Acta Cryst. (1988). A**44**, 482-485

Distribution Fitting Methods Used as Figures of Merit for Non-Centrosymmetric Structures

BY J. HAŠEK

Institute of Macromolecular Chemistry, 162 06 Praha 6, Czechoslovakia

AND H. SCHENK

*Laboratory for Crystallography, University of Amsterdam, Nieuwe Achtergracht 166,
1018 WV Amsterdam, The Netherlands*

(*Received 4 November 1987; accepted 16 February 1988*)

Abstract

Experimental tests have shown that for non-centrosymmetric structures distribution fitting methods can be used successfully as figures of merit for the determination of the most probable set of phases. This applies to both symbolic addition procedures and multisolution methods.

1. Introduction

In a preceding paper (Hašek, Schenk, Kiers & Schagen, 1985) it has been shown that distribution fitting methods (DFM's), theoretically described by Hašek (1984*b*, *c*, *d*), can be successfully applied as figures of merit (FOM's) to centrosymmetric structures when used in the symbolic addition program *SIMPEL*83 (Schenk, 1983) or in the multisolution method *MULTAN*80 (Main, Fiske, Hull, Lessinger, Germain, Declercq & Woolfson, 1980). In this paper, similar results are presented for non-centrosymmetric structures.

2. The use of DFM's for non-centrosymmetric seminvariants

The best procedure for the solution of the phase problem consists in fitting the empirical distributions of seminvariant phase sums as a function of the unknown phases to the sharpest theoretical distributions available at the particular stage of the structure determination (Hašek, 1974). In this way the whole *a priori* structure information contained in the distributions is utilized and as a result this procedure will yield more precise estimates of phases of the reflexions than standard methods. It comprises eventually the most successful route to the solution of the structure.

On the other hand, both the multisolution methods and the symbolic addition procedure assume that the value of a seminvariant is equal to the value at the maximum (mode) of its theoretical distribution. For example, the triplet relation is used in the form

$$\varphi_H+\varphi_K+\varphi_{-H-K}=0, \qquad (1)$$

and quartets as

$$\varphi_H+\varphi_K+\varphi_L+\varphi_{-H-K-L}=0 \text{ or } \pi. \qquad (2)$$

Both methods calculate the phases by solving the over-determined set of linear relations (1) and sometimes (2). When for a solution all the equations are satisfied, the corresponding probability distributions of triplets are δ functions, *i.e.* all seminvariants are 0 or π. In early stages of symbolic addition, before substitution of numerical values to symbolic phases, only a small number of the most reliable relations, without contradictions, are usually used. In general this leads to 'pseudocentrosymmetric' solutions, *i.e.* all trial phases are either 0 or π. After substitution of numerical values to symbols, the tangent refinement may compensate partly for this 'pseudocentric bias'. However, the actual distributions of the invariants attain their true forms only after structure-factor least-squares refinement of the structure.

Because the symbolic addition and multisolution methods tend to yield restricted values (0, π) for the phases, the simplest way to apply DFM's is to accept the bias and to test the distributions just for 0 and π. Thus, as for centrosymmetric seminvariants, the comparison of the distributions of non-centrosymmetric seminvariants is reduced to the test of the fit between probabilities that the seminvariant values lie inside or outside the interval ($-\pi/2$, $\pi/2$). Therefore, the formulae for centrosymmetric and special seminvariants (Hašek, 1984*b*) can be used directly; naturally at the price that information about the distribution profiles is lost.

3. DFM's in the symbolic addition procedure

As in the centrosymmetric case (Hašek *et al.*, 1985), DFM's were tested in *SIMPEL*83 (Schenk, 1983) for a number of non-centrosymmetric structures, one of which, K18JAP (Peschar, 1980) will be used as an example in this paper. All other results are comparable. The K18JAP structure crystallizes in $P2_12_12_1$, with $a=b$, $b=15$ and $c=15$ Å and $N=72$. The testing program *DEM* was implemented in parallel with

0108-7673/88/040482-04$03.00

Table 1. *Numbers of reflexions which appeared in more than* 40, 20, 10, 5 *or* 0 *seminvariants for the structure K*18*JAP* (*Peschar*, 1980)

	$\sum_1$ relation	Non-centrosymmetric triplets	Centrosymmetric triplets	Non-centrosymmetric quartets
More than 40	0	45	0	2
More than 20	0	124	0	6
More than 10	0	201	10	21
More than 5	0	223	47	27
More than 0	25	230	79	58
Number of seminvariants	25	2121	154	133
Number of untested phases	205	0	150	172

the FOM calculation *CRITS* before the tangent refinement routine *TANREF*. It uses the fitting of theoretical and empirical distributions by comparing their function values (Hašek, 1984*b*).* In order to do so the seminvariants were divided into four groups: $\sum_1$ relations, centrosymmetric and non-centrosymmetric triplets and non-centrosymmetric quartets (Table 1). Their distribution characteristics were calculated by means of the well known expressions (*e.g.* Giacovazzo, 1980). All tests of DFM's were carried out using the set of seminvariants selected by *SIMPEL*83 for the determination of the structure. No procedure for optimizing the choice of seminvariants using the theory of graphs (Hašek, Huml, Schagen & Schenk, 1983) was used in order to improve the number of reflections appearing in a particular invariant. For example, for K18JAP (Table 1), 172 phases had no influence on the distribution of quartets because 58 reflections appear in the generated set of quartets only.

