{2.6}$$

$$\sigma_n = \sum_{j=1}^{N} Z_j^n. \tag{2.7}$$

A specified number of the largest $E_\mathbf{h}$ was used as input to *MULTAN* and the triple-product relations, up to a maximum of 2000, were used in the phase determination. The maximum number is merely a convenient limitation for the Oxford computing system and is likely to be different on other machines. Two atoms per asymmetric unit were assumed in the computation of $\alpha_\mathbf{h}$. The starting set for phase determination comprised reflections to fix the origin and enantiomorph, those $\sum_1$ indications with associated probability greater than 0·9 (Hauptman & Karle, 1953) and two or three known phases which were assigned multiple values.

A particular problem in the application of direct methods to the isomorphous derivatives of proteins is the evaluation of the results. No prior knowledge of the structure is available, unlike the situation in small-molecule studies. It is therefore important that a reliable internal criterion for 'correctness' of a phase set be available. Three figures of merit (ABSFOM, PSIZERO and RESID) were computed for each phase set, and their usefulness in indicating the best phase set was investigated.

ABSFOM represents the internal consistency of the phase set:

$$\text{ABSFOM} = \frac{\sum_\mathbf{h} \alpha_\mathbf{h} - \sum_\mathbf{h} \alpha_r}{\sum_\mathbf{h} \alpha_e - \sum_\mathbf{h} \alpha_r}, \tag{2.8}$$

where $\sum_\mathbf{h} \alpha_e$ is the estimated self consistency, $\sum_\mathbf{h} \alpha_\mathbf{h}$ the experimental value and $\sum_\mathbf{h} \alpha_r$ is computed assuming random phases. For structures of small molecules the value of ABSFOM for the correct phase set is generally in the range 1·0–1·2.

PSIZERO is defined by:

$$\text{PSIZERO} = \sum_{\mathbf{h}} |\sum_{\mathbf{k}} E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}|, \qquad (2.9)$$

where the inner summation is over the $E_{\mathbf{h}}$ in the phase set, and the outer is over the fifty $E_{\mathbf{h}}$ with values nearest to 0·0. PSIZERO should be a minimum for the correct set.

RESID is a conventional crystallographic residual for the equations

$$E_{\mathbf{h}} = K\langle E_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}\rangle_{\mathbf{k}}, \qquad (2.10)$$

where K is an empirically determined scale factor:

$$K = \frac{\sum_{\mathbf{h}} |E_{\mathbf{h}}|}{\sum |\langle E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}\rangle_{\mathbf{k}}|}. \qquad (2.11)$$

RESID should be a minimum for the best phase set, and lies in the range 20–25% for the correct set in most structural analyses of small molecules.

Sets of phases, $\varphi_{\mathbf{h}}$, derived by *MULTAN* were used for the calculation of weighted E maps with coefficients:

$$W_{\mathbf{h}} |E_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}), \qquad (2.12)$$

where $W_{\mathbf{h}}$ is defined in equation 2.5. In the following results the 'background' level of the E maps is that level of density which corresponds to the larger (but not significant) positive and negative features common to most sections.

3. Results

The crystal forms of the four proteins studied are reported in Table 2. The structures are representative of several problems which may be encountered in an isomorphous replacement study. Phosphorylase b crystallizes in a unit cell containing 800 000 daltons of protein: the observational errors in the measurements are large. There is one highly isomorphous derivative of phosphoglycerate kinase, together with two derivatives of much lower quality. Hagfish insulin is a small protein with a molecular weight of 6300, but the crystals are small and the data have large errors of observation. For triose phosphate isomerase at least one heavy atom lies close to a 'special position' (defined below) in all of the four derivatives. For each protein the data used corresponded to spacings greater than 6 Å.

3.1. *Phosphorylase b*

Two derivatives were studied: diaminodichloroplatinum with one major binding site and ethylmercury thiosalicylate with two major sites (Johnson, Madsen, Mosley & Wilson, 1974). The mean isomorphous difference was three times the mean standard deviation estimated from counting statistics – almost certainly an underestimate of the true error. Location of the heavy atoms was thought to pose a stringent test for the application of the tangent formula to very large proteins.

For comparison with the present study I note that the interpretation of the Patterson synthesis was easy for the Pt, and possible for the Hg – but complex with twenty independent vectors present.

The application of the tangent formula to the Pt derivative was immediately successful. Phases were generated for the 300 largest $E_{\mathbf{h}}$ corresponding to a spacing greater than 6 Å. The figures of merit suggest the phase sets are of considerably lower quality than those expected in studies of small molecules (Table 3).

