Acta Cryst. (1990). A46, 560-567

Direct Methods with Single Isomorphous Replacement Data. I. Reduction of Systematic Errors

BY W. FUREY JR,* K. CHANDRASEKHAR, F. DYDA† AND M. SAX

Biocrystallography Laboratory, PO Box 12055, *Veterans Administration Medical Center, University Drive C, Pittsburgh, PA* 15240, *USA and Department of Crystallography,* 304 *Thaw Hall, University of Pittsburgh, Pittsburgh, PA* 15260, *USA*

(*Received* 20 *March* 1989; *accepted* 28 *February* 1990)

Abstract

The direct-methods procedure for single isomorphous replacement (SIR) data [Hauptman (1982). *Acta Cryst.* A38, 289-294], as modified by Fortier, Moore & Fraser [*Acta Cryst.* (1985), A41, 571-577] has been implemented and tested with a large number of known structures. It was found that the modified procedure greatly reduces the bias toward 'unresolved' SIR invariant values associated with estimates of 0 or π, but does not remove it entirely. If the heavy atoms are not in a centrosymmetric array the centroid of the distribution of invariant estimates is not centered on true protein values, but is biased toward conventional SIR values by up to 15°, thus errors in the estimates are not random but systematic. When the heavy atoms are in a centrosymmetric array (or single heavy-atom site in space group $P2_1$), the distribution of estimates is often sharply bimodal, with peaks centered at both true invariant values and pure 'unresolved' SIR values. Simple procedures are given which can be applied in both situations to reduce significantly the bias with no overall loss of accuracy. An additional correction factor is then described which can be used to remove nearly all of the bias, and improve the accuracy as well. The result is that errors in the corrected invariant estimates are small in magnitude, but are now also random instead of systematic. Since the number of estimates greatly exceeds the number of phases, the remaining random errors should have little impact in phasing processes.

Introduction

In recent years, theoretical developments in the area of direct methods as applied to protein crystallography have advanced considerably. In particular, a theory for the integration of direct methods with single isomorphous replacement (Hauptman, 1982) looked very promising in that it was possible accurately to identify large numbers of three-phase structure invariants with values of 0 or π, even for very large structures. Other procedures capable of identifying invariants with values of 0 or π from single-isomorphous-replacement data were also developed (Karle, 1983; Giacovazzo, Cascarano & Zheng, 1988). Unfortunately, it was shown (Xu, Yang, Furey, Sax, Rose & Wang, 1984) that invariant values of 0 or π are not particularly useful for protein crystallography since they generally correspond to the heavy-atom invariants (or heavy-atom invariants plus π) of the included derivative. Any procedure which forces individual phases to satisfy such invariants therefore results in producing classical 'unresolved' SIR (single isomorphous replacement) phases, since the invariants themselves are actually SIR invariants (*e.g.* invariants produced by summing over three SIR phases). The realization of the correspondence with SIR phases prompted a re-examination and modification of Hauptman's formulation (Fortier, Moore & Fraser, 1985) resulting in a new procedure which should be considerably more powerful. With this modification it is possible accurately to identify large numbers of invariants with absolute values anywhere in the range 0-π, however, only the magnitude of the angle can be identified (*i.e.* cosine invariant). By moving away from 0 and π values the bias toward SIR invariants should be diminished and the resulting estimates should become more useful for the determination of individual protein phases.

In all previous studies the proposed methods were tested with error-free data, usually for a single structure; thus the general applicability has not been demonstrated. In the current study we have applied the modified formulation of Fortier, Moore & Fraser to numerous structures taken from the Protein Data Bank (Bernstein *et al.*, 1977) to determine whether the accuracy of the estimates is sensitive to space group, structure size and heavy-atom substitution parameters. It was found that although the Fortier modification greatly reduces the bias towards SIR invariants, it does not remove it entirely since a residual bias of up to 15° remains. Several alternative modifications to the procedure are now reported, all of which lead to further reductions in the bias towards SIR, and one which can significantly improve the accuracy of the estimates as well. With the new

* To whom correspondence should be addressed.

† In partial fulfilment of the Doctor of Philosophy Degree.

0108-7673/90/070560-08$03.00

modifications the errors have been 'randomized' whereas they originally were systematic. Since the number of estimates greatly exceeds the number of phases, the remaining random error should have little impact on individual phase determination.

Definitions

$|F_{\mathbf{H}}|_P, \varphi_{\mathbf{H},P}$ Native protein structure-factor amplitude and phase, respectively, for reflection **H**.

$|F_{\mathbf{H}}|_H, \varphi_{\mathbf{H},H}$ Heavy-atom structure-factor amplitude and phase, respectively, for reflection **H**.

$|F_{\mathbf{H}}|_D, \varphi_{\mathbf{H},D}$ Derivative structure-factor amplitude and phase, respectively, for reflection **H** such that

$$|F_{\mathbf{H}}|_D \exp(i\varphi_{\mathbf{H},D}) = |F_{\mathbf{H}}|_P \exp(i\varphi_{\mathbf{H},P}) + |F_{\mathbf{H}}|_H \exp(i\varphi_{\mathbf{H},H}).$$

$\varphi_{\mathbf{H},S}$ 'Unresolved' SIR phase for reflection **H**. This is the classical SIR phase and corresponds to the centroid of the bimodal SIR probability distribution. In general, it is the heavy-atom phase or the heavy-atom phase $+\pi$ according to

$$\varphi_{\mathbf{H},S} = \varphi_{\mathbf{H},H} \quad \text{if } |F_{\mathbf{H}}|_D > |F_{\mathbf{H}}|_P$$

and

$$\varphi_{\mathbf{H},S} = \varphi_{\mathbf{H},H} + \pi \quad \text{if } |F_{\mathbf{H}}|_D < |F_{\mathbf{H}}|_P.$$

$\psi_P = \varphi_{\mathbf{H},P} + \varphi_{\mathbf{K},P} + \varphi_{\mathbf{L},P}$ Protein invariant when $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$

$\psi_H = \varphi_{\mathbf{H},H} + \varphi_{\mathbf{K},H} + \varphi_{\mathbf{L},H}$ Heavy-atom invariant when $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$.

