Acta Cryst. (1983). A**39**, 916–920

Largest Likely Values for *R* Factors Calculated After Phase Refinement by Non-crystallographic Symmetry Averaging

BY D. C. REES

Department of Chemistry and Biochemistry and Molecular Biology Institute, University of California, Los Angeles, CA 90024, *USA*

(*Received* 22 *April* 1983; *accepted* 18 *July* 1983)

Abstract

The progress of phase refinement by non-crystallographic symmetry averaging is often described by the behavior of a crystallographic *R* factor expressing agreement between the observed structure factors and structure factors calculated from an averaged electron density map. An upper limit for this *R* factor is evaluated for the case of incorrectly positioned non-crystallographic symmetry operators. Depending on the degree of non-crystallographic symmetry, the upper limit on *R* varies from 0·29 to 0·43, for acentric structures. Incorrect structures with the correct non-crystallographic symmetry are anticipated to converge to even lower values of *R*. In all cases, *R* values calculated for incorrect structures will be significantly lower than the value of 0·586 characteristic of wrong structures lacking non-crystallographic symmetry [Wilson (1950). *Acta Cryst.* **3**, 397–399].

Introduction

When multiple copies of identical molecules are located in the asymmetric unit of a crystal, constraints exist between the phases of the associated structure factors (Rossmann & Blow, 1963). These constraints arise from redundancies in the X-ray intensity data generated by the non-crystallographic symmetry relationships between molecules. The real-space formulation of these relationships (Bricogne, 1974) forms the basis of a powerful technique for phase refinement, which has been utilized for macromolecular structure determinations in two principal fashions: (*a*) to refine experimental (multiple isomorphous replacement) phases (Buehner, Ford, Moras, Olsen & Rossmann, 1974; Bloomer, Champness, Bricogne, Staden & Klug, 1978; Harrison, Olson, Schutt, Winkler & Bricogne, 1978; Rees & Lipscomb, 1980; Abad-Zapatero *et al.*, 1981; Wilson, Skehel & Wiley, 1981) and (*b*) to refine trial phases calculated from some initial structural model (Rayment, Baker, Caspar & Muratami, 1982; Robinson & Harrison, 1982).

Structure determinations using either approach are susceptible to systematic errors resulting from incorrect identification of the non-crystallographic symmetry operations. The possibilities for such errors are not negligible, since non-crystallographic symmetry operations are often obtained from rotation and translation functions and from molecular symmetry and packing considerations – techniques which may suffer from significant ambiguities in interpretation. In addition, systematic errors may be introduced into approach (*b*) if the initial structural model is poorly correlated with the true structure. The correctness of the refined structure may be most easily ascertained for structure determinations at high resolution by the ability to follow the path of the macromolecular chain. At low resolution, however, criteria for correctness of a refined structure are much less definitive (Eisenberg, 1982). The crystallographic *R* factor describing the agreement between observed structure factors and structure factors calculated from the final symmetrized electron density map is often presented as an indication of the correctness of the final structure. In this paper, we evaluate an upper limit to this *R* factor for the 'worst' case of a structural model having incorrect non-crystallographic symmetry. It is shown that the largest likely value for *R* in this incorrect case is significantly lower than the value of 0·586 characteristic of random non-centrosymmetric structures lacking non-crystallographic symmetry.

Theoretical aspects

The transformations relating the coordinates of *N* identical copies of a molecule to one particular copy may be expressed by

$$x_n = C_n x_1 + d_n, \quad 1 \le n \le N, \tag{1}$$

where C_n and d_n are the rotation matrix and translation vector describing the *n*th transformation. The following discussion will be restricted to the particular case in which a structure exhibits 'proper' non-crystallographic symmetry, in which the structure is invariant

0108-7673/83/060916-05$01.50

under a group of local rotations about a point. If this point is selected as the origin, $d_n = 0$.

The electron density, ρ, at points x_1 and x_n related by (1) must be equal:

$$\rho(x_n) = \rho(x_1). \tag{2}$$

Averaging an electron density map with the non-crystallographic symmetry generates a new map with density $\rho_c(x_1)$, where

$$\begin{aligned}\rho_c(x_1) &= \frac{1}{N}\sum_n \rho(x_n) \\ &= \frac{1}{N}\sum_n \rho(C_n x_1).\end{aligned} \tag{3}$$

The Fourier transform of this expression yields

$$F_c(h) = \frac{1}{N}\sum_n F_o(hC_n), \tag{4}$$

where $F_o(h)$ and $F_c(h)$ are the observed and calculated structure factors, respectively.