An E map was calculated for each phase set on the same arbitrary scale. Phase sets (5), (7), (10) and (12) had the highest – and closely similar – values of ABSFOM and the E maps computed with these sets

Table 2. *A summary of the crystal forms for the proteins studied (V is the volume of the unit cell)*

Protein	Space group	Cell dimensions	Number of reflections	V	Protein content of asymmetric unit (daltons)	Protein content of unit cell (daltons)
Phosphorylase b	$P4_32_12$	$a = b = 128{\cdot}5$ Å $c = 115{\cdot}9$	2500	$1{\cdot}93 \times 10^6$ Å^3	100 000	800 000
Phosphoglycerate kinase	$P2_1$	$a = 50{\cdot}8$ Å $b = 106{\cdot}9$ $c = 36{\cdot}3$ $\beta = 98{\cdot}6°$	1100	$1{\cdot}95 \times 10^5$	40 000	80 000
Hagfish insulin	$P4_12_12$	$a = b = 38{\cdot}4$ Å $c = 85{\cdot}3$	215	$1{\cdot}25 \times 10^5$	6 000	48 000
Triose phosphate isomerase	$P2_12_12_1$	$a = 106{\cdot}0$ Å $b = 74{\cdot}8$ $c = 61{\cdot}7$	1300	$4{\cdot}89 \times 10^5$	50 000	200 000

Table 3. *The figures of merit of the phase sets generated by MULTAN for the platinum derivative of phosphorylase b*

The 300 largest E_h to a spacing of 6 Å were included. Sets five, seven, ten and twelve correspond to the correct heavy-atom structure.

Phase set	ABSFOM	PSIZERO × 10^{-4}	RESID
1	0·2178	0·2852	47·75
2	0·2661	0·4091	48·21
3	0·2189	0·3166	49·81
4	0·3520	0·4760	45·96
5	0·3998	0·5411	47·23
6	0·2238	0·3068	49·20
7	0·4131	0·5239	44·84
8	0·2340	0·3084	48·74
9	0·2537	0·3347	49·95
10	0·4142	0·5378	46·21
11	0·2168	0·3187	49·25
12	0·3730	0·5257	49·64
13	0·2125	0·3342	52·39
14	0·2443	0·3135	46·91
15	0·3014	0·4735	47·76
16	0·2217	0·2922	48·94

contained a single significant feature corresponding to the correct structure (Table 4 and Fig. 1). The interpolated atomic position was 1–2 Å from the refined minimum, a distance well within the requirement of successful least-squares refinement at this resolution.

ABSFOM most clearly denoted the best phase sets. Inspection of the remaining E maps, including set (4) indicated by RESID, revealed that none contained the correct structure. The highest spurious peak in any map was roughly half that of the site in the 'correct' maps, whilst several maps contained no significant features.

The data were restricted to smaller Bragg angles in two further calculations. At 9 Å resolution the phase sets indicated by ABSFOM (and, in fact, RESID) again corresponded to the true structure. At 12 Å resolution none of the E maps contained the true structure. The ratio of triple-product relations to reflections was ten and this should be sufficient to allow adequate phase development. The difficulty most probably arises from the large solvent contributions to 'isomorphous' differences of very low Bragg angle. It may be advantageous to omit entirely those terms corresponding to a spacing greater than about 20·0 Å from the direct-methods analysis.

The application to the Hg derivative gave less convincing results. The highest 300 terms corresponding to spacings greater than 6 Å were phased. All the phase sets had similar figures of merit. An E map was computed for each of the sixteen phase sets, and all

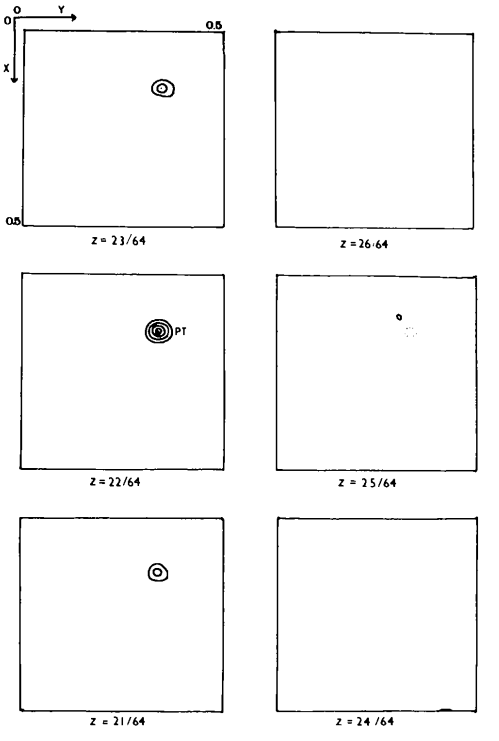


Fig. 1. Six sections of the E map calculated with phase set (5) for the platinum derivative of phosphorylase b. The contour interval is 100 density units. The zero level is omitted and dotted contours represent negative regions. On only two of the remaining sections were any contour levels present.

