SHORT COMMUNICATIONS

Acta Cryst. (1999). A**55**, 396–398

Probability distribution of the four-phase structure invariants of isomorphously related structure factors and its applications

Yong-Sheng Liu,* Ning-Hai Hu and Shu-Yun Wang *at Applied Spectroscopy Laboratory, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China. E-mail: peifk@ns.ciac.jl.cn*

(*Received 6 May 1998; accepted 6 July 1998*)

Abstract

The probability distribution of the four-phase structure invariants (4PSIs) involving four pairs of structure factors is derived by integrating the direct methods with isomorphous replacement (IR). A simple expression of the reliability parameter for 16 types of invariant is given in the case of a native protein and a heavy-atom derivative. Test calculations on a protein and its heavy-atom derivative using experimental diffraction data show that the reliability for 4PSI estimates is comparable with that for the three-phase structure invariants (3PSIs), and that a large-modulus invariants method can be used to improve the accuracy.

1. Introduction

Estimates of three-phase structure invariants (3PSIs) based on integrating the techniques of direct methods and isomorphous replacement (IR) have been studied extensively (Hauptman, 1982; Giacovazzo *et al.*, 1988) and some encouraging results have been obtained by their application to macromolecular structures (Hauptman *et al.*, 1982; Giacovazzo *et al.*, 1994). Probabilistic approaches leading to estimates of four-phase structure invariants (4PSIs) have been developed by several authors (Schenk, 1973; Hauptman, 1974; Giacovazzo, 1975). The results have been used to improve starting-set and figures-of-merit procedures (Schenk, 1973; De Titta *et al.*, 1975) and, in combination with 3PSIs, to solve some small macromolecular structures (Sheldrick, 1993). More recently, the probability theory of 4PSIs in the IR case was proposed separately by Kyriakidis *et al.* (1996) and Giacovazzo & Siliqi (1996*a*). Kyriakidis *et al.* (1996) applied a technique that assumes the difference structure factors of two isomorphous structures as random variables, and Giacovazzo & Siliqi (1996*a*) derived the probability distribution of seven pairs of isomorphous structure factors.

It was observed from the results of Giacovazzo & Siliqi (1996*b*) that the 4PSI probability distribution derived from seven pairs of structure factors, four main terms and three cross terms, depends mainly on the four main terms. The cross terms contribute to the distribution in such a way that only those with small $|\Delta|$ values change the sign, with poor reliability, provided by the main terms, while those with large $|\Delta|$ values do not. Thus the formula based on four main pairs of structure factors as the first approximation for the 4PSI distribution seems to be worth studying in greater detail. We present here the probability distribution of four pairs of isomorphously related structure factors, taking the structure factors themselves as random variables, and examine the differences between the 3PSI and 4PSI estimates. Moreover, we suggested recently a method of so-called large-modulus invariants (LMIs) in the phasing process (Hu & Liu, 1997). It makes use of all types of the invariants, *e.g.* eight types for 3PSIs, rather than that consisting only of the native structure factors. This work aims also at providing a theoretical basis for 16 types of 4PSIs and checking how the LMIs work in the 4PSI estimates.

2. The conditional probability distribution of 4PSIs given four pairs of structure-factor magnitudes

Assume that R_i, φ_i and S_i, ψ_i, $i = 1, 2, 3, 4$ are the magnitudes and phases of the corresponding four pairs of structure factors $E_{\mathbf{H}}$, $E_{\mathbf{K}}$, $E_{\mathbf{L}}$, $E_{\mathbf{M}}$ and $G_{\mathbf{H}}$, $G_{\mathbf{K}}$, $G_{\mathbf{L}}$, $G_{\mathbf{M}}$ in the IR case. When a quartet of reciprocal-lattice vectors **H**, **K**, **L**, **M** satisfies $\mathbf{H}+\mathbf{K}+\mathbf{L}+\mathbf{M}=0$, there exist 16 4PSIs of the type:

$$\begin{aligned} \omega_1 &= \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}} + \varphi_{\mathbf{M}}, \\ &\vdots \\ \omega_{16} &= \psi_{\mathbf{H}} + \psi_{\mathbf{K}} + \psi_{\mathbf{L}} + \psi_{\mathbf{M}}. \end{aligned} \tag{1}$$