As expressed, this averaging operation is strictly valid only for the continuous molecular transform corresponding to a non-repeating system. In a crystalline sample, this molecular transform is sampled only at grid points of the reciprocal lattice. Since the C_n do not correspond to crystallographic symmetry operators, hC_n will not in general coincide with a reciprocal-lattice point, and so will not be observed. In principle, however, these values could be obtained from the observed structure factors by interpolation (Sayre, 1952).

Since the matrices C_n form a closed point group, we may also write

$$\begin{aligned}F_c(hC_i) &= \frac{1}{N}\sum_n F_o(hC_iC_n), \quad 1 \le i \le N, \\ &= \frac{1}{N}\sum_n F_o(hC_n) \\ &= F_c(h).\end{aligned} \tag{5}$$

This equivalence requires that both the phases and the amplitudes of the calculated structure factors for the N reflections hC_n be identical. If these calculated phases are applied to the $|F_o(h)|$ in an iterative refinement process, we will have, at convergence,

$$|F_c(h)|e^{i\alpha_h} = \frac{1}{N}\left\{\sum_n |F_o(hC_n)|\right\} e^{i\alpha_h}. \tag{6}$$

Consequently, the final calculated amplitude for a structure factor will be equal to the mean value of the amplitudes of the non-crystallographically related reflections. The ability to separate the phase and amplitude components in this summation is due entirely to the non-crystallographic symmetry relationships. This property is an extension of the case where the C_n are crystallographic symmetry operations, so that (6) follows immediately from the required identity of all the $|F_o(h)|$.

If the rotation matrices C_n are incorrect, then the structure factors for the reflections hC_n related by the non-crystallographic symmetry will be uncorreleated. Evaluation of the crystallographic R factor

$$R = \frac{\sum \big||F_o(h)| - |F_c(h)|\big|}{\sum |F_o(h)|} \tag{7}$$

as a function of N is of considerable interest in this case, since R is often quoted as an important indicator of the correctness of a structural model (Rayment, 1983). Misinterpretation of the non-crystallographic symmetry transformations is a serious fundamental error in a structure determination, so that the convergence behavior for this example should provide a 'worst-case' estimate for the largest likely value of R for any given N.

Values for $|F_c(h)|$, when $N = 2$, may be obtained from (6):

$$|F_c(h)| = \tfrac{1}{2}[\,|F_o(h)| + |F_o(hC_2)|\,]. \tag{8}$$

Substitution of (8) into (7) yields

$$R = \frac{1}{2}\left\{\frac{\sum \big||F_o(h)| - |F_o(hC_2)|\big|}{\sum |F_o(h)|}\right\}. \tag{9}$$

Since the model is assumed to be random, $|F_o(h)|$ and $|F_o(hC_2)|$ are independent. The bracketed expression in (9) was shown by Wilson (1950) to equal 0·586 for an acentric structure, so that $R = 0{\cdot}293$ when $N = 2$.

For values of N greater than 2, numerical estimates of R were obtained by the following calculation. N independent data sets, corresponding to the $|F_o(hC_n)|$, $1 \le n \le N$, and obeying acentric Wilson statistics, were obtained using a random number generator. For these calculations, 5000 reflections (different h values) were included in each data set. The results obtained below were essentially unaffected by halving or doubling this number. To obtain $|F_c(h)|$, the average of the N different $|F_o(hC_n)|$ for a given h was calculated. This is equivalent to the averaging operation expressed in (4). R was then evaluated from (7) by taking data set 1 as the $|F_o(h)|$. The resulting dependence of R on N appears in Table 1. The limiting value for R may also be evaluated by noting that as $N \to \infty$ $|F_c(h)| \to \langle|F_o(h)|\rangle$. By standard techniques (Wilson, 1950) and numerical methods (Stroud & Secrest, 1966) to evaluate (7), a value of 0·43 for R is calculated in the limit of $N \to \infty$. Consequently, the value of R at convergence for structures with incorrect non-crystallographic symmetry is much lower than the value $R =$ 0·586 characteristic of a completely incorrect structure without non-crystallographic symmetry.

Table 1. *Dependence of the largest likely value for R on the number N of equivalent molecules related by non-crystallographic symmetry*

R is defined in equation (7) of the text.

N	R
1	0·00
2	0·29
3	0·34
4	0·35
5	0·38
15	0·42
∞	0·43

A second feature of (6) for incorrect solutions is that the probability distribution function for the calculated structure-factor amplitudes will differ from the observed amplitude distribution. The averaging operation in (6) implies that the calculated amplitudes will have both fewer weak and fewer strong reflections than the observed structure factors. The influence of incorrectly averaging a structure on the calculated structure factors is similar to the effect of twinning by merohedry on the intensity distribution (Rees, 1982). Random-walk methods used to obtain the intensity distribution function for the twinning problem are also applicable to the averaging problem. In the present case, however, it does not seem possible to obtain simple analytical solutions for the distribution functions, so that numerical calculations must be employed.