Table 4. *The parameters for the derivatives of phosphorylase b from E maps and after least-squares refinement at 6 Å resolution with the first data sets (see text)*

The E map for the platinum derivative corresponds to phase set (5) of Table 3. Both E maps were computed on the same scale, had a background level of 100 density units, and have been transformed to the origin and enantiomorph of the refined parameters. The refined occupancies are on an arbitrary scale.

		E map			Least-squares refinement				
	Peak	Fractional coordinates			Fractional coordinates				
Site	height	x	y	z	x	y	z	Occupancy	B (Å^2)
Diaminodichloroplatinum (R_c = 55·7%)									
1	509	0·155	0·355	0·345	0·157	0·359	0·345	0·362	156
Ethylmercury thiosalicylate (R_c = 54·1%)									
1	–				0·195	0·224	0·038	0·132	45
2	188	0·023	0·094	0·400	0·028	0·088	0·385	0·213	89

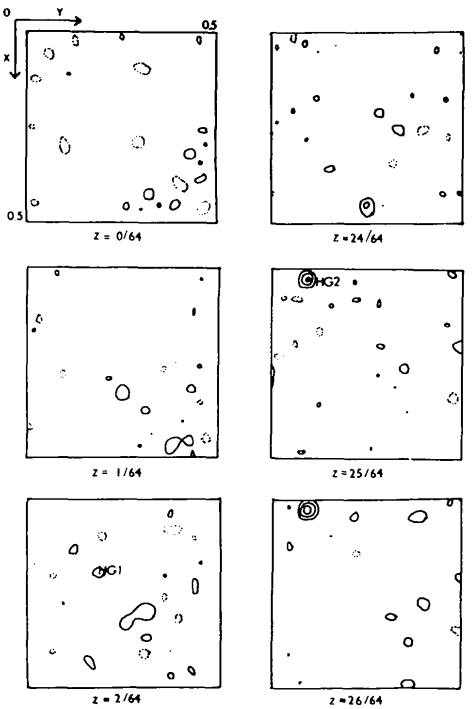


Fig. 2. Six sections of the E map containing a peak at the major site for the mercury derivative of phosphorylase b. The same scale factor was used in the Fourier transformation as in Fig. 1. The contour level is 50 density units. The zero level is omitted and dotted contours represent regions of negative density. There is no significant feature at the position corresponding to the second site (Table 4).

had a background level of 100 density units. The highest peak (188 units) occurred in E map (11), roughly 2 Å from the most highly occupied site (Table 4 and Fig. 2). There was no significant density at the position of the second site. Least-squares refinement of the first site proceeded to the correct minimum and allowed identification of the second site in the difference synthesis.

This 'correct' phase set was not denoted as best by any of the figures of merit. However, the next highest feature in any map was of height 167 units and did not correspond to a true structural feature. This suggests that an additional criterion for the selection of the best phase set might be the one which gives an E map with the most significant feature.

An extensive series of computations was carried out in an attempt to improve upon this result. The following variations on the first calculations were tried, both singly and in some of the many possible combinations.

(1) Modification of the number of terms phased in the range 100–300.

(2) Restriction of the data to a spacing greater than 9 Å to increase the density of sampling of the sphere of reflection and hence the proportion of phase relations with high reliability.

(3) Exclusion of those data where F_P or F_{PH} was small (up to a maximum of half of the data were excluded) as these terms possessed the largest fractional errors.

(4) Use of the isomorphous and anomalous information in the estimation of F_H (Dodson & Vijayan, 1971).

(5) Variation of the number of atoms assumed to be present in the range 8–80 per unit cell which caused the reliabilities of the phase relations to change by a factor of three.

(6) Variation of the number and identity of the terms in the starting set.

(7) Rejection of triple-product relations involving three reflections in the same centric zone.

In none of these trials was more structural information obtained than in the original calculation. In most calculations no E map contained a peak at either Hg site. The initial phase development involved a fortuitous choice of starting set and other parameters.

The Hg sites do not appear to lie on special positions of the space group or in special relationship to one another. The problem in the phasing may lie in the

Table 5. *The parameters for the derivatives of phosphorylase b determined using the photographic data from tangent-formula calculations at* 6 Å *and least-squares refinement at* 3 Å *resolution.*

The E maps were calculated on the same scale as in Table 4. Other details as in Table 4.