The conditional probability distribution of (1) can be found from the joint probability distribution $P(R_1, R_2, R_3, R_4, S_1, S_2, S_3, S_4, \varphi_1, \varphi_2, \varphi_3, \varphi_4, \psi_1, \psi_2, \psi_3, \psi_4)$, which is derived *via* its characteristic function, by fixing $R_1, \ldots, S_4$ and integrating with respect to different φ_i or ψ_i variables.† The final results are expressed as

$$\begin{aligned} &P_i(\omega_i|R_1, R_2, R_3, R_4, S_1, S_2, S_3, S_4) \\ &\quad \simeq K_i \exp(A_i \cos\omega_i), \qquad i = 1, 2, \ldots, 16, \end{aligned} \tag{2}$$

where K_i is a normalizing constant. The distribution (2) has a unique maximum at $\omega_i = 0$ or π when $A_i > 0$ or $A_i < 0$, respectively. The reliability parameter

† The full derivation of the joint probability distribution of the eight structure factors $E_{\mathbf{H}}$, $E_{\mathbf{K}}$, $E_{\mathbf{L}}$, $E_{\mathbf{M}}$, $G_{\mathbf{H}}$, $G_{\mathbf{K}}$, $G_{\mathbf{L}}$, $G_{\mathbf{M}}$, where $\mathbf{H}+\mathbf{K}+\mathbf{L}+\mathbf{M}=0$, along with the conditional probability of the 4PSIs given the eight magnitudes $|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, $|E_{\mathbf{L}}|$, $|E_{\mathbf{M}}|$, $|G_{\mathbf{H}}|$, $|G_{\mathbf{K}}|$, $|G_{\mathbf{L}}|$, $|G_{\mathbf{M}}|$ is available from the IUCr electronic archives (Reference: AU0150). Services for accessing this information are described at the back of the journal.

Table 1. *Comparisons between the 4PSI and 3PSI estimates using different reflection sets from the experimental data of cytochrome c_{550} and its $PtCl_4^{2-}$ derivative*

NR is the number of the relationships (4PSIs or 3PSIs) having $|A| > |A_{\min}|$, % is the percentage of the invariants whose cosine signs are correctly estimated, ERR = $\langle|\varphi_t - \omega|\rangle$ (°) with φ_t being the true value and ω the estimated value of the invariant. The reflection subsets are defined as follows: (I) 198 reflections with $|\Delta_R| > 1.1$; (II) 207 reflections with $|E| > 1.4$ and $|\Delta_R| > 0.7$; (III) 210 reflections with $|E| > 2.0$.

	(I)			(II)			(III)		
$\|A_{\min}\|$	NR	%	ERR	NR	%	ERR	NR	%	ERR
4PSI									
0.0	513066	80.8	48.8	411069	64.7	71.1	433082	52.4	86.6
1.0	499399	81.1	48.5	87855	74.9	57.8	3188	82.7	39.1
2.0	227522	85.6	40.9	15722	84.0	43.2	823	89.9	26.1
3.0	83883	89.3	33.3	4318	91.1	28.6	390	93.8	15.3
4.0	33639	91.2	28.3	1617	94.1	20.3	210	98.1	4.6
6.0	6629	93.1	21.7	432	99.1	9.4	68	100.0	0.0
3PSI									
0.0	3471	91.6	30.5	2407	77.2	53.9	2531	54.4	85.0
1.0	3471	91.6	30.5	1481	81.0	47.5	166	69.9	65.4
2.0	2949	92.3	28.6	414	89.6	33.9	20	90.0	22.8
3.0	1569	95.3	22.5	138	89.9	30.3	9	77.8	29.2
4.0	700	96.4	19.1	48	100.0	18.6	–	–	–
6.0	166	98.8	12.6	4	100.0	0.0	–	–	–

$$\begin{aligned} A_i = 2[&\beta_0 R'_1R'_2R'_3R'_4 - \beta_1(S'_1R'_2R'_3R'_4 + R'_1S'_2R'_3R'_4 \\ &+ R'_1R'_2S'_3R'_4 + R'_1R'_2R'_3S'_4) + \beta_2(S'_1S'_2R'_3R'_4 \\ &+ S'_1R'_2S'_3R'_4 + S'_1R'_2R'_3S'_4 + R'_1S'_2S'_3R'_4 \\ &+ R'_1S'_2R'_3S'_4 + R'_1R'_2S'_3S'_4) - \beta_3(S'_1S'_2S'_3R'_4 \\ &+ S'_1S'_2R'_3S'_4 + S'_1R'_2S'_3S'_4 + R'_1S'_2S'_3S'_4) \\ &+ \beta_4 S'_1S'_2S'_3S'_4], \end{aligned} \tag{3}$$