This averaging effect may be demonstrated in a particularly clear fashion by the cumulative function $N(z)$, which gives the probability that a particular normalized intensity is less than z (Howells, Phillips & Rogers, 1950). The intensity z is equal to the square of the structure-factor amplitude and the normalization procedure scales the z such that $\langle z \rangle = 1$. Fig. 1 presents $N(z)$ plots for squared amplitudes calculated from (6) for various values of N, where the unaveraged structure factors obey acentric Wilson statistics. The differences between the averaged and unaveraged distributions are quite pronounced, even for $N = 2$. For comparison, the size of this effect exceeds the difference in $N(z)$ for acentric and centric distributions. Since the $N(z)$ test is routinely used to detect the presence of a center of symmetry in a crystal structure, the test should be sufficiently sensitive to detect errors in non-crystallographic symmetry averaging.

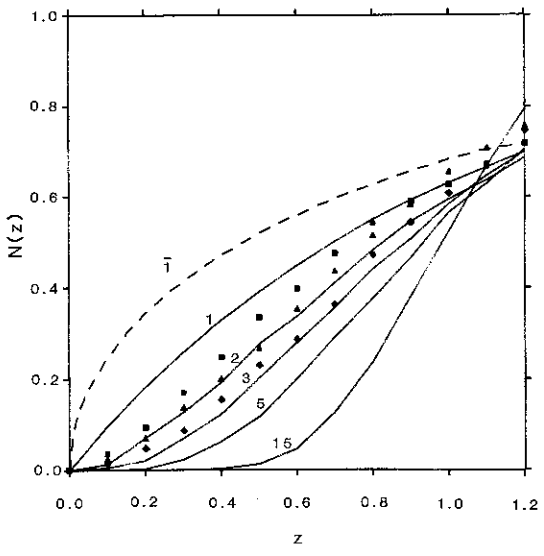

Fig. 1. Cumulative distribution function $N(z)$ for various data sets. Curves designated $\bar{1}$ and 1 are the theoretical distributions for centric and acentric Wilson distributions. Curves 2, 3, 5, and 15 are for acentric distributions averaged by the indicated factor. Data points ■, ▲, ♦ are the $N(z)$ distributions from the model calculations, cases 1, 2 and 3, respectively.

In the preceding discussion, the effects of the molecular envelope and solvent regions on the calculated structure factors have been neglected. This omission should not seriously alter the theoretical expectations, however, since Crowther (1967) has demonstrated that the vanishing of density between subunits provides relatively little phasing information, except with low-resolution data, or unless the subunits occupy only a small fraction of the unit cell. Consequently, the non-crystallographic symmetry relationships which form the foundation of the present analysis should make the dominant contribution to the calculated structure factors.

Model calculations

Model calculations were performed to test the validity of the theoretical results. Random structure-factor amplitudes (obeying acentric Wilson statistics) and phases were generated to 6 Å resolution for a $P1$ lattice with unit-cell dimensions $a = b = c = 40$ Å, $\alpha = \beta = \gamma = 90°$. An overall temperature factor of 30 Å^2 was applied to the data. The non-crystallographic symmetry axis was parallel to the y axis, with fractional coordinates $x = 0{\cdot}5$, $z = 0{\cdot}5$. The molecular envelope had cylindrical symmetry about this axis, with a radius of 0·3 for $0{\cdot}1 \leq y < 0{\cdot}5$ and 0·5 for $0{\cdot}5 \leq y < 1{\cdot}0$, 49% of the unit-cell volume was contained within this envelope.

The system of real-space averaging programs developed by Bricogne (1976) were used to 'refine' the phases of this random structure. Three cases were considered: (1) the structure was assumed to have no non-crystallographic symmetry (only the envelope was used for the refinement); (2) a threefold non-crystallographic symmetry axis was assumed; (3) a fivefold non-crystallographic symmetry axis was assumed. No evenfold axes were used, as this would generate an apparent center of symmetry in y projections. Five cycles of averaging were calculated for each case.

Table 2. *Statistics of phase refinement for the three test calculations described in the text*

R is the crystallographic R factor between the original and final calculated structure-factor amplitudes; $\Delta\varphi$ is the phase difference between the original and final phases.