	E map				Least-squares refinement				
	Peak	Fractional coordinates			Fractional coordinates				
Site	height	x	y	z	x	y	z	Occupancy	B (Å^2)
Diaminodichloroplatinum ($R_c = 51{\cdot}3\%$)									
1	603	0·154	0·358	0·343	0·150	0·364	0·340	0·515	10
2					0·172	0·360	0·351	0·328	10
3	151	0·255	0·379	0·255	0·252	0·382	0·254	0·201	25
Ethylmercury thiosalicylate ($R_c = 43{\cdot}34\%$)									
1	169	0·198	0·227	0·036	0·195	0·225	0·038	0·461	32
2	472	0·030	0·090	0·386	0·028	0·089	0·384	0·648	27
3	–		–		0·158	0·380	0·325	0·187	25

Table 6. *The analysis of the derivatives of phosphoglycerate kinase*

The refined parameters are taken from Blake, Evans & Scopes (1972). The refined occupancies are given in electrons. The origin and enantiomorph of the E maps have been transformed to correspond to the refined models. The E maps were all computed on the same scale and had a background of 100 density units.

		E map			Least-squares refinement			
	Peak	Fractional coordinates			Fractional coordinates			Occupancy
Site	height	x	y	z	x	y	z	(electrons)
Ethyl mercury phosphate ($R_c = 24\%$)								
1	291	0·946	0·000	0·869	0·943	0·000	0·876	62·4
2	333	0·904	0·431	0·713	0·908	0·430	0·711	65·6
3	430	0·475	0·069	0·769	0·480	0·071	0·776	78·9
Mercury *p*-chlorobenzoate ($R_c = 59\%$)								
1	521	0·952	0·997	0·887	0·961	0·997	0·867	25·2
2	152	0·883	0·445	0·741	0·887	0·448	0·728	16·6
Gold cyanide ($R_c = 41\%$)								
1	129	0·925	0·989	0·894	0·945	0·999	0·868	30·6
2	487	0·444	0·120	0·744	0·439	0·120	0·763	54·6
3	–		–		0·589	0·058	0·154	29·2
4	180	0·038	0·487	0·862	0·072	0·480	0·880	28·0
5	146	0·054	0·120	0·788	0·062	0·114	0·798	15·7

quality of the data: good enough to specify a single major site but not two, or perhaps with errors in just the wrong terms! This is confirmed by recent calculations with data of considerably higher quality collected on an oscillation camera.

Identical calculations to those described above were carried out with the new data. For both derivatives the results were significantly better. The phase sets indicated by ABSFOM contained the correct structures for both derivatives (Table 5). For the Hg the two major sites were both present and for the Pt the major site and a minor site located during the high-resolution study (unpublished results).

This appears to confirm that the method is limited in its application by experimental errors in the moduli F_P and F_{PH}.

3.2. *Phosphoglycerate kinase*

6 Å data for the native protein and three derivatives were provided by Drs C. C. F. Blake and P. R. Evans. The derivatives were EMP with three binding sites, PCMB with two and AUCN with five.

The refined heavy-atom parameters are given in Table 6. The R factors show that the derivative quality decreases in the order: EMP $\gg$ AUCN > PCMB. This reflects the high occupancy of the EMP sites and the imperfect isomorphism and large observational errors for the other two derivatives. The interpretation of the difference Patterson synthesis was simple for EMP. *Ab initio* interpretation of the Patterson synthesis for AUCN or PCMB would have been difficult (Evans, private communication): the derivatives were analysed from difference Fourier syntheses.

The 200 highest E_h were phased for each derivative. Two atoms per asymmetric unit were assumed in the calculation. An E map was computed for each phase set on the same arbitrary scale. The background level in all the maps was 100 density units. The results are summarized in Table 6.

For the EMP derivative twelve phase sets from the sixteen generated possessed roughly the same 'best' values of ABSFOM and RESID: 0·86 and 22% respectively. The E map with the highest value of ABSFOM contained three significant features, which were all within 0·5 Å of real atomic positions. The relative peak heights in the map were close to the refined occupancies.

For PCMB a similar group of 'best' phase sets was clearly indicated by ABSFOM and RESID. The value of ABSFOM (0·68) was lower and of RESID (35%) higher than for EMP. In the E map corresponding to the highest value of ABSFOM there were two significant features with peak heights of 521 and 152 units. These corresponded to the two sites of Evans's model.