$\psi_S = \varphi_{\mathbf{H},S} + \varphi_{\mathbf{K},S} + \varphi_{\mathbf{L},S}$ SIR invariant when $\mathbf{H}+\mathbf{K}+\mathbf{L}=0$.

ψ_E Estimated value of ψ_P produced by various formulae.

Absolute error Magnitude of the error in ψ_E, given in degrees, equal to $|\psi_P - \psi_E|$ where ψ_P is computed from a known structure and angular differences are measured such that they are always less than or equal to 180°.

Absolute deviation Magnitude of the difference between a reference and target invariant computed as for the absolute error. Reference and target invariants may be ψ_P, ψ_S or ψ_E.

Detection of bias toward SIR invariants

There are several ways to evaluate the accuracy of invariant estimates, some of which are more useful than others. We found that, when dealing with isomorphous-replacement or anomalous-scattering methods, it is not sufficient to compute only absolute errors between invariant estimates ψ_E and their associated true values ψ_P, since it does not detect systematic error towards invariant values corresponding to conventional 'unresolved' phases. Elimination of such systematic error is very important for the following reason. The classical method for determining protein phases from SIR data is simple to apply and yields phases which are already quite accurate when compared only to true values (mean error of about 45° and maximum error of 90°, provided only the sign of $|F_{\mathbf{H}}|_D - |F_{\mathbf{H}}|_P$ is correct). Nevertheless, such phases are not particularly useful since they are severely biased towards the heavy-atom phases and correspond to extremely noisy electron-density maps. On the other hand, multiple isomorphous replacement phases typically have similar error magnitudes when compared to true values, yet yield interpretable electron-density maps. The implication is that the error magnitudes not only must be reasonably small, the distribution of errors must be appropriately centered as well. We interpret this to mean that, for any phasing method based on SIR data alone to offer significant improvements over the conventional method, it must yield a phase-error distribution which not only has a small variance but is also centered on true protein phases with no bias towards classical SIR phases. The same logic is assumed to apply when dealing with invariants rather than individual phases.

To determine the nature of the distribution of errors in the invariant estimates we use the following procedure: for SIR data we compute absolute deviations between (1) estimated and true invariants; (2) estimated and SIR invariants ψ_S; and (3) true and SIR invariants. We then determine distributions of *signed* deviations with respect to both true protein and SIR invariants as follows:

(*a*) Using SIR invariants ψ_S as a reference point, we associate a sign with each absolute deviation (2) such that it is negative if the estimate is in the direction of the true protein invariant and positive if in the opposite direction. A histogram is then accumulated showing the distribution of all *signed* deviations.

(*b*) Using true protein invariants ψ_P as a reference point, we associate a sign with each absolute deviation (1) such that it is positive if the estimate is in the direction of the SIR invariant and negative if in the opposite direction. A histogram is then accumulated showing the distribution of all *signed* deviations.

For each of the distributions (*a*) and (*b*), the mean and variance are computed. For unbiased estimates of protein invariants the distribution (*a*) should be centered on a negative value corresponding to the mean absolute deviation between protein and SIR invariants. The distribution (*b*) should be centered on zero and have a small variance. Any deviation from this pattern would indicate a systematic bias towards (or away from) SIR invariants.

Test results

The procedure described above was applied to ideal (error-free) data for the Bence Jones protein Rhe (Furey, Wang, Yoo & Sax, 1983) (space group $P2_12_12$) after generating estimates of the three-phase invariants and the A values indicative of their variances according to the procedure of Fortier, Moore & Fraser (1985). A 3 site Au derivative was used, with heavy-atom parameters as reported by Wang, Yoo & Sax (1979). Out of the 2 043 196 invariants linking 2213 phases (3 Å data with $E > 0{\cdot}35$), the 27 443 most reliable ($A > 1{\cdot}00$) were used in the test. Since the

procedure produces only magnitude estimates $|\psi_E|$, both possible values were tested for each invariant and the choice resulting in minimum deviation from ψ_P was used. The mean absolute error $\langle|\psi_P - \psi_E|\rangle$ is 23·4°. However, analysis of the *signed* deviations as described above indicates that nearly half of that error can be attributed to a systematic shift towards SIR values. The signed distributions are shown in Fig. 1. The mean signed deviation of +10·22° and the shoulder on the positive side for the distribution with true protein invariants as a reference point illustrate the systematic bias towards SIR invariants. For the original formulation of Hauptman (invariants 0 or π) the situation is much worse, as the distribution is centered precisely on SIR values. Thus the Fortier modification has indeed improved the results, but a systematic bias towards SIR still remains. Further analysis of the results indicate that the bias increases with decreasing A magnitude (Table 1). To determine whether the problem is general in nature or specific for the test structure, extensive test calculations were performed on 3 Å error-free data for 16 additional protein structures, using 33 different heavy-atom-derivative combinations. Coordinates were obtained from the Protein Data Bank (Bernstein *et al.*, 1977), while heavy-atom parameters were taken from the original references given in the data bank. Results are given in Table 2 and show that the bias is generally present, although in varying amounts, regardless of the structure size, space group and number of heavy-atom sites. It was also found that the overall accuracy in the estimates is relatively constant with mean absolute errors of 21 ± 4°, provided an A-value cutoff of 1·0 is used. The mean bias towards SIR values is 7°.