Case	N	R	$\Delta\varphi$ (°)
1	1	0·199	58·0
2	3	0·370	68·1
3	5	0·408	72·6

Phases generated in one cycle were transferred to the 'observed' structure-factor amplitudes in the Fourier-map calculation for the following cycle; no phase combination methods were employed. The refinement had essentially converged by this stage, since the R factor between structure factors calculated for cycles 4 and 5 was less than 0·04, with phase differences of under 4·4° for each case. Statistics describing the refinement are presented in Table 2. A section of the initial Fourier map and the corresponding sections of maps phased using the 'refined' phases of the three test cases are illustrated in Fig. 2.

In each test calculation, the final R factor is well below 0·586, and close to the values expected from the approximate theory presented above (for the $N = 3$ and 5 cases). In addition, the $N(z)$ distributions for the calculated structure factors deviate appreciably from the original distribution (Fig. 1). These effects are consistent with the theoretical arguments described earlier. Discrepancies between theory and test calculations are most probably due to neglect of the envelope in the theoretical analysis.

Phase refinement by proper non-crystallographic symmetry averaging yields identical phases for structure factors related by the non-crystallographic symmetry operations. This effect exists independently of the

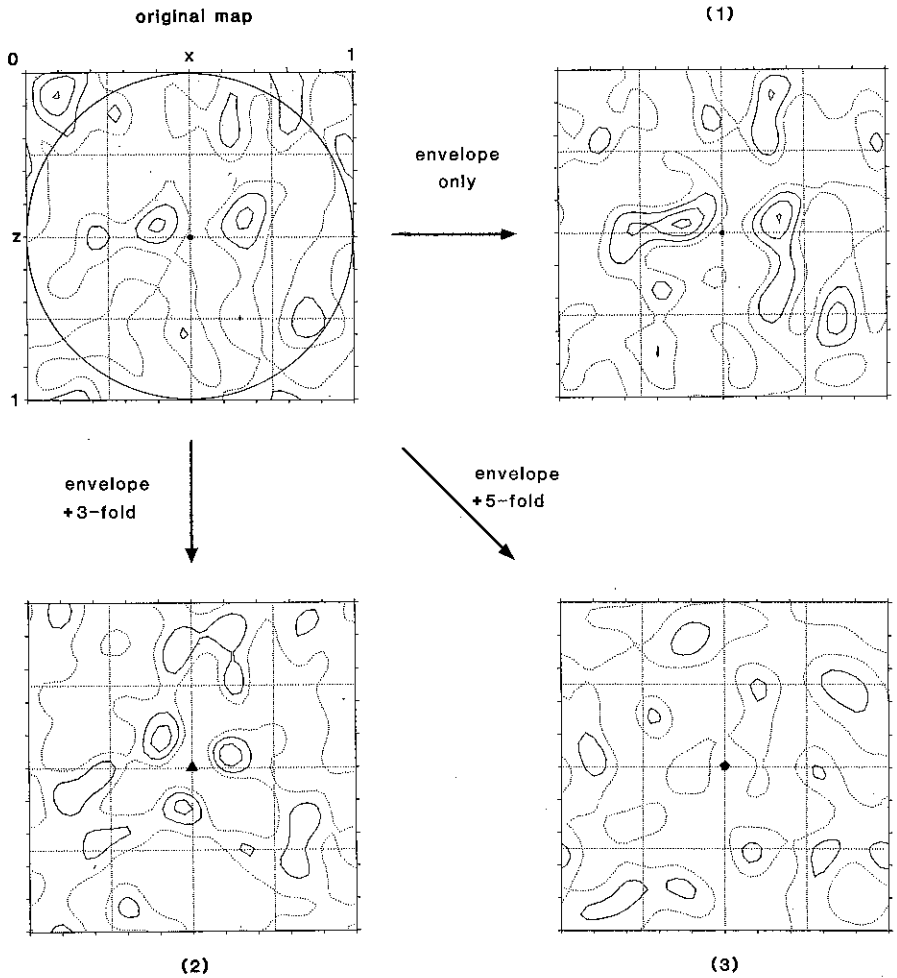

Fig. 2. $y = 35/40$ section of original and 'refined' electron-density maps for test cases 1, 2 and 3. The non-crystallographic symmetry axis is at the center of each section. The envelope boundary is indicated for the original map. Contour levels are at equal and arbitrary levels above zero, and are identical for all maps.

correctness of the non-crystallographic symmetry operators. When these operations are incorrect, this phase identity results in calculated structure factors whose amplitudes are the mean of the amplitudes of observed structure factors related by the non-crystallographic symmetry. As a direct consequence, the R factor between observed and calculated acentric structure factors will be significantly below 0·586, the value which is characteristic of incorrect structures lacking non-crystallographic symmetry. Still lower values for R are anticipated during phase refinement of incorrect structures with the correct non-crystallographic symmetry, since the observed structure-factor amplitudes in this case will automatically satisfy (6) for obtaining the calculated magnitudes, irrespective of the associated phases. Neglect of the envelope in this treatment will modify the quantitative details somewhat, but model calculations presented here suggest this effect will be small for $N \geq 3$. Consequently, low R values during phase refinement by non-crystallographic symmetry averaging do not necessarily imply correctness of resulting structures.