In Evans's analysis the occupancies of the two sites were 25·2 and 16·6 electrons. This difference was emphasized by the squaring effect inherent in the tangent formula. The domination of the E map with tangent phases by the largest site was more marked for the poor quality PCMB than for EMP.

For AUCN all eight phase sets had closely similar values of ABSFOM and RESID. The values again indicated poorer phase sets than for EMP, comparable with PCMB. All eight maps contained a single large peak of height 400–470 units. In no map did this coincide with a real atomic position. The peaks generally had the same x coordinate as site B, but with

z displaced from $\frac{3}{4}$ to $\frac{1}{8}$ or 0. In none of the E maps did the major feature correspond to the largest peak in the Harker section of the Patterson synthesis.

As the AUCN derivative was known to contain five sites per asymmetric unit two more experiments were carried out. These were identical to those described above in all respects but the assumption of first eight and then twenty atoms per asymmetric unit. The effect of increasing the number is to decrease the reliability estimates for the phase relations (see § 4). This confers more flexibility upon the phase development, and in particular may allow erroneous indications in the early stages to be modified to improved, more consistent, values as the phase determination proceeds. This effect was apparent in the increase of ABSFOM for the best set from 0·68 through 0·85 to 1·16 as the number of atoms was increased from two to eight to twenty.

The E maps for all the phase sets again each contained a single large peak. In one map in the eight-atom and two maps in the twenty-atom experiments this peak corresponded to the major site, B. Inspection of the E maps with the correct major peak revealed three other significant features. These closely corresponded to sites A, D and E. There was no significant density at site C.

The only indication that these three phase sets were optimal was given by the PSIZERO figure of merit. Thus PSIZERO may be useful in precisely those situations where ABSFOM and RESID are similar for all the phase sets. However, PSIZERO indicated two other, incorrect, phase sets as equally good.

For completeness similar calculations were performed for EMP and PCMB. The correct structures were obtained whether two, eight, or twenty atoms per asymmetric unit were assumed.

The result for AUCN is at first sight disappointing. It is in exactly this situation with a multi-site derivative and a complex Patterson synthesis that the tangent formula should be most useful. However, it is probable that a combination of the direct-methods and Patterson information would have led to the correct structure: only the correct E maps predicted the largest peak in the Harker section, and closer inspection of these maps revealed three more of the sites.

The results suggest that one parameter which may be usefully varied is the number of atoms assumed, and emphasize the need to show that the largest peaks in the E maps do correspond to the largest peaks in the Patterson synthesis.

3.3. *Hagfish insulin*

Data for the native protein and two single-site derivatives, lead acetate and uranyl acetate, were made available by J. Cutfield and G. Dodson. The large observational errors are shown by the high R factors after the refinement of the derivatives: about 55% for the Pb and 65% for the uranyl using the centric terms (J. Cutfield, private communication).

For both derivatives the 61 normalized moduli greater than unity were phased, and an E map was computed with the phase set that had the highest ABSFOM. The synthesis for the Pb derivative contained a single significant feature three times, and for the uranyl a single feature five times, the background level.

The coordinates for the sites after least-squares refinement and from the E maps are given in Table 7. The largest discrepancy lies in the z coordinate of the uranyl site which in the E map is displaced by 1·8 Å from its refined position. This is most probably a result of the proximity of the site to the twofold axis, and hence to a symmetry-related atomic position, from which it is less than 7 Å.

The E map clearly defined the location of the uranyl site relative to the origin, with all three coordinates

Table 8. *Distribution of* $\langle E_{\mathbf{h}}^2 \rangle$ *with parity group for the triose phosphate isomerase isomorphous differences*

The numbers are the mean values of $E_{\mathbf{h}}^2$, which are normalized to unity overall.

Parity group	PCMB	EMP	BAKERS	$PtCl_4$
EEE	1·7463	1·3071	1·6367	1·3137
0*EE*	0·5810	0·8466	0·8886	0·9563
*E*0*E*	0·4987	0·7626	0·9193	0·8062
00*E*	1·2198	1·0608	0·9617	1·2060
*EE*0	0·5400	0·8549	0·6910	1·0903
0*E*0	1·3984	1·1208	1·0437	0·7477
*E*00	1·2581	1·2229	1·1362	0·8656
000	0·6163	0·6612	0·6228	0·9202

Table 7. *The parameters after least-squares refinement and from the E maps for the lead and uranyl derivatives of hagfish insulin at* 6 Å

The E maps were computed on the same scale, and have been scaled so that the height of the uranyl site corresponds to the refined occupancy.