The coefficients describing the fit of the 'pseudocentric distributions' are defined by

$$K_k = \sum_i w_i(p_i^{\text{theor}} - p_i^{\text{trial}})^2 \qquad (3)$$

where the summation runs over all regions of seminvariants, p_i^{theor} is the theoretical probability that the seminvariant in the ith region falls into the interval $(-\pi/2, \pi/2)$ and p_i^{trial} is the relative frequency that the seminvariants from the ith region fall into the interval $(-\pi/2, \pi/2)$, and w_i is a weight representing the reliability of p_i^{trial}. The combined coefficient of the fit between 'pseudocentric distributions' is defined by the linear combination

$$K = \sum_k a_k K_k \qquad (4)$$

where K_k are the coefficients of the fit for $\sum_1$ relations, centrosymmetric and non-centrosymmetric triplets and quartets and a_k are coefficients giving them different weights.

Generally, no large differences in the selectivity of 'the best trial set' were found between the 'two-point distribution fitting methods' and the combined figure of merit (CFOM) of *SIMPEL*83. The details of the behaviour of the two-point DFM's are given below for the example of K18JAP.

Table 2. *Efficiency of the SIMPEL*83 *figures of merit for K*18*JAP for solution no.* 94 *which reveals the structure*

	*S*1	HK	*Q*	PQ	NQ	CFOM
FOM for the best solution	99	76	90	97	73	90
Sequence no. of the best solution	1	4	2	1	27	1–2

In a default run of *SIMPEL*83 for the example of K18JAP, the program automatically used three general and two special symbolic phases, giving rise to 128 trial solutions. Two sets (nos. 64 and 94) were selected as the most plausible ones with CFOM = 0·90 in both cases. The values of the individual figures of merit are given in Table 2. Solution no. 64 gave no reasonable structure fragment but the *E* map of set no. 94, based on 198 phases refined by weighted tangent formula, yielded the eight highest peaks at correct atomic positions. Further, out of the 30 highest peaks 18 correct ones identified uniquely all the atoms of the ten-membered ring and its substituents.

A comparison of the relative frequencies of non-centrosymmetric triplets (Table 3*c*) clearly shows 'the pseudocentric character' of the best trial set of phases. The triplets were divided according to their weights into eight regions (characteristics are given in the tenth and eleventh columns of Table 3*a*) and according to their actual phase sums into eight intervals. The pseudocentric distribution of non-centrosymmetric triplets for 'the best trial set' of phases (Table 3*c*) shows that all the phase relations $\varphi_H + \varphi_K + \varphi_{-H-K} = 0$ are either exactly fulfilled or exactly contradictory, *i.e.* $\varphi_H + \varphi_K + \varphi_{-H-K} = \pi$. This partly disappeared after tangent refinement of phases, where the averaging over a large number of phase indications takes place. The bias of the actual distributions disappeared after the structure-factor least-squares refinement (Table 3*a*).

The results of the 'two-point fitting' of the theoretical and empirical distribution of triplets, divided into the eight regions of Table 3, is shown for K18JAP in

* Note that for 'two-point fitting' the distribution moments can be expressed as a simple function of probability of a positive sign, *e.g.* $\langle\cos\varphi\rangle = 2P^+ - 1$ *etc.* The fitting of distribution moments is then equivalent to the fitting of their function values.

Table 3. *Relative frequencies of non-centrosymmetric triplets* (%) *for K18JAP in the following intervals*: 1: $(0, \pi/8)$ *and* $(15\pi/8, 2\pi)$; 2: $(\pi/8, 2\pi/8)$ *and* $(14\pi/8, 15\pi/8)$; . . . ; 8: $(7\pi/8, 9\pi/8)$

Characteristics of the regions of seminvariants *a*, *b*, *c*, . . . , *h* are in the tenth and eleventh columns.

