Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 11 August 2000
Accepted 3 November 2000

Improvement of SAS triple invariant estimates for macromolecular direct-methods phasing

David A. Langs,* Robert H. Blessing and Dongyao Guo

Hauptman–Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, NY 14203, USA. Correspondence e-mail: langs@hwi.buffalo.edu

Single-wavelength anomalous dispersion (SAS) data can in principle be phased by direct methods since *a priori* estimates of the three-phase structure invariants can be computed from these data. The mean phase error of the most reliable triple estimates for a small protein, however, is typically no better than 60°, and does not bode well for applications to larger structures. A procedure is described that can substantially lower the error in these estimates and introduce a larger number of useful triple invariants into the phasing process. The mean phase error of the most reliable triples for a 2.5 Å resolution data set from a Pt derivative of a 115-residue protein was reduced from 55 to 25° by this method. It was also possible to identify a significant number of the poorest triple estimates, those with mean phase errors approaching 90°, such that they could be reliably down-weighted or excluded from the phasing process.

1. Introduction

Furey and co-workers (Furey *et al.*, 1985) were the first group to demonstrate that direct methods utilizing single-wavelength anomalous dispersion (SAS) estimates for the three-phase invariants (Hauptman, 1982) were feasible provided one had adequate computing facilities and reasonable criteria to help identify potentially good well phased maps. Several methods have been proposed to further take advantage of these SAS triple estimates (Langs, 1986; Han *et al.*, 1991; Hauptman & Han, 1993; Langs & Han, 1995) to ultimately obtain the native crystal phases. Multisolution tangent-formula procedures are easily modified to use SAS invariant estimates and suitable solutions are often identified in a reasonable number of trials by low values of their SAS phase-refinement residual [Hauptman & Han, 1993, equation (14)]. These procedures are best suited to smaller structures that contain a few strong anomalously scattering atoms. Larger more complex structural applications will require SAS triple estimates that are significantly more accurate than are currently available.

A number of algebraic (Karle & Hauptman, 1957; Vaughan, 1958; Hauptman, 1964; Hauptman *et al.*, 1969; Karle, 1970) and probabilistic (Hauptman, 1975; Giacovazzo, 1977; Karle, 1982) formulae have been proposed to obtain more reliable triple invariant estimates for native data sets based on the six E magnitudes in the first neighborhood or so-called quadrupole relationship (Viterbo & Woolfson, 1973) of the triple. A subsequent application (Langs & Han, 1995) used quadrupole averaging to obtain better SAS triple estimates, but its success depended critically upon the average error in the initial SAS estimates being reasonably small and Gaussian in distribution. In fact, the errors in the estimates were strongly correlated through quadrupoles. An alternative method to improve the SAS estimates will be described that is more independent of this phase-error correlation.

We note that numerous techniques have been devised to resolve the SIR/SAS phase ambiguity given that the substructure is known, in contrast to the methods described above, which seek an *ab initio* direct-methods solution through reliably estimated triple phase invariants, and as such do not require this additional structural information. It is appropriate, however, to acknowledge several recent papers (Fan *et al.*, 1990; Liu *et al.*, 1999) that have successfully used earlier algebraic probabilistic principles (Fan, 1965) for resolving the phase doublet ambiguity based on the value of the triplet phase invariant computed from the known substructure.

2. Analysis

Quadrupole relationships have been a useful framework for the evaluation of three-phase structure invariants. A quadrupole is defined by phases and amplitudes of six pairs of Friedel-related E values which can be partitioned into four triple relationships as

$$\begin{aligned}
\Phi_1 &= \varphi_{\mathbf{h}} - \varphi_{\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}} + t_1 \\
\Phi_2 &= -\varphi_{\mathbf{h}} + \varphi_{\mathbf{l}} + \varphi_{\mathbf{h}-\mathbf{l}} + t_2 \\
\Phi_3 &= \varphi_{\mathbf{k}} - \varphi_{\mathbf{l}} + \varphi_{\mathbf{l}-\mathbf{k}} + t_3 \\
\Phi_4 &= -\varphi_{\mathbf{k}-\mathbf{h}} - \varphi_{\mathbf{h}-\mathbf{l}} - \varphi_{\mathbf{l}-\mathbf{k}} + t_4 \\
\textstyle\sum \Phi_i &= \Phi_1 + \Phi_2 + \Phi_3 + \Phi_4 = t_1 + t_2 + t_3 + t_4.
\end{aligned}$$