Table 1. *Results from the procedure of Fortier, Moore & Fraser* (1985), *when applied to* 3 Å *Bence Jones protein Rhe data*

The mean A values, mean absolute errors and mean signed deviations (using ψ_P as a reference point, as described in the text) are given for ranges of invariant estimates grouped in descending order of A. All errors are given in degrees. Positive entries in the last column indicate a bias towards 'unresolved' SIR invariant values.

Number of invariants	$\langle A\rangle$	$\langle\lvert\psi_P - \psi_E\rvert\rangle$	$\langle S\lvert\psi_P - \psi_E\rvert\rangle$
2000	3·546	12·76	2·73
2000	2·359	16·26	3·99
2000	1·997	17·78	6·13
2000	1·776	21·11	8·49
2000	1·617	21·55	8·79
2000	1·496	22·38	10·05
2000	1·401	23·09	10·25
2000	1·320	25·27	11·72
2000	1·248	26·18	12·20
2000	1·187	25·76	11·76
2000	1·134	27·41	13·45
2000	1·087	28·84	13·59
2000	1·046	30·08	14·52
1443	1·013	31·00	15·56

Procedure modifications

On analysis of the results, it was apparent that the bias might be correctable, since its magnitude is highly correlated with the A values and is relatively constant across the range of structures examined. Two simple methods were explored to achieve this.

(1) Determine the magnitude of the bias towards SIR as a function of the A value (*via* polynomial fit) with a few known structures. The sign of the correction can be deduced by noting that SIR invariants ψ_S, being heavy-atom invariants or heavy-atom invariants plus π, will tend to have values near 0 or

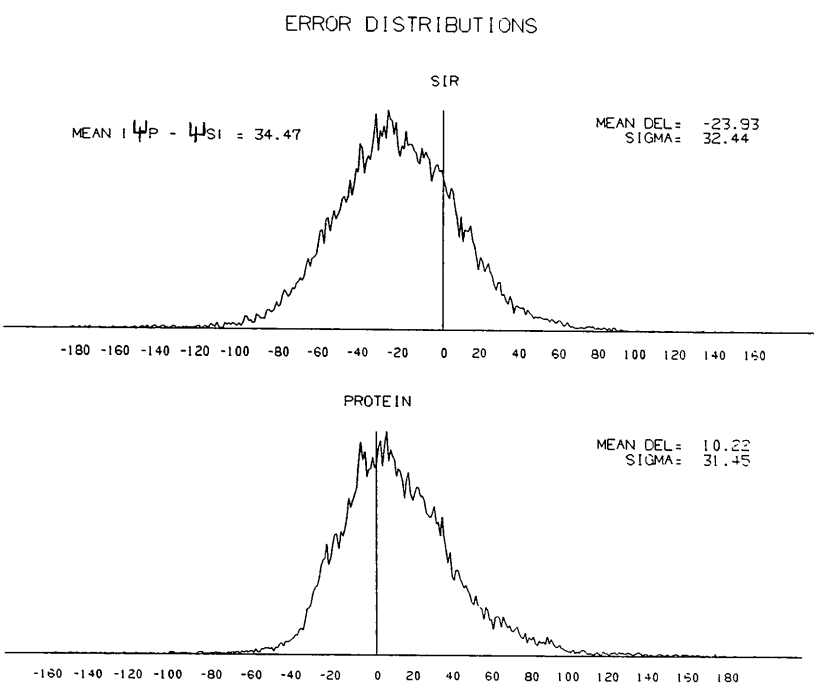

Fig. 1. Distribution of signed errors in three-phase invariant estimates for Bence Jones protein Rhe, when estimates are produced by the Fortier modification of Hauptman's theory. Upper curve uses classical 'unresolved' SIR invariants as reference point. Lower curve uses true protein invariants as reference point.

Table 2. *Summary of results of test calculations on* 16 *structures with* 33 *different heavy-atom substitution patterns, taken from the Protein Data Bank*

When the Fortier modification of Hauptman's formulation is used, the mean absolute error in invariant estimates is 21° (column 9) and the mean bias towards 'unresolved' SIR values is 7° (column 11). After inclusion of the heavy-atom invariant ψ_H as described in this paper, the mean absolute errors and bias towards SIR become 15 and 1°, respectively (columns 10 and 12). Inconsistent triples of the type reported by Han & Langs (1988) were excluded from the results.