(*a*) The actual distribution based on the final phases after structure-factor least-squares refinement

	Sequence no. of interval								Mean triple product	Number of triplets
Region	1	2	3	4	5	6	7	8	in region	in region
a	70	24	6	0	0	0	0	0	0·753	17
b	60	21	15	3	1	0	0	0	0·640	67
c	48	31	9	8	4	0	0	0	0·541	261
d	46	24	16	8	4	1	1	0	0·479	252
e	37	27	19	8	3	3	2	1	0·439	414
f	37	24	16	10	7	2	2	2	0·405	397
g	26	27	18	11	6	4	3	4	0·369	624
h	21	29	16	12	9	4	4	3	0·348	89

(*b*) The theoretical distribution

	Sequence no. of interval							
Region	1	2	3	4	5	6	7	8
a	51	31	13	4	1	0	0	0
b	43	30	16	7	3	1	1	0
c	37	28	17	9	4	2	1	1
d	33	27	17	10	6	3	2	2
e	31	25	18	11	6	4	3	2
f	29	24	17	11	7	5	3	3
g	28	23	17	12	8	5	4	3
h	26	23	17	12	8	6	4	4

(*c*) The actual distribution based on phases corresponding to 'the best trial set' after symbolic addition procedure for K18JAP

	Sequence no. of interval							
Region	1	2	3	4	5	6	7	8
a	100	0	0	0	0	0	0	0
b	99	0	0	0	0	0	0	1
c	91	0	0	0	0	0	0	9
d	88	0	0	0	0	0	0	12
e	82	0	0	0	0	0	0	18
f	79	0	0	0	0	0	0	21
g	77	0	0	0	0	0	0	23
h	78	0	0	0	0	0	0	22

Table 4. The similarity of all three distributions is striking and the corresponding coefficient of the 'two-point fit' works satisfactorily for triplets (Table 5). The best fit between theoretical and trial distributions was found for set no. 64 (Table 5). The correct phase set was second in the ranking order after the combined *K* FOM.

The comparison of the empirical quartet distributions with the theoretical formulae taken from Hauptman (1975) and Giacovazzo (1976) has not led to satisfactory results. The reasons are summarized in papers by Peschar & Schenk (1986, 1987), in which also more adequate theoretical distributions are described and tested. This will lead to improved results for DFM's.

The coefficients K_k (3) for different seminvariant types and the combined coefficient of the fit *K* (4) are given in Table 5 as a summary for different trial sets of phases. The results are comparable with those described in Table 2.

Table 4. *Relative frequencies of non-centrosymmetric triplets* (%) *in interval* $(-\pi/2, \pi/2)$ *for K18JAP*

Characteristics of regions of seminvariants are in Table 3(*a*).

(*A*) Empirical distribution based on refined phases.
(*B*) Theoretical distribution.
(*C*) Empirical distribution based on the best trial set of phases.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
A	100	99	96	94	91	87	83	80
B	99	95	92	87	85	82	80	78
C	100	99	91	88	82	79	77	78

4. Multisolution methods

A different formulation of the DFM's has been used in the multisolution methods *MULTAN*80 (Main *et al.*, 1980). It is based on a comparison of distribution moments and in particular on the fitting of the widths of the empirical and theoretical distributions. Instead of variances, we tested mean values of three phase cosine invariants [$\langle \cos \varphi \rangle$ calculated only for one side of the symmetrical distribution, *i.e.* on the interval $(0, \pi)$].

Only triplets produced by a standard default run of *MULTAN*80 were used in the DFM program. After grouping the 3000 triplets into 15 regions according to their weights, the most probable set of phases was determined as the lowest value of

$$K' = \sum_k (\langle \cos \varphi_{\rm trial} \rangle_k - \langle \cos \varphi \rangle_k)^2 / {\rm var}\,(\langle \cos \varphi \rangle_k), \quad (5)$$

where the summation runs over all 15 regions, $\langle \cos \varphi_{\rm trial} \rangle_k$ is the mean value of cosine invariant in the *k*th region calculated from the trial phases, the theoretical value $\langle \cos \varphi \rangle_k$ is calculated for the mean weight of triplets $E_3 = 2|E_H E_K E_{-H-K}|/N^{1/2}$ in the *k*th region according to the relation:

$$\langle \cos \varphi \rangle_k = \min \{[(0{\cdot}0106E_3 - 0{\cdot}1304)E_3 + 0{\cdot}5658]E_3, 1\} \quad (6)$$

and the corresponding variance

$${\rm var}\,(\langle \cos \varphi \rangle_k) = \max\,(1 - \langle \cos \varphi \rangle / E_3 - \langle \cos \varphi \rangle^2, 0{\cdot}004).$$

No special attention was paid to symmetrically restricted triplets: they were handled in the same way as the non-centrosymmetric ones. Because all tests were carried out after the tangent refinement, the pseudocentric bias of the trial distributions is not so obvious as in symbolic addition. The coefficient (5) has proved to be successful for all structures routinely solved by *MULTAN*80 in the Institute of one of us (JH). No tests were made using quartets because they are not supported by *MULTAN*80.