This work was supported by a USPHS Biomedical Research Support Grant to UCLA, and a Dreyfus Foundation Starter Grant in Chemistry.

References

ABAD-ZAPATERO, C., ABDEL-MEGUID, S. S., JOHNSON, J. E., LESLIE, A. G. W., RAYMENT, I., ROSSMANN, M. G., SUCK, D. & TSUKIHARA, T. (1981). *Acta Cryst.* B**37**, 2002–2018.

BLOOMER, A. C., CHAMPNESS, J. N., BRICOGNE, G., STADEN, R. & KLUG, A. (1978). *Nature (London)*, **276**, 362–368.

BRICOGNE, G. (1974). *Acta Cryst.* A**30**, 395–405.

BRICOGNE, G. (1976). *Acta Cryst.* A**32**, 832–847.

BUEHNER, M., FORD, G. C., MORAS, D., OLSEN, K. W. & ROSSMANN, M. G. (1974) *J. Mol. Biol.* **82**, 563–585.

CROWTHER, R. A. (1967). *Acta Cryst.* **22**, 758–764.

EISENBERG, D. S. (1982). *Nature (London)*, **295**, 99–100.

HARRISON, S. C., OLSON, A. J., SCHUTT, C. E., WINKLER, F. K. & BRICOGNE, G. (1978). *Nature (London)*, **276**, 368–373.

HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* **3**, 210–214.

RAYMENT, I. (1983). *Acta Cryst.* A**39**, 102–116.

RAYMENT, I., BAKER, T. S., CASPAR, D. L. D. & MURAKAMI, W. T. (1982). *Nature (London)*, **295**, 110–115.

REES, D. C. (1982). *Acta Cryst.* A**38**, 201–207.

REES, D. C. & LIPSCOMB, W. N. (1980). *Proc. Natl Acad. Sci. USA*, **77**, 4633–4637.

ROBINSON, I. K. & HARRISON, S. C. (1982). *Nature (London)*, **297**, 563–568.

ROSSMANN, M. G. & BLOW, D. M. (1963). *Acta Cryst.* **16**, 39–45.

SAYRE, D. (1952). *Acta Cryst.* **5**, 843.

STROUD, A. H. & SECREST, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs: Prentice-Hall.

WILSON, A. J. C. (1950). *Acta Cryst.* **3**, 397–398.

WILSON, I. A., SKEHEL, J. J. & WILEY, D. C. (1981). *Nature (London)*, **289**, 366–373.

Acta Cryst. (1983). A**39**, 920–924

Moments of the Probability Density Function of R_2 Approached *Via* Conditional Probabilities. IV. Influence of the Elimination of (Low-Intensity) Data on the Applicability of R_2 in Automated Structure Evaluation

BY W. K. L. VAN HAVERE AND A. T. H. LENSTRA

University of Antwerp (UIA), Department of Chemistry, Universiteitsplein 1, *B*-2610 *Wilrijk, Belgium*

(*Received* 7 *September* 1982; *accepted* 17 *July* 1983)

Abstract

The average value of the residual R_2 and its spread $\sigma(R_2)$ is described as a function of a threshold a, below which E_o^2 values are omitted from the data set. Theoretical expressions, valid for finite data sets in the space groups $P1$ and $P\bar{1}$, are derived for $\langle R_2 \rangle$ and $\sigma^2(R_2)$ as functions of a for models containing atoms correctly as well as incorrectly positioned. Use of a threshold causes a decrease in the resolving power of R_2-based strategies used in automated structure evaluations. Random elimination of E_o values gives rise to a larger loss of resolving power than does the elimination of small E_o values.

1. Introduction

Automation in X-ray single-crystal analysis requires criteria discriminating correct from incorrect models of the structure. The residual function R_2, defined as

$$R_2 \equiv \sum_H (E_o^2 - \eta^2 E_c^2)^2 / \sum_H E_o^4 \qquad (1.1)$$

may be used as such a criterion. E_o represents the observed and E_c the calculated magnitude of the

0108-7673/83/060920-05$01.50