Reference No.	Protein	Space group	Number of atoms	Number and type of heavy atom(s)	Number of unique invariants	Number of accepted invariants	Number of phaseable reflections	$\langle\lvert\psi_P-\psi_E\rvert\rangle$ (Fortier) (°)	$\langle\lvert\psi_P-\psi_E\rvert\rangle$ (°)	$\langle s\lvert\psi_P-\psi_E\rvert\rangle$ (Fortier) (°)	$\langle s\lvert\psi_P-\psi_E\rvert\rangle$ (°)
1.	Hen egg white lysozyme	$P2_1$	2002	4; Hg	3 636 804	68 069	2186	19·1	15·2	6·8	0·8
	Hen egg white lysozyme	$P2_1$	2002	7; Hg	3 413 780	24 112	1826	24·6	15·7	10·9	0·9
	Hen egg white lysozyme	$P2_1$	2002	2; U	3 579 012	65 042	1904	17·3	16·7	4·6	2·8
2.	Ribonuclease A	$P2_1$	1039	4; Hg	886 738	10 680	1009	24·1	17·7	11·6	5·0
	Ribonuclease A	$P2_1$	1039	3; U	950 070	17 384	1161	22·6	16·2	10·7	3·4
3.	Human deoxyhemoglobin	$P2_1$	4779	2; Hg	14 949 510	219 085	4774	19·5	15·2	6·8	0·5
4.	Subtilisin novo	$P2_1$	1948	1; Tl	3 464 566	151 006	2277	15·5	15·5	−0·8	−0·8
5.	Alpha chymotrypsin	$P2_1$	3719	7; Hg	13 949 388	115 773	4737	24.5	15·0	10·4	1·9
	Alpha chymotrypsin	$P2_1$	3719	6; U	14 540 117	432 282	5129	20·6	16·3	9·7	3·3
6.	Human carbonic anhydrase C	$P2_1$	2040	1; Au	4 660 100	228 944	3571	16·0	16·3	2·6	2·3
	Human carbonic anhydrase C	$P2_1$	2040	1; Hg	4 642 268	209 689	3543	15·1	15·3	0·2	0·0
	Human carbonic anhydrase C	$P2_1$	2040	3; I	4 917 148	113 687	3688	20·0	15·5	7·7	0·7
7.	Carboxypeptidase A	$P2_1$	2445	1; Hg	6 648 522	287 338	3169	15·9	15·9	2·0	1·9
	Carboxypeptidase A	$P2_1$	2445	2; Pb	6 447 745	105 278	3038	19·5	15·0	7·4	1·3
8.	Acid proteinase	$P2_1$	2732	2; Hg	9 431 160	377 538	4808	18·3	17·3	5·4	3·2
	Acid proteinase	$P2_1$	2732	1; Pt	9 411 317	434 823	4709	17·3	17·2	3·1	3·2
9.	L-Arabinose binding protein	$P2_12_12_1$	2335	4; Pt	13 192 502	118 618	3268	25·4	13·5	10·6	2·3
10.	Phospholipase A2	$P2_12_12_1$	1080	2; Cd	1 865 987	37 889	1413	21·8	13·2	9·5	2·8
	Phospholipase A2	$P2_12_12_1$	1080	2; Pt	1 691 813	31 653	1191	19·1	13·5	8·3	2·9
11.	Triose phosphate isomerase	$P2_12_12_1$	3740	4; Hg	30 397 408	225 779	5161	25·2	12·3	8·5	−2·8
	Triose phosphate isomerase	$P2_12_12_1$	3740	2; Pt	30 278 624	294 871	5058	22·3	12·8	7·6	−2·3
12.	Cytoplasmic malate dehydrogenase	$P2_12_12_1$	4748	3; Hg	21 432 932	111 863	4860	25·6	13·0	10·3	−0·1
	Cytoplasmic malate dehydrogenase	$P2_12_12_1$	4748	8; Pt	21 975 732	28 645	3777	36·7	24·7	14·4	7·3
13.	Tuna cytochrome c	$P2_12_12$	900	1; Pt	1 839 366	50 381	1511	17·7	13·0	1·4	−5·8
	Tuna cytochrome c	$P2_12_12$	900	5; Pt	1 772 406	26 484	1597	26·0	13·3	7·0	−3·7
14.	Staphylococcal nuclease complex	$P4_1$	1151	1; Ba	3 306 780	53 047	1449	18·8	14·3	5·6	−1·1
	Staphylococcal nuclease complex	$P4_1$	1151	1; I	3 177 162	57 243	1521	18·4	13·8	4·8	−1·7
15.	Proteinase A	$P4_2$	1259	1; Hg	4 469 478	97 316	1582	15·8	15·8	0·3	0·3
	Proteinase A	$P4_2$	1259	2; Hg	4 469 473	46 651	2036	22·6	14·7	8·4	−0·2
	Proteinase A	$P4_2$	1259	1; Re	4 246 539	102 457	1614	15·8	15·8	1·1	1·2
16.	Oxidized cytochrome c	$P4_3$	1743	3; Au	5 675 079	85 019	2059	21·0	13·8	7·8	−0·7
	Oxidized cytochrome c	$P4_3$	1743	2; Ir	5 730 515	122 528	2141	20·4	14·1	6·4	−2·2
	Oxidized cytochrome c	$P4_3$	1743	2; Pt	5 516 784	66 301	1941	23·3	13·8	10·5	0·8

Heavy-atom parameters for each of the structures were obtained from the following references:

1. Hogle, J., Rao, S. T., Mallikarjunan, M., Bedell, C., McMullan, R. K. & Sundaralingam, M. (1981). *Acta Cryst.* B**37**, 591-597.
2. Carlisle, C. H., Palmer, R. A., Mazumdar, S. K., Gorinsky, B. A. & Yeates, D. G. R. (1974). *J. Mol. Biol.* **85**, 1-18.
3. Ten Eyck, L. F. & Arnone, A. (1976). *J. Mol. Biol.* **100**, 3-11.
4. Drenth, J., Hol, W. G. J., Jansonius, J. N. & Koekok, R. (1972). *Eur. J. Biochem.* **26**, 177-181.
5. Tulinsky, A., Mani, N. V, Morimoto, C. N. & Vandlen, R. L. (1973). *Acta Cryst.* B**29**, 1309-1322.
6. Liljas, A., Kannan, K. K., Bergsten, P.-C., Waara, I., Fridborg, K., Strandberg, B., Carlbom, U., Jarup, L., Lovgren, S. & Petef, M. (1972). *Nature (London) New Biol.* **235**, 131-137.
7. Quiocho, F. A. & Lipscomb, W. N. (1971). *Adv. Protein Chem.* **25**, 1-78.
8. Jenkins, J. A., Blundell, T. L., Tickle, I. J. & Ungaretti, L. (1975). *J. Mol. Biol.* **99**, 583-590.
9. Gilliland, G. L. & Quiocho, F. A. (1981). *J. Mol. Biol.* **146**, 341-362.
10. Dijkstra, B. W., Kalk, K. H., Hol, W. G. & Drenth, J. (1981). *J. Mol. Biol.* **147**, 97-123.
11. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Wiley, S. G. (1975). *Nature (London)*, **255**, 609-614.
12. Tsernoglou, D., Hill, E. & Banaszak, L. J. (1972). *J. Mol. Biol.* **69**, 75-87.
13. Takano, T., Kallai, O. B., Swanson, R. & Dickerson, R. E. (1973). *J. Biol. Chem.* **248**, 5234-5255.
14. Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E. Jr, Richardson, D. C., Richardson, J. S. & Yonath, A. (1971). *J. Biol. Chem.* **246**, 2302-2316.
15. Brayer, G. D., Delbaere, L. T. J. & James, M. N. G. (1978). *J. Mol. Biol.* **124**, 243-259.
16. Swanson, R., Trus, B. L., Mandel, N., Mandel, G., Kallai, O. B. & Dickerson, R. E. (1977). *J. Biol. Chem.* **252**, 759-775.