The t_i are fractional shifts of 2π which result from transforming the phase values of general reflections back to some standard reference form. In most instances, $\sum \Phi_i$ will exactly equal 0 (mod 2π), inferring that all four values of Φ_i may be reliably close to zero. But on occasion $\sum \Phi_i$ may equal some large fraction of 2π to indicate that it is impossible for all four Φ_i values to be reliably close to zero. It follows that, if a particular triple Φ_i forms a relatively large percentage of these 'aberrant' relationships, it may be indicative that it has some value far from zero. In the case that one has initial SAS estimates, ω_i, for the full basis set of triples Φ_i, where $-\pi \leq \omega_i \leq \pi$ rad, it follows that the estimates will tend to be good when $\sum \omega_i \approx \sum \Phi_i$, regardless of the particular value of $\sum \Phi_i$. Improved ω estimates

$$\exp i\omega_1 = K \sum_j wt_j \exp i\left[\sum \Phi_j - \omega_{2j} - \omega_{3j} - \omega_{4j}\right] \qquad (1)$$

were reported in an earlier paper (Langs & Han, 1995), where wt_j is a weighting factor:

$$wt_j = 0.5\left[1 + \cos\left(\sum_j -\omega_{1j} - \omega_{2j} - \omega_{3j} - \omega_{4j}\right)\right], \qquad (2)$$

which tends towards 1.0 as the quadrupole's closure, ω_{1j}–ω_{2j}–ω_{3j}–ω_{4j}, approaches 0 (mod 2π), and tends toward 0.0 when the closure approaches π (mod 2π). The actual results obtained from (1) suffer from the fact that the mean phase errors in the ω estimates are not randomly distributed in a Gaussian manner but are strongly correlated through the quadrupole itself.

An alternative strategy for evaluating the reliability of the SAS triples is proposed based on the conditional frequency distribution of triples occurring in quadrupoles (Langs, 1993). The v^+ frequency statistic for the triple invariant Φ_1 was formulated with reference to one particular E magnitude, in the group $E_{\mathbf{l}}$, $E_{\mathbf{h}-\mathbf{l}}$ and $E_{\mathbf{k}-\mathbf{l}}$, exceeding some threshold value t, say $|E_{\mathbf{l}}| > 1.75$, on the condition that the other two, $|E_{\mathbf{h}-\mathbf{l}}|$ and $|E_{\mathbf{k}-\mathbf{l}}|$, also exceed this value.

$$v^+ = \left[\sum_{\mathbf{l}} \#\text{qds}(|E_{\mathbf{l}}| > t \big| |E_{\mathbf{h}-\mathbf{l}}|, |E_{\mathbf{k}-\mathbf{l}}| > t)\right] \times \left[\sum_{\mathbf{l}} \#\text{qds}(|E_{\mathbf{l}}| = \text{obs} \big| |E_{\mathbf{h}-\mathbf{l}}|, |E_{\mathbf{k}-\mathbf{l}}| > t)\right]^{-1}. \qquad (3)$$

The denominator of this fraction records the total number of quadrupoles that have any two of the three $|E|$'s exceeding the chosen threshold and the third $|E|$ magnitude known to be recorded within the observed data set.

It is not possible to formulate a similar statistic with regard to analyzing SAS triples based simply on the E magnitudes, since the A values indicating the reliability of the triples are related to their six composite E values [$|E_{\mathbf{h}}|$, $|E_{-\mathbf{h}}|$, $|E_{\mathbf{k}}|$, $|E_{-\mathbf{k}}|$, $|E_{\mathbf{k}-\mathbf{l}}|$, $|E_{-\mathbf{k}+\mathbf{l}}|$] in a more complex manner. Rather than use $|E|$ amplitudes, it may be more appropriate to derive a quadruple-based frequency expression for Φ_1 that is similar to (3) in that it is conditioned on the A values of any one of the other three triples (Φ_2, Φ_3, Φ_4) exceeding some threshold value, say 1.0, on the condition that the other two are known to exceed that threshold.

$$v(\text{SAS})^+ = \left[\sum_{\mathbf{l}} \#\text{qds}(A_4 > t \big| A_2, A_3 > t)\right] \times \left[\sum_{\mathbf{l}} \#\text{qds}(A_4 = \text{obs} \big| A_2, A_3 > t)\right]^{-1}. \qquad (4)$$