where $R'_j = C_{jR}R_j$, $S'_j = C_{jS}S_j$; if the jth phase of the invariant ω_i is φ, then $C_{jR} = 1$, $C_{jS} = I_1(x)/I_0(x)$; if the jth phase of the invariant ω_i is ψ, then $C_{jR} = I_1(x)/I_0(x)$, $C_{jS} = 1$; $x = 2\gamma R_jS_j$, $j = 1, 2, 3, 4$ and I_0 and I_1 are the modified Bessel functions.† Equations (2) and (3) allow the estimates of 4PSIs from any two isomorphous structures. In the special case that the derivative (D) is obtained by addition of heavy atoms (H) to the native protein (P), (3) can be expressed in a quite simple form:

$$A_i = 2\sigma_{4P}\sigma_{2P}^{-2}R'_1R'_2R'_3R'_4 + 2\sigma_{4H}\sigma_{2H}^{-2}\Delta_1\Delta_2\Delta_3\Delta_4, \tag{4}$$

where $\Delta_j = (C_{jS}|F_{jD}| - C_{jR}|F_{jP}|)/\sigma_{2H}^{1/2}$, $j = 1, 2, 3, 4$, $\sigma_{2P} = \sum_P Z_j^2$, $\sigma_{2H} = \sum_H Z_j^2$, $\sigma_{4P} = \sum_P Z_j^4$, $\sigma_{4H} = \sum_H Z_j^4$, Z_j is the atomic number of the jth atom in the unit cell and the summations over P and over H indicate that the indices j vary over protein atoms (N) and over heavy atoms (N_H), respectively. Because $\sigma_{4P}\sigma_{2P}^{-2} \ll \sigma_{4H}\sigma_{2H}^{-2}$, we obtain a practically useful formula for the most common case involving a native and a heavy-atom derivative:

$$A_i \simeq 2\sigma_{4H}\sigma_{2H}^{-2}\Delta_1\Delta_2\Delta_3\Delta_4, \quad i = 1, 2, \ldots, 16. \tag{5}$$

Defining $\Delta_j = \Delta_{jR}$ when $C_{jR} = 1$ and $\Delta_j = \Delta_{jS}$ when $C_{jS} = 1$, we further have the formulae of A_i for each invariant in (1):

$$\begin{aligned} A_1 &= 2\sigma_{4H}\sigma_{2H}^{-2}\Delta_{1R}\Delta_{2R}\Delta_{3R}\Delta_{4R}, \\ &\vdots \\ A_{16} &= 2\sigma_{4H}\sigma_{2H}^{-2}\Delta_{1S}\Delta_{2S}\Delta_{3S}\Delta_{4S}. \end{aligned} \tag{6}$$

† The parameters γ and β are defined in the supplementary material. See previous footnote.

The 3PSI estimate is in general believed to be more reliable than the 4PSI estimate in the case of non-isomorphous replacement because the reliability parameters are a function of $N^{-1/2}$ and N^{-1} for 3PSIs and 4PSIs, respectively. It is clearly shown from (5) that the 4PSI distribution is mainly related to the contribution from heavy atoms in the derivative and the quality of the estimate can be comparable with that of 3PSIs because the coefficient $\sigma_{4H}\sigma_{2H}^{-2} \simeq N_H^{-1}$ does not differ so much from $N_H^{-1/2}$ in the 3PSI formula when N_H is not large. Equation (5) has a definitive physical significance: since C_{jR} or C_{jS} is the expected value of $\cos(\psi_j - \varphi_j)$ (Fortier *et al.*, 1985), Δ_j is nothing but a projection of the normalized structure-factor vector of heavy-atom structure on the structure-factor vector of protein structure or on the structure-factor vector of derivative structure when the jth phase of the invariant is φ or ψ. Thus we have $|\Delta_j| \leq |E_{jH}|$. When the positions of heavy atoms are known, we can make use of the actual value of Δ_j derived from the triangle composed of $|E_{jP}|$, $|E_{jD}|$ and $|E_{jH}|$ to enhance the effectiveness of (5).