π. This is a consequence of the fact that the heavy-atom invariants themselves should be near 0 since they correspond to a structure with only a few atoms in the unit cell. Accordingly, the appropriate sign to use is that which moves the invariant estimate away from the real axis, *i.e.* pushes the estimate away from SIR. Alternatively, one can simply compute the heavy-atom invariants and choose the sign which maximizes the separation. Results for the Bence Jones protein Rhe are shown in Fig. 2 when the estimates are corrected according to

$$|\Psi|_{\text{corr}} = |\Psi_E| + S \times |\text{BIAS}|,$$

where |BIAS| is obtained from the polynomial fit and $S = 1{\cdot}0$ if $|\Psi_E| < 90°$ and $S = -1{\cdot}0$ if $|\Psi_E| > 90°$. The mean absolute error is essentially unchanged (23·8°), but the bias towards SIR is reduced to about +4·0°. A shoulder remains on the positive side, however, as a residual bias still exists.

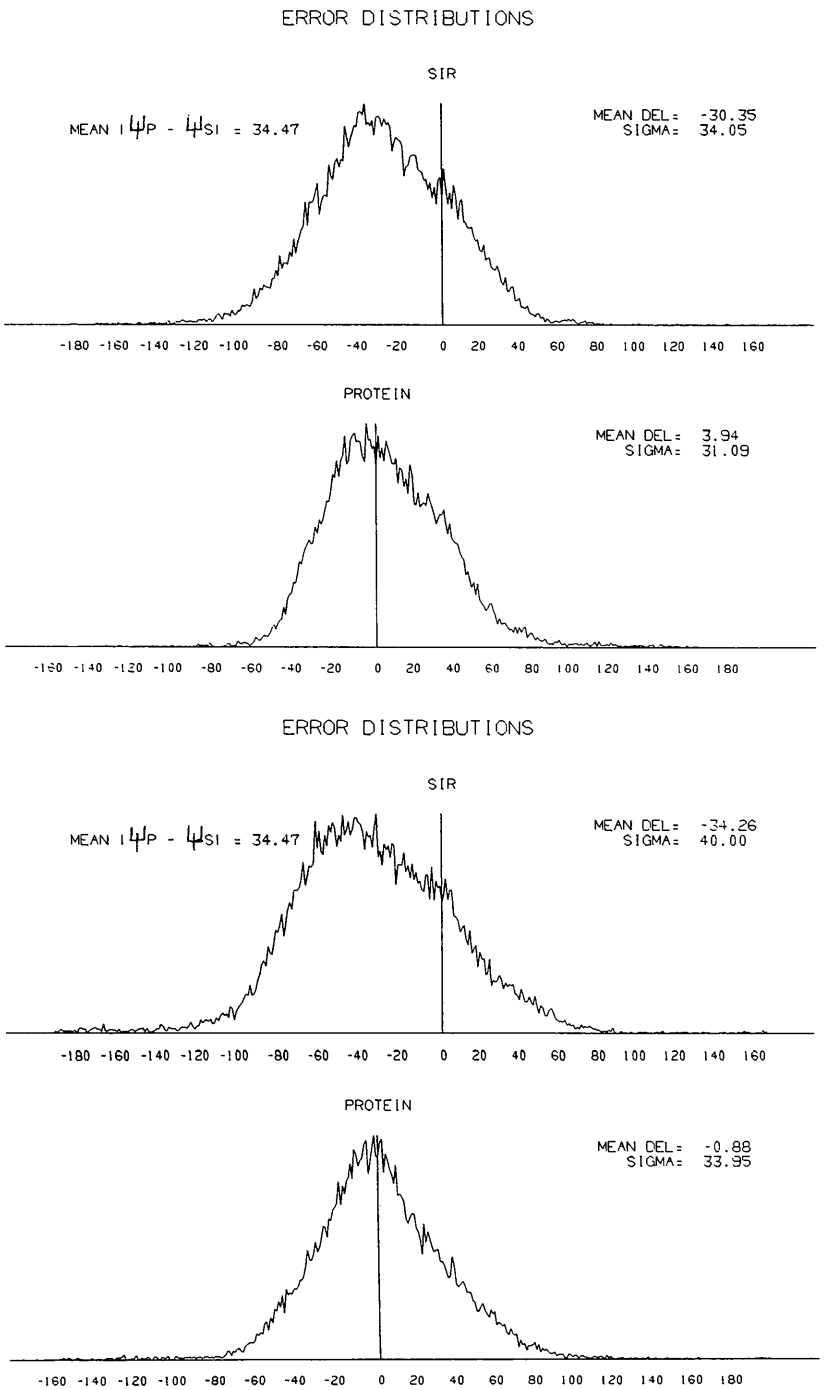

Fig. 2. Distribution of signed errors for Bence Jones protein Rhe (as in Fig. 1), when invariant estimates are corrected according to $|\Psi|_{\text{corr}} = |\Psi_E| + S \times |\text{BIAS}|$, where $S = 1{\cdot}0$ if $|\Psi_E| < 90°$ and $S = -1{\cdot}0$ if $|\Psi_E| > 90°$. |BIAS| is determined from a polynomial fit to A values from several known structures. Note the bias towards SIR has been reduced to +4·0°.