To compute this statistic, first generate all the ω estimates for a basis set of E values, which have A_{sas} values exceeding 1.0. Next count the number of quadrupoles that each triple in that list can form, which exclusively involves other triples within the list. This is the numerator of (4). The denominator of (4) counts the number of quadrupoles that can be formed for which three triples are in the list but the fourth is not, simply because its A_{sas} value was less than 1.0. As a practical matter, it may not be necessary to compute the actual denominator, which would require one to scan through a large number of triples not actively used in the phasing process. It may be just as effective to normalize the numerator of (4) using relative numbers based on quadrupoles generated on the basis of $|E|$'s:

$$\text{Scale} = \left\langle \sum_{\mathbf{l}} \#\text{qds}(|E_{\mathbf{l}}| = \text{obs} \big| |E_{\mathbf{h}-\mathbf{l}}|, |E_{\mathbf{k}-\mathbf{l}}| > t) \right\rangle_{\mathbf{h}} \times \left[\sum_{\mathbf{l}} \#\text{qds}(|E_{\mathbf{l}}| = \text{obs} \big| |E_{\mathbf{h}-\mathbf{l}}|, |E_{\mathbf{k}-\mathbf{l}}| > t)\right]^{-1}, \qquad (5)$$

where (4) may be approximated by

$$v(\text{SAS})^+ = \text{Scale} \times \sum_{\mathbf{l}} \#\text{qds}(A_4 > t \big| A_2, A_3 > t).$$

3. Trial calculations

SAS data were provided for the $Pt(NO_2)_4^{2-}$ derivative of macromomycin (Van Roey & Beerman, 1989) which diffracted to 2.5 Å resolution for Cu $K\alpha$ radiation. The protein crystallizes in space group $P2_1$ with a single 115-residue molecule in the asymmetric unit. Data were normalized to E values (Blessing *et al.*, 1996) and locally scaled (Matthews & Czerwinski, 1975) to minimize systematic errors in measurements between Friedel-related reflections. The SAS data set consisted of 3028 Friedel pairs of data for which $k \neq 0$. The largest 1500 E magnitudes were then used to generate 111 670 triples that had A_{sas} values exceeding 0.75 according to Hauptman's (1982) formula. This calculation was repeated, this time using error-free E values in place of the experimental E values.

The list of triples generated from the experimental SAS data was sorted in descending order on A_{sas} and grouped into shells for which the number of triples, average A_{sas}, and mean phase error $\langle|\delta\Phi|\rangle$ between the ω estimates and their 'true' Φ values are reported in section (*a*) of Table 1. The fraction of triples, F_{sin}, for which sin ω and sin Φ_{true} are the same sign, and thus conform to the correct enantiomorph, are in column 4 of the table. Analogous statistics for the triples generated from error-free data are given in section (*b*) of Table 1.

With further analysis of the experimental SAS triples list, the number of quadrupoles (#Qd's) that each triple could form using three other triples within the list was next computed.

Table 1
SAS triple ω value phase-error statistics; comparison between (*a*) experimental and (*b*) error-free *E* values.

The 111 670 triples are ranked in descending order on the magnitude of the *A* value associated with each estimate and then partitioned into groups simulating a normal distribution based on the number of triples in each group. The average *A* value, $\langle A_{\rm sas}\rangle$, mean phase-invariant error, $\langle|\delta\Phi| = |\omega - \Phi_{\rm true}|\rangle$, and fraction of triples, $F_{\rm sin}$, having the same sign of $\sin\omega$ and $\sin\Phi_{\rm true}$ values are indicated. Better than 90% of the top 9000 triples (*) from set (*b*) are consistent with the correct enantiomorph as compared to ~70% for the experimental data set (†).

	(*a*) Experimental *E*'s			(*b*) Error-free *E*'s		
No. of triples	$\langle A_{\rm sas}\rangle$	$\langle\delta\Phi\rangle$	$F_{\rm sin}$	$\langle A_{\rm sas}\rangle$	$\langle\delta\Phi\rangle$	$F_{\rm sin}$
5	16.83	46.7	0.80	17.85	11.3	1.00
15	12.24	52.2	0.80	15.72	23.9	0.99
112	8.74	56.9	0.72	13.42	27.8	1.00
244	6.75	57.4	0.74	11.53	26.5	0.98
726	5.36	55.5	0.73	9.78	28.9	0.97
2032	4.07	57.1	0.70	8.09	32.3	0.94
5977	2.94	58.3	0.69†	6.34	36.9	0.88*
18309	1.98	63.9	0.65	4.54	41.5	0.77
20103	1.42	68.8	0.61	3.30	44.7	0.77
29878	1.09	71.8	0.59	2.39	48.2	0.74
18714	0.89	73.9	0.58	1.70	55.3	0.68
9302	0.81	75.7	0.57	1.29	65.3	0.62
3077	0.78	76.3	0.56	1.06	71.0	0.59
1946	0.76	75.1	0.58	0.94	74.2	0.57
923	0.76	76.2	0.58	0.84	79.2	0.53
307	0.75	72.1	0.56	0.78	84.6	0.51