		E map			Least-squares refinement			
	Peak	Fractional coordinates			Fractional coordinates			
Site	height	*x*	*y*	*z*	*x*	*y*	*z*	Occupancy
Lead	0·25	0·43	0·18	0·185	0·43	0·17	0·185	0·27
Uranyl	0·40	0·01	0·03	0·015	0·01	0·06	0·036	0·40

positive and y greater than x. The Patterson synthesis did not allow such a straightforward positioning of this site. The tangent formula provided a useful aid in the analysis of the derivatives of hagfish insulin.

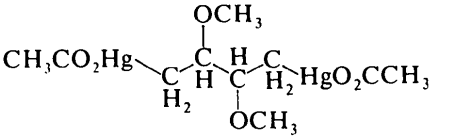
3.4. *Triose phosphate isomerase*

Data corresponding to a spacing greater than 6 Å for the native protein and four isomorphous derivatives were provided by Professor D. C. Phillips and co-workers. In each derivative at least one site had one or more fractional coordinate close to $\frac{1}{4}n$ of the cell edge, where n is integral (Bloomer, 1972). These will henceforth be termed 'special' positions. This led to problems in the analysis and refinement by conventional techniques. The interpretation of the Patterson functions led to homometric sets of structures which required correct positioning of atoms relative to these 'special' positions for effective discrimination.

The isomorphous differences were normalized for each derivative. The irregular distribution of the value of $\langle E_{\mathbf{h}}^2 \rangle$ with parity group is symptomatic of the 'special positions' of the atoms (Table 8). The 200 largest terms were phased and an E map was computed for the phase set with the highest value of ABSFOM. The results are summarized in Table 9.

For the PCMB and EMP derivatives the significant features in the maps represented the complete heavy-atom structures.

The Bakers' dimercurial molecule has the structure:



with two Hg atoms roughly 5·5 Å apart in the most probable conformation. The low-resolution model for the derivative comprised two pairs of sites, each pair representing one dimercurial molecule.

The two significant features in the E map corresponded to the pairs of sites in the model. The peaks are spherically symmetric and give no indication that they represent two discrete atoms. The centre of each peak was closer to the larger of the pair of atoms.

There were two significant features in the E map for the $PtCl_4$ derivative, corresponding to the two major sites. There was no evidence of the four minor sites. In the structural analysis these minor sites were not located from the Patterson synthesis, but rather from the difference Fourier synthesis.

The coordinates obtained from the E maps were closely similar to the refined models. The correct location of the atoms relative to the special positions was clearly shown for all but the Bakers' dimercurial derivative and the x coordinate of site A of the EMP derivative. This is a fundamental improvement on the analysis from the Patterson syntheses which did not completely resolve this problem.

Table 9. *The analysis of the derivatives of triose phosphate isomerase after least-squares refinement (D. C. Phillips, P. S. Rivers & I. A. Wilson, personal communication) and from the E maps*

The E maps were all computed on the same scale, had a background of 40 units, and have been transformed to the origin and enantiomorph of the refined parameters.

		E map			Least-squares refinement				
	Peak	Fractional coordinates			Fractional coordinates				
Site	height	x	y	z	x	y	z	Occupancy	B (Å^2)
Mercury *p*-chlorobenzoate (R_c = 40·6%)									
1	172	0·124	0·008	0·744	0·124	0·012	0·737	0·242	17
2	184	−0·006	0·245	0·166	−0·007	0·244	0·167	0·268	17
Ethyl mercury phosphate (R_c = 45·0%)									
1	66	0·115	0·016	0·737	0·126	0·015	0·741	0·238	33
2	148	−0·009	0·245	0·148	−0·010	0·243	0·160	0·277	5
3	119	0·009	0·047	0·047	0·007	0·051	0·049	0·239	17
4	98	0·061	0·189	0·666	0·060	0·194	0·671	0·247	17
Bakers' dimercurial (R_c = 44·4%)									
1	160	0·115	0·008	0·740	0·122	0·010	0·702	0·256	26
2	–		–		0·080	−0·027	0·789	0·178	13
3	66	−0·004	0·253	0·172	−0·003	0·244	0·174	0·236	41
4	–		–		0·045	0·270	0·160	0·135	80
Tetrachloroplatinum (R_c = 49·8%)									
1	85	0·122	−0·005	0·017	0·125	−0·001	0·014	0·307	80
2	137	0·454	0·276	0·219	0·453	0·270	0·220	0·300	60
3	–		–		0·194	0·252	0·029	0·163	96
4	–		–		0·236	0·496	0·319	0·072	48
5	–		–		0·350	0·123	0·022	0·074	9
6	–		–		0·438	−0·007	0·195	0·047	4

For each derivative the phase set indicated by ABSFOM corresponded to the correct structure. RESID indicated the same 'best' phase set for the EMP and PCMB, but a different, and incorrect, set for the Bakers' and $PtCl_4^{2-}$ derivatives.