Fig. 3. Distribution of signed errors for Bence Jones protein Rhe (as in Fig. 1), when invariant estimates are obtained by averaging only over the two vectors with greatest deviation from 0 or π. Note that the resulting curve is symmetrical about the true protein invariants, with essentially no bias towards SIR (actually a slight bias of less than 1·0° **away** from SIR). The mean absolute error has only slightly increased (to 25·1°).

(2) The procedure as outlined by Fortier, Moore & Fraser (1985) involves taking a weighted average over four possible vectors to estimate the value of each invariant. Since we know that this results in a bias towards SIR values (and therefore towards 0 or π), we can instead average only over a subset of the four vectors, selectively rejecting those that are close to 0 (or π). The resulting estimate should then slowly move away from the real axis as more contributing vectors are rejected. Best results are obtained when only the two vectors nearest the imaginary axis are used (Fig. 3). Note that in comparison to Figs. 1 and 2 the signed distribution is now centered on the true protein invariants, with little bias towards SIR. Equally important is the fact that the mean absolute error in the estimates has only slightly increased (to 25°), thus we have not introduced any major errors; we merely have 'randomized' what originally was systematic error towards 'SIR'.

Although effective, procedures (1) and (2) are semi-empirical. A more powerful correction may be obtained if one examines the modification procedure of Fortier *et al.*, and recasts it in terms of conventional protein crystallographic notation. For single isomorphous replacement measurements, from a knowledge of native and derivative amplitudes and heavy-atom information, protein phases can be determined from

$$\begin{aligned}\varphi_{\mathbf{H},P} &= \varphi_{\mathbf{H},H} + \cos^{-1}\left[(|F_{\mathbf{H}}|_D^2 - |F_{\mathbf{H}}|_P^2 - |F_{\mathbf{H}}|_H^2)\right. \\ &\quad \left.\times (2|F_{\mathbf{H}}|_P|F_{\mathbf{H}}|_H)^{-1}\right] \\ &= \varphi_{\mathbf{H},H} \pm |\Delta_{\mathbf{H}}|. \end{aligned} \tag{1}$$

By direct substitution, ψ_P is then given by

$$\begin{aligned}\psi_P &= \varphi_{\mathbf{H},H} + \varphi_{\mathbf{K},H} + \varphi_{\mathbf{L},H} \pm |\Delta_{\mathbf{H}}| \pm |\Delta_{\mathbf{K}}| \pm |\Delta_{\mathbf{L}}| \\ &= \psi_H \pm |\Delta_{\mathbf{H}}| \pm |\Delta_{\mathbf{K}}| \pm |\Delta_{\mathbf{L}}|. \end{aligned} \tag{2}$$

Because of the sign ambiguities there are eight versions of (2), only one of which can be correct:

$$\psi_H + |\Delta_{\mathbf{H}}| + |\Delta_{\mathbf{K}}| + |\Delta_{\mathbf{L}}| = \psi_{P1} \tag{2a}$$
$$\psi_H + |\Delta_{\mathbf{H}}| + |\Delta_{\mathbf{K}}| - |\Delta_{\mathbf{L}}| = \psi_{P2} \tag{2b}$$
$$\psi_H + |\Delta_{\mathbf{H}}| - |\Delta_{\mathbf{K}}| + |\Delta_{\mathbf{L}}| = \psi_{P3} \tag{2c}$$
$$\psi_H + |\Delta_{\mathbf{H}}| - |\Delta_{\mathbf{K}}| - |\Delta_{\mathbf{L}}| = \psi_{P4} \tag{2d}$$
$$\psi_H - |\Delta_{\mathbf{H}}| + |\Delta_{\mathbf{K}}| + |\Delta_{\mathbf{L}}| = \psi_{P5} \tag{2e}$$
$$\psi_H - |\Delta_{\mathbf{H}}| + |\Delta_{\mathbf{K}}| - |\Delta_{\mathbf{L}}| = \psi_{P6} \tag{2f}$$
$$\psi_H - |\Delta_{\mathbf{H}}| - |\Delta_{\mathbf{K}}| + |\Delta_{\mathbf{L}}| = \psi_{P7} \tag{2g}$$
$$\psi_H - |\Delta_{\mathbf{H}}| - |\Delta_{\mathbf{K}}| - |\Delta_{\mathbf{L}}| = \psi_{P8}. \tag{2h}$$

The eight versions can be partitioned into two groups which are enantiomorphically related about the heavy-atom invariant $\psi_{\mathbf{H}}$. Although not cast in this form, the procedure of Fortier *et al.* also involves generation of eight possibilities which are separated into two groups (enantiomorphically related about zero). We now seek to relate the procedures.

In the reported procedure (Fortier) weights (A values) and invariant estimates are computed for all eight possibilities and, since they are enantiomorphically related, a weighted average of the appropriate four are used to obtain an estimate of the invariant magnitude $|\psi_E|$. The weighting function [equation (16) of Fortier, Moore & Fraser (1985)] depends on the spread of values, and is such that the resulting weighted average is deemed accurate (large A value) if all four contributors cluster near a single value. In this way the average will always be a good approximation to whichever of the four is actually correct (in magnitude).