Table 2
Experimentally measured Pt-MCRM data (*a*) before and (*b*), (*c*) after quadrupole analysis.

After the analysis, the triples were ranked in decreasing order on the number of quadrupole interactions (#Qd). The number of triples in each ⟨#Qd⟩ shell is the same as reported in column 1. The mean phase-invariant error $\langle\delta\Phi\rangle_q$ and $\langle A_{\rm sas}\rangle$ associated with this reordered list are also given. Finally, the number of quadrupoles are renormalized ($\#{\rm Qd}_n$) according to the relative number of quadrupoles which can be formed that have $|E_{\mathbf{l}}|$, $|E_{\mathbf{h}-\mathbf{l}}|$ and $|E_{\mathbf{k}-\mathbf{l}}| \geq 1.65$ regardless of whether the invariants Φ_2, Φ_3 and Φ_4 have $A_{\rm sas} \geq 0.75$ and qualify to be included in the original list of 111 670 triples. Although $\langle|\delta\Phi|\rangle$ has been significantly reduced, the $F_{\rm sin}$ ratio is only marginally improved (†) as compared to column 4 in Table 1.

	(*a*) Before	(*b*) After Quad analysis			(*c*) After renormalization			
No. of triples	$\langle\delta\Phi\rangle$	⟨#Qd⟩	$\langle A_{\rm sas}\rangle$	$\langle\delta\Phi\rangle_q$	$\langle\#{\rm Qd}_n\rangle$	$\langle A_{\rm sas}\rangle$	$\langle\delta\Phi\rangle_n$	$F_{\rm sin}$
5	46.7	603	8.75	24.4	348	12.62	27.6	1.00
15	52.2	480	8.38	39.5	327	9.68	41.6	0.87
112	56.9	383	5.83	52.2	288	6.95	48.1	0.70
244	57.4	320	5.13	54.3	255	6.09	49.4	0.74
726	55.5	270	4.25	55.8	228	4.83	51.2	0.70
2032	57.1	220	3.43	55.9	194	3.63	55.6	0.72
5977	58.3	169	2.58	57.6	158	2.66	58.2	0.69†
18309	63.9	119	1.82	61.8	118	1.82	61.8	0.66
20103	68.8	85	1.39	66.2	88	1.36	66.7	0.63
29878	71.8	61	1.16	70.8	65	1.13	71.7	0.59
18714	73.9	42	1.03	75.8	46	1.01	76.5	0.56
9302	75.7	29	0.97	79.2	33	0.97	80.1	0.54
3077	76.3	22	0.95	82.9	24	0.94	81.9	0.52
1946	75.1	17	0.93	83.0	19	0.93	83.2	0.50
923	76.2	12	0.93	85.6	14	0.92	88.3	0.50
307	72.1	7	0.93	89.3	8	0.90	89.8	0.56

This list was then sorted in decreasing order on the #Qd and ordered in shells having the same number of triples as re-iterated in column 1 of Table 2. The ⟨#Qd⟩, $\langle A_{\rm sas}\rangle$ and $\langle|\delta\Phi|\rangle$ are tabulated in the three columns in section (*b*) of the table. Finally, the scale (5) was computed for each triple using a value $t = 1.65$, and the #Qd for each triple was normalized as $\#{\rm Qd}_n$ to put these numbers on a more comparative scale. This last file was re-sorted in descending order on the value of $\#{\rm Qd}_n$, and the shell averages of $\#{\rm Qd}_n$, $A_{\rm sas}$ and $|\delta\Phi|$ are given in the last three columns in section (*c*) of Table 2.