3. Test calculations and discussion

Test calculations were performed on protein cytochrome c_{550}, which crystallizes in space group $P2_12_12_1$ with molecular weight ~ 14 500 Da, and its $PtCl_4^{2-}$ derivative (Timkovich & Dickerson, 1976). In order to assess the estimate quality of (5), three subsets of the data were selected, as defined in Table 1, from 2807 experimental data pairs up to 2.5 Å resolution. For the sake of comparison, the statistical results of native invariants $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}} + \varphi_{\mathbf{M}}$ for 4PSIs and $\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{L}}$ for 3PSIs are given in Table 1 for each subset. We note the following. (*a*) The $|A|$ values for 4PSIs and 3PSIs are substantially comparable but, at the same $|A_{\min}|$ level, the number of the relationships (NR) for 3PSIs is much smaller. (*b*) The overall accuracy is in the order (I) > (II) > (III) for both the 3PSI and the 4PSI cases. The results suggest that, as predicted, the reliability depends on $|\Delta|$ values and is in effect a function of N_H^{-1} rather than of N^{-1}. Therefore, the 4PSI estimate accuracy is greatly improved in the IR case. It is known (Schenk, 1973)

Table 2. *Comparisons between LMIs and the native invariants* ω_1 *for the 4PSI estimated results using 192 reflections with* $|\Delta_R|$ *or* $|\Delta_S| > 1.4$ *from the calculated data of cytochrome* c_{550} *and its* $PtCl_4^{2-}$ *derivative*

See Table 1 for definitions of NR, % and ERR.

	LMI			ω_1		
$\lvert A_{\min}\rvert$	NR	%	ERR	NR	%	ERR
0.0	996792	100.0	8.5	996792	85.9	37.6
2.0	996357	100.0	8.5	590335	98.6	18.3
4.0	268732	100.0	5.9	124878	100.0	10.8
5.0	53561	100.0	3.7	25972	100.0	6.9
6.0	6766	100.0	1.0	3754	100.0	1.4

that a 4PSI satisfying $\mathbf{H} + \mathbf{K} + \mathbf{L} + \mathbf{M} = 0$ can be obtained by the combination of two 3PSIs with one phase in common, the index of which corresponds to the sum of cross terms among **H**, **K** and **L**. In order to check how many 4PSIs are obtainable from 3PSIs in a given data set, 513 066 4PSIs from the subset (I) were examined. The results show that more than two thirds of the 4PSIs are independent of 3PSIs. They may give additional information in practical applications within a given data range.

The use of LMIs has proved to be effective in improving the accuracy of the 3PSI estimates (Hu & Liu, 1997). For 4PSIs, the point of the method is that for each quartet relationship in a set of reflection data with the largest $|\Delta|$ values, only the LMI, which consists of those φ or ψ corresponding to large R or S values, is calculated. This method offers the opportunity for all 16 types of the invariants to be used actively. The reflection set used to check the effect of LMIs in Table 2 comes from 4159 calculated structure-factor pairs of cytochrome c_{550} and its $PtCl_4^{2-}$ derivative. For each quartet relationship, a LMI from (1) was calculated by a corresponding formula in (6). It is evident that LMIs are highly effective in improving the accuracy. The test provides leads for further study on how to make use of all 16 types of 4PSIs, which has not been clear before.

A potential use of 4PSIs would be to estimate some phases with special values when some symmetric relations are incorporated in the probability formula. For such phase-restricted reflections, the LMI method will be easily put into effect to improve the estimate quality. A relevant study is in progress.

This work was supported by the National Natural Science Foundation of China (No. 29573127 and No. 29773045).

References

De Titta, G. T., Edmonds, J. W., Langs, D. A. & Hauptman, H. (1975). *Acta Cryst.* A**31**, 472–479.
Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.
Giacovazzo, C. (1975). *Acta Cryst.* A**31**, 252–259.
Giacovazzo, C., Cascarano, G. & Zheng, C. D. (1988). *Acta Cryst.* A**44**, 45–51.
Giacovazzo, C. & Siliqi, D. (1966*a*). *Acta Cryst.* A**52**, 133–142.
Giacovazzo, C. & Siliqi, D. (1996*b*). *Acta Cryst.* A**52**, 143–151.
Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* A**50**, 609–621.
Hauptman, H. (1974). *Acta Cryst.* A**30**, 472–476.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Hauptman, H., Potter, S. & Weeks, C. M. (1982). *Acta Cryst.* A**38**, 294–300.
Hu, N.-H. & Liu, Y.-S. (1997). *Acta Cryst.* A**53**, 161–167.
Kyriakidis, C., Peschar, R. & Schenk, H. (1996). *Acta Cryst.* A**52**, 77–87.
Schenk, H. (1973). *Acta Cryst.* A**29**, 77–82.
Sheldrick, G. M. (1993). *Crystallographic Computing 6. A Window on Modern Crystallography*, edited by H. D. Flack, L. Parkanyi & K. Simon, pp. 100–110. IUCr/Oxford University Press.
Timkovich, R. & Dickerson, R. E. (1976). *J. Biol. Chem.* **251**, 4033–4046.