Were it not for the term ψ_H, it is obvious that similar results should be obtainable by identifying cases of clustering of the appropriate four estimates above, and again taking a suitable weighted average. It is important to note that the term ψ_H destroys the symmetry about zero, thus the two procedures are not equivalent. One could, however, form the weighted average using only the $|\Delta|$ terms, and then add and subtract it from ψ_H to again form two possible estimates. Note that the two procedures are equivalent if the heavy-atom invariant ψ_H is 0 or π. Since the equations above which include ψ_H are exact, one can conclude that the procedure of Fortier must implicitly assume all heavy-atom invariants are zero (or π). While this is probably not a bad assumption given the small number of heavy-atom sites typically involved, the procedure could be improved by modifying it to include the ψ_H term.

To determine the effectiveness of the modification outlined above, new sets of corrected invariant estimates ψ'_E were generated according to

$$\psi'_E = \pm|\psi_E| + \psi_H$$

for the same set of structures as before, where $|\psi_E|$ is computed using the procedure of Fortier. The results were evaluated in a manner similar to that described earlier; however, the two possible estimates for each invariant are no longer enantiomorphically related. For Bence Jones protein Rhe the resulting distributions of signed deviations are shown in Fig. 4 and for other structures the results are summarized in Table 2. For most structures the bias towards SIR invariants is significantly reduced, and the accuracy is improved as well. For some structures the absolute deviations from true protein invariants were cut in half, while in no cases did the errors increase by more than 0·3°. After applying the correction, the mean absolute error over all structures was 15 ± 2°, and the mean bias towards SIR was 0·8°. In some situations when there was an extremely small number of heavy atoms included in the cell (usually space group $P2_1$ with only one or two sites of substitution) no detectable change occurred. In the orthorhombic space groups apparently even one or two heavy-atom sites are enough to lead to significant deviations from zero

for heavy-atom invariants, and thus to improvements in the estimates.

Since the procedure is easy to apply, never degrades the estimates and often significantly improves them, it probably should be used in all cases regardless of the space group and number of heavy-atom sites.

Space-group and heavy-atom considerations

With the procedure of Fortier, Moore & Fraser (1985), it was claimed that there should be no problem with situations in which the heavy-atom derivatives form a centrosymmetric array. We found that in such cases the distribution of errors in the invariant estimates is quite different, frequently resulting in a sharply bimodal pattern (Fig. 5). One of the peaks is centered near the true protein invariants (but still displaying a residual bias as in the general case); the other is very sharp and is centered essentially on pure SIR invariants. This situation exists in polar space groups when there is only a single heavy-atom site (or two sites with the same y coordinate in $P2_1$). Fortunately

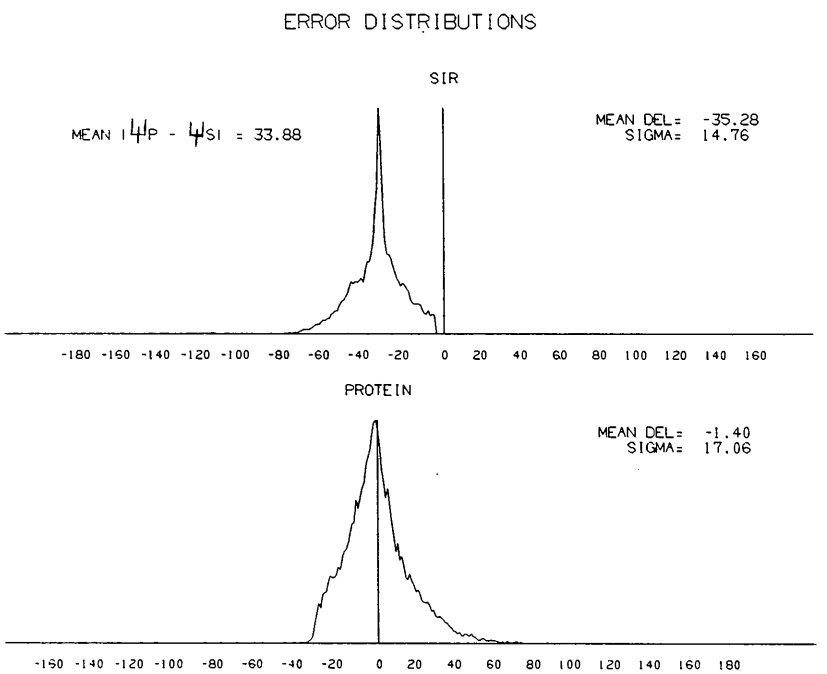

Fig. 4. Distribution of signed errors for Bence Jones protein Rhe (as in Fig. 1), when heavy-atom invariant correction is included. The mean absolute error is only 13·8°, and the bias is actually **away** from SIR by 1·4°.