4. Conclusion

MULTAN has been shown to be generally useful in the rapid elucidation of the heavy-atom binding in isomorphous derivatives. Normalization of the isomorphous differences, assuming two atoms per asymmetric unit, phase determination with *MULTAN* and computation of an E map for the phase set with the highest value of ABSFOM led directly to an essentially correct structure for nine out of the eleven derivatives studied.

The three-dimensional data need only extend to a maximum Bragg angle sufficient to resolve the atomic sites. Thus for the Pt derivative of phosphorylase b (a single-site derivative) data corresponding to a maximum spacing of 9 Å proved sufficient. This will be vital for application to problems with large unit cells. Use of the data to a spacing of 12 Å did not lead to the correct structure. This may be dependent upon the effect of solvent changes at low Bragg angle.

The ABSFOM figure of merit has proved itself an excellent indicator of the best phase set. RESID substantiated the best set in a number of examples – but gave invalid indications in the others. PSIZERO in general presented the inverse, and incorrect, indications to ABSFOM. Only for the AUCN derivative of phosphoglycerate kinase, where ABSFOM and RESID were similar for all the phase sets, did PSIZERO provide valid information regarding the phases.

The referee has indicated that PSIZERO depends upon reflections with $|E| \doteqdot 0$. These cannot be identified among the E's used because a zero value of ΔF does not necessarily correspond to a zero heavy-atom contribution. Very little (or no) reliability would, therefore, be expected for PSIZERO as a figure of merit under these conditions – as was found in the results described above.

Inspection of the E maps suggested that the synthesis containing the highest feature is usually the closest approximation to the true structure. However, this is by no means an infallible guide, and will always be suspect if different large peaks of similar height appear in different E maps or if the peak lies on a special position of the space group.

The relationship between the Patterson synthesis and the E maps is of interest. For each of the 'correct' E maps the largest features provided an excellent interpretation of the Patterson synthesis, that is they explained the majority of the higher peaks. For 'incorrect' E maps which contained significant features at sites other than the real atomic positions, these features did not correspond to the larger features of the Patterson synthesis. This is an enigmatic result. The information used in the phase determination by *MULTAN* and in the computation of a Patterson synthesis is the same – namely, the amplitudes of the structure factors. The difference lies in the effect of observational errors in the values of $|F_P|$ and $|F_{PH}|$ (and hence in ΔF) upon the methods. For the direct method errors in the terms involved in the first steps of the phase propagation may cause a completely incorrect pathway to be followed. For the Patterson syntheses one hopes that a small number of erroneous values will be averaged out during the summation over a large number of terms. I conclude that a useful check on the validity of the information from the E maps is its ability to rationalize the Patterson synthesis.

It may be of interest to examine more closely the two examples where the approach was least successful, namely the Hg derivative of phosphorylase b and the AUCN derivative of phosphoglycerate kinase.

For the AUCN derivative the effect of varying the numbers of atoms assumed to be present in the asymmetric unit was investigated. For the high-quality EMP (and lower-quality PCMB) the correct structure was evident in the E maps whether two, eight or twenty atoms were assumed. For AUCN the calculation with two atoms did not lead to the correct structure. However, AUCN did indeed contain five independent sites and was not highly isomorphous with the native protein.

The effect of varying the number of atoms, N, is to change the reliability estimates $\kappa(\mathbf{h},\mathbf{k})$ associated with the triple-product relations. We see:

$$\kappa(\mathbf{h},\mathbf{k}) = 2\sigma_3\sigma_2^{-3/2}|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|, \tag{4.1}$$

where

$$\sigma_n = \sum_{j=1}^{N} Z_j^n, \tag{4.2}$$

and Z_j is the atomic scattering factor of the jth of N atoms. On the assumption of equi-sized atoms this reduces to:

$$\kappa(\mathbf{h},\mathbf{k}) = \frac{2}{N^{1/2}}|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|. \tag{4.3}$$

Increasing N thus reduces the reliabilities. Improved phase determination on such a reduction may reflect a truly better estimate of κ (*i.e.* a better guess at the number of atoms) or an effective amelioration of the problem of observational errors in the $E_{\mathbf{h}}$. A large error in the observed magnitude of an $E_{\mathbf{h}}$ involved in the early stages of phase determination will cause phase errors to be propagated throughout the phase set. If high values of κ are assumed then once such an incorrect indication is accepted, phase generation will irrevocably follow the wrong pathway. A lower value of κ may confer additional flexibility on the phasing and

allow effective discrimination against such erroneous indications.