Table 5. *Coefficient K of the fit between pseudocentric distributions for different types of seminvariants for K*18*JAP*

Seminvariant	Combined K	Non-centrosymmetric triplets	Centrosymmetric triplets	Non-centrosymmetric quartets	$\sum_1$ relation
Sequence no. of the best set	2	2	11	12	20
Refined phases	6·7	3·2	1·6	8·7	3·7
The best set of phases	19·3	8·4	3·5	7·4	7·6
The lowest K	13·0	2·7	0·8	5·1	2·7
The highest K	193·3	161·1	47·8	20·1	24·7

5. Concluding remarks

When DFM's are used as figures of merit for the determination of the best trial set of phases in early stages of the phase determination, the description of the distribution profiles has to be adapted in order to overcome the problems with the bias of the phases in symbolic addition or multisolution methods. Therefore, in the case of symbolic addition, the distributions were calculated in two points only, as in the centrosymmetric case. Naturally, in this way the distribution profiles are neglected and, as a consequence, the discriminating power of DFM's is not fully utilized.

The significantly lower values of the coefficient of the fit for the refined phases, compared with those for pseudocentric solutions, indicate the possibility of obtaining better results by using the method *ab initio*, *i.e.* when the phases are refined directly by minimization of a criterion based on DFM's as described by Hašek (1985*b*, *c*, *d*) without the preceding step of multisolution or symbolic addition procedures. The main problem of this approach will be to find a sufficiently fast and converging algorithm. Also, theoretical probability distributions have to be used which describe the true distributions of seminvariants more adequately, in particular for quartets and quintets, where the existing formulae (*e.g.* Hauptman, 1975; Giacovazzo, 1976) are not sufficiently exact for the description of the profiles (Peschar 1987; Peschar & Schenk, 1986, 1987).

JH thanks all members of the Laboratory for Crystallography at the University of Amsterdam for their cooperation and Dr K. Huml (IMC) for supporting this research.

References

GIACOVAZZO, C. (1976). *Acta Cryst.* **A32**, 91-99.
GIACOVAZZO, C. (1980). *Direct Methods in Crystallography*, p. 316. New York: Academic Press.
HAŠEK, J. (1974). *Acta Cryst.* **A30**, 576-579.
HAŠEK, J. (1984*a*). *Acta Cryst.* **A40**, 338-340.
HAŠEK, J. (1984*b*). *Acta Cryst.* **A40**, 340-346.
HAŠEK, J. (1984*c*). *Acta Cryst.* **A40**, 346-350.
HAŠEK, J. (1984*d*). *Acta Cryst.* **A40**, 350-352.
HAŠEK, J., HUML, K., SCHAGEN, J. D. & SCHENK, H. (1983). 8th Eur. Crystallogr. Meet., Liege, Belgium. Abstracts, p. 4.04-P.
HAŠEK, J., SCHENK, H., KIERS, C. TH. & SCHAGEN, J. D. (1985). *Acta Cryst.* **A41**, 333-340.
HAUPTMAN, H. (1975). *Acta Cryst.* **A31**, 680-687.
MAIN, P., FISKE, S. J., HULL, S. E., LESSINGER, L., GERMAIN, G., DECLERCQ, J.-P. & WOOLFSON, M. M. (1980). *MULTAN80. A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Univs. of York, England, and Louvain, Belgium.
PESCHAR, R. (1980). Unpublished results.
PESCHAR, R. (1987). Thesis. Univ. of Amsterdam, The Netherlands.
PESCHAR, R. & SCHENK, H. (1986). *Acta Cryst.* **A42**, 309-317.
PESCHAR, R. & SCHENK, H. (1987). *Acta Cryst.* **A43**, 84-92.
SCHENK, H. (1983). *Recl Trav. Chim. Pays-Bas*, **102**, 1-8.

Acta Cryst. (1988). **A44**, 485-495

The Isomorphous Pseudo-Derivative Technique for Phase Refinement by Density Modification

BY CH. ZELWER

Centre de Génétique Moléculaire, CNRS, 91190 *Gif-sur-Yvette, France*

(*Received* 24 *December* 1987; *accepted* 26 *February* 1988)

Abstract

Density modification techniques try to improve the phases of poorly resolved electron density maps given by isomorphous replacement by correcting the systematic errors of the maps according to known physical properties. The phases computed from the corrected maps are combined with the observed moduli through a suitable weighting scheme. A new refinement strategy is proposed which considers the observed moduli and the moduli of the Fourier coefficients of the 'best' map as isomorphous pairs, the Fourier transform of the known systematic errors being a 'heavy-atom contribution'. The lack of closure