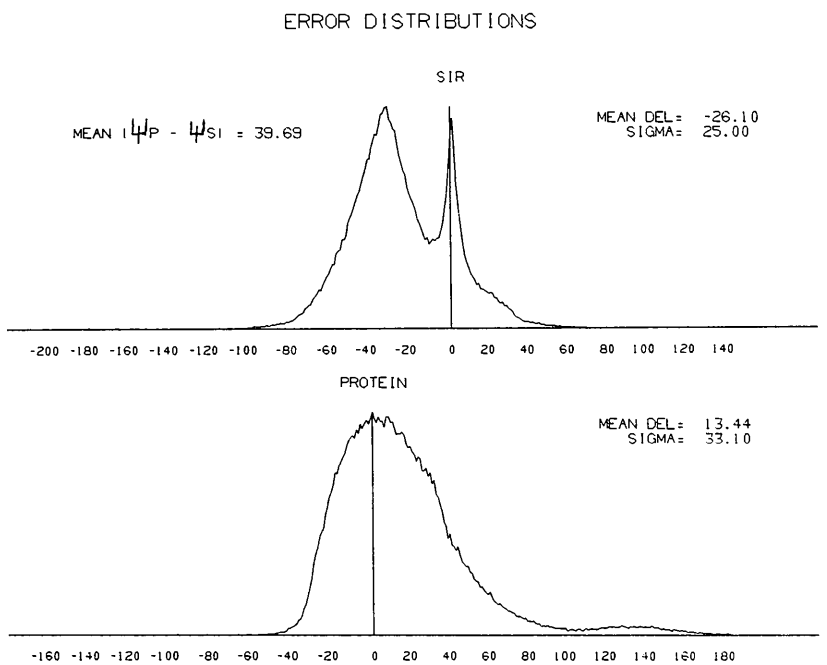

Fig. 5. Distribution of signed errors in invariant estimates for human deoxyhemoglobin, determined from a two-site mercury derivative. The space group is $P2_1$ with the two sites having the same y coordinate. The heavy-atom invariant correction has not been applied. Note the bimodal nature of the distribution when SIR values are used as a reference point. Estimates near SIR values are readily identified, and should probably not be used for phasing.

the peak centered on SIR values usually represents only a small fraction of the total number of invariants (typically <10%), and they can be readily identified. We found that in these cases simply rejecting all estimates near the corresponding SIR invariants (within 5°) results in error distributions similar to those found in the general case. Thus it should still be possible to obtain unbiased protein phases from the remaining invariant estimates. It is useful to do this for all space groups even when the heavy-atom-invariant correction is included, except when the triplet consists of all centric reflections.

Summary

The procedure for estimating three-phase structure invariants from single isomorphous replacement data (Hauptman, 1982) as modified by Fortier, Moore & Fraser (1985) has been extensively tested on over 260 million invariants computed from 17 protein structures and 34 heavy-atom derivatives. It was found that the procedure can provide reasonably accurate values for any protein and derivative combination. Although the modification of Fortier greatly reduces systematic bias towards 'unresolved SIR values', a residual bias still remains. This residual bias can be further reduced, or eliminated, by one of several procedures described in this manuscript. If a correction term is added to account for the heavy-atom invariant, the accuracy is often improved as well. When applied to systems with heavy atoms in a centrosymmetric arrangement, the distribution of errors in the estimates is frequently bimodal, with the major peak centered on the true protein invariants and the minor peak on their SIR counterparts. The estimates near SIR values are readily identified and can be removed, leaving an acceptable distribution of errors in the remaining estimates. The new estimates, being unbiased with randomly distributed errors, should be better suited for use in phase-determining procedures.

Finally, it must be noted that although the procedures outlined can be used to reduce both the absolute errors and bias towards SIR inherent in the estimates, they do not resolve the twofold ambiguity in that there are still two equally probable estimates for each invariant. Work is under way in our laboratory and elsewhere (Hao Quan & Fan Hai-Fu, 1988; Klop, Krabbendam & Kroon, 1987; Langs, 1986; Fan Hai-Fu, Han Fu-son, Qian Jin-zi & Yao Jia-xing, 1984) to resolve this difficulty by various procedures.

The authors are grateful to Dr S. Fortier for helpful discussions. This work was supported by NIH grant GM32918-01A1.

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.

FAN HAI-FU, HAN FU-SON, QIAN JIN-ZI & YAO JIA-XING (1984). *Acta Cryst.* A**40**, 489-495.

FORTIER, S., MOORE, N. J. & FRASER, M. E. (1985). *Acta Cryst.* A**41**, 571-577.

FUREY, W. JR, WANG, B. C., YOO, C. S. & SAX, M. (1983). *J. Mol. Biol.* **167**, 661-692.

GIACOVAZZO, C., CASCARANO, G. & ZHENG, C. D. (1988). *Acta Cryst.* A**44**, 45-51.

HAN, F. & LANGS, D. A. (1988). *Acta Cryst.* A**44**, 563-566.

HAO QUAN & FAN HAI-FU (1988). *Acta Cryst.* A**44**, 379-382.

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

KARLE, J. (1983). *Acta Cryst.* A**39**, 800-805.

KLOP, E. A., KRABBENDAM, H. & KROON, J. (1987). *Acta Cryst.* A**43**, 810-820.

LANGS, D. A. (1986). *Acta Cryst.* A**42**, 362-368.

WANG, B. C., YOO, C. S. & SAX, M. (1979). *J. Mol. Biol.* **129**, 657-674.

XU, Z. B., YANG, D. S. C., FUREY, W. JR, SAX, M., ROSE, J. & WANG, B. C. (1984). Proc. Am. Crystallogr. Assoc. Meet., Lexington, Kentucky, 20-25 May 1984. Abstr. PC2, 50.

Acta Cryst. (1990). A**46**, 567-576

Phase Effects in Three-Beam Grazing-Incidence X-ray Diffraction

BY TZE-PING TSENG AND SHIH-LIN CHANG

Department of Physics, National Tsing Hua University, Hsinchu, Taiwan 30043

(*Received* 31 *October* 1989; *accepted* 28 *February* 1990)

Abstract

The effects of X-ray reflection phases on the surface-reflected intensity, the dispersion surface and the excitation of modes of wave propagation of three-beam grazing-incidence X-ray diffraction are investigated *via* numerical calculations, based on the dynamical theory. Possible ways of determining the triplet phases involved are demonstrated. The *Aufhellung* and *Umweganregung* interactions and the

0108-7673/90/070567-10$03.00