4. Results and discussion

The shell average mean phase errors $\langle|\delta\Phi|\rangle$ in the original *A*-sorted SAS triple invariant estimates increased from about 45 to 75° as the SAS *A* values decreased from 20 to 0.75 as shown in column 3 of Table 1. The error tends to be smallest for the larger *A* values, but is not consistent with the expected errors, which should range from about 10 to 65° over the listed range of *A* values based on the $I_1(A_{\rm sas})/I_0(A_{\rm sas})$ estimate. The fraction of the triples, $F_{\rm sin}$, consistent with the correct enantiomorph, progressively decreased from 0.80 to 0.56 as shown in column 4 of the table. Error-free SAS data produce $\langle|\delta\Phi|\rangle$ results (*b*) more in line with their $A_{\rm sas}$ values as expected by theory.

When the list of ω estimates is sorted on ⟨#Qd⟩ and arranged in shells having the same number of triples as the original *A*-sorted list, we note the $\langle|\delta\Phi|\rangle_q$ decreases significantly for the first 10 shells of data having *A* values from 20 to ~1.0, and then increases for the last 6 shells which have *A* values from ~1.0 to 0.75. The phase error for the first shell is reduced from 46.7 to 24.4°, while for the last shell it increased from 72.1 to 89.3°. An additional small improvement is noted when the ⟨#Qd⟩ is renormalized with respect to the ability of each triple to form quadrupoles, $\langle\#{\rm Qd}_n\rangle$. As a result of the quadruple analysis (Table 2*b*, *c*), the experimental SAS triple estimates can be improved to the point that they approach the accuracy demonstrated by error-free SAS *E* values (Table 1*b*).

In summary, we note that the quadrupole analysis scheme does help identify which triples in our original list are more reliable estimates than the others. This can only be achieved if we simultaneously isolate a group of triples for which the triple estimates are worse than that observed at the bottom of the original *A*-sorted list. Since the individual ω estimates are not changed by this procedure, the $\langle|\delta\Phi|\rangle$ over the entire set of triples is unchanged and any reduction in this value for some subset of triples must be accompanied by an increase for some other group. Appropriate reassignment of *A* values for the quadrupole ordered list, even to the exclusion of the least reliable triples invariants, should improve the results of phase determination methods which use these estimates.

We thank Dr Patrick Van Roey for the use of various MIR/SAS data sets from his analysis of macromomycin. Support which this research received through NIH grant GM-46733 is gratefully acknowledged.

References

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.
Fan, H.-F. (1965). *Acta Phys. Sin.* **21**, 1114–1118.
Fan, H.-F., Hao, Q., Gu, Y.-X., Qian, J.-Z., Zheng, C.-D. & Ke, H. (1990). *Acta Cryst.* A**46**, 935–939.
Furey, W., Robbins, A. H., Clancy, L. L., Winge, D. R., Wang, B. C. & Stout, C. D. (1985). *Science*, **231**, 704–710.
Giacovazzo, C. (1977). *Acta Cryst.* A**33**, 933–944.
Han, F., DeTitta, G. & Hauptman, H. (1991). *Acta Cryst.* A**47**, 484–490.
Hauptman, H. (1964). *Acta Cryst.* **17**, 1421–1433.
Hauptman, H. (1975). *Acta Cryst.* A**31**, 680–687.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 632–641.
Hauptman, H., Fisher, J., Hancock, H. & Norton, D. (1969). *Acta Cryst.* B**25**, 811–814.
Hauptman, H. A. & Han, F. (1993). *Acta Cryst.* D**49**, 3–8.
Karle, J. (1970). *Acta Cryst.* B**26**, 1614–1617.
Karle, J. (1982). *Acta Cryst.* A**38**, 327–333.
Karle, J. & Hauptman, H. (1957). *Acta Cryst.* **10**, 515–524.
Langs, D. A. (1986). *Acta Cryst.* A**42**, 362–368.
Langs, D. A. (1993). *Acta Cryst.* A**49**, 545–556.
Langs, D. A. & Han, F. (1995). *Acta Cryst.* A**51**, 542–547.
Liu, Y.-D., Harvey, I., Gu, Y.-X., Zheng, C.-D., He, Y.-Z., Fan, H.-F., Hasnain, S. S. & Hao, Q. (1999). *Acta Cryst.* D**55**, 1620–1622.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* A**31**, 480–497.
Van Roey, P. & Beerman, T. A. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 6587–6591.
Vaughan, P. A. (1958). *Acta Cryst.* **11**, 111–115.
Viterbo, D. & Woolfson, M. M. (1973). *Acta Cryst.* A**29**, 205–208.