The problem of observational errors in $|F_P|$ and $|F_{PH}|$, and hence in E_h, is likely to be the major limitation to the technique, and is presumed to underlie the difficulties encountered with the Hg derivative of phosphorylase b. Errors in the observed magnitudes of the E_h will reduce the reliability of the triple-product phase relations. To arbitrarily increase the value of N is not the ideal solution to this problem, as the ensuing fractional reduction in $\kappa(\mathbf{h},\mathbf{k})$ is independent of the moduli and identity of the terms involved – while the errors will almost certainly be dependent upon them. Furthermore this approach does not provide an objective evaluation of the reliability.

A fundamental improvement should result if the errors inherent in the observations and those inherent in the relations were to be combined to produce an overall reliability measure. This would allow a rational decision as to whether phase determination by *MULTAN* would be likely to produce satisfactory results for a particular problem. A modification to the tangent formula which will allow for experimental error in the observations will be proposed later (French & Wilson, in preparation).

I acknowledge the support of the Medical Research Council during the course of this study. I am grateful to Drs C. Blake, J. Cutfield, G. Dodson and P. R. Evans, and to Professor D. C. Phillips and co-workers for making their data freely available. I thank Professor Phillips for the pleasure of working in his laboratory and Dr L. N. Johnson for many helpful discussions.

References

BLAKE, C. C. F., EVANS, P. R. & SCOPES, R. K. (1972). *Nature (London) New Biol.* **235**, 195–198.

BLOOMER, A. C. (1972). DPhil Thesis, Oxford Univ.

DODSON, E. & VIJAYAN, M. (1971). *Acta Cryst.* B**27**, 2402–2411.

GERMAIN, G., MAIN, P. & WOOLFSON, M. M. (1970). *Acta Cryst.* B**26**, 274–285.

GERMAIN, G., MAIN, P. & WOOLFSON, M. M. (1971). *Acta Cryst.* A**27**, 368–376.

GERMAIN, G. & WOOLFSON, M. M. (1968). *Acta Cryst.* B**24**, 91–96.

HAUPTMAN, H. & KARLE, J. (1953). *Solution of the Phase Problem. I. The Centrosymmetric Crystal.* ACA Monograph No. 3.

JOHNSON, L. N., MADSEN, N. B., MOSLEY, J. & WILSON, K. S. (1974). *J. Mol. Biol.* **90**, 703–717.

NAVIA, M. A. & SIGLER, P. B. (1974). *Acta Cryst.* A**30**, 706–712.

SCHEVITZ, R. W., NAVIA, M. A., BANTZ, D. A., CORMICK, G., ROSA, J. J., ROSA, M. D. H. & SIGLER, P. B. (1972). *Science*, **177**, 429–431.

STEITZ, T. A. (1968). *Acta Cryst.* B**24**, 504–507.

Acta Cryst. (1978). B**34**, 1608–1612

The Crystal Structures of Trimesic Acid, its Hydrates and Complexes. III.* Trimesic Acid–H_2O–1,4-Dioxane

BY F. H. HERBSTEIN AND M. KAPON

Department of Chemistry, Technion–Israel Institute of Technology, Haifa, Israel

(*Received* 28 *October* 1977; *accepted* 2 *December* 1977)

$C_9H_6O_6.H_2O.C_4H_8O_2$ is triclinic, $a = 9{\cdot}528\ (3)$, $b = 9{\cdot}535\ (3)$, $c = 8{\cdot}031\ (3)$ Å, $\alpha = 89{\cdot}68\ (5)$, $\beta = 95{\cdot}01\ (5)$, $\gamma = 92{\cdot}03\ (4)°$, $Z = 2$, space group $P\bar{1}$. The structure was refined to $R = 0{\cdot}113$ for 2890 counter reflections. It consists of planar ribbons of composition TMA.H_2O extending along [100]; these ribbons are hydrogen-bonded to one another *via* two crystallographically independent groups of bridging dioxane molecules. The framework formed in this way can be described in terms of an infinite series of steps running approximately about the (011) planes.

1. Introduction

Trimesic acid (TMA) forms complexes with many organic molecules. The two structures reported to date (Herbstein & Marsh, 1977) are of the channel inclusion type without hydrogen bonding between host and guest. However, other arrangements are possible and we are studying some of them to obtain an overall view of the crystal chemistry of TMA molecular complexes.

TMA crystallized from dioxane gives a ternary

* Part II: Herbstein & Marsh (1977).