Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Detection and treatment of twinning: an improvement and new results

P. Dumas,* E. Ennifar and P. Walter

Institut de Biologie Moléculaire et Cellulaire du CNRS, UPR9002, 15 rue René Descartes, F67084 Strasbourg CEDEX, France

Correspondence e-mail: dumas@ibmc.u-strasbg.fr

Received 22 July 1998
Accepted 31 March 1999

This work deals with two aspects of the twinning problem. Firstly, an improvement of a known statistical test aimed at detecting twinning is presented and, secondly, a new parametrization of twinning is described, as well as a new method to obtain an accurate estimate of the degreee of twinning. During work on crystals of the dimerization-initiation site of the HIV-1 genomic RNA, perfectly twinned crystals were obtained which were not immediately recognized as such by use of a known statistical method. This method, reminiscent of Wilson tests for the detection of centrosymmetric space groups, relies on the calculation of $\langle F^2\rangle/\langle F\rangle^2$ or, equivalently, of $\langle I^2\rangle/\langle I\rangle^2$. It is shown that overlooking experimental errors may lead to erroneously large values of this index and, in turn, to ambiguous or incorrect conclusions. An immediate solution to this problem is presented. Independently, an alternative parametrization which expresses both the effect of twinning on intensities and the operation of untwinning to recover the correct intensities is proposed. A new method for estimating the degree of twinning is also presented. It is based upon maximization of the cross-correlation coefficients between intensities of all available data sets, and yields a fully analytical solution. Tests made with experimental data are quite satisfactory. It is suggested that the latter results could be used efficiently within the MIR method by allowing refinement, through *one* additional parameter only, of the twinning ratios of *all* data sets considered for phasing. Finally, the new parametrization of twinning has striking consequences in this correlation-based twinning determination: very unexpectedly, it yields a novel estimate of the 'twinning ratio' of a potentially twinned crystal which is fully independent of the data set used for the comparison.

1. Introduction

Twinning by hemihedry consists of a growth defect leading to two or more crystal domains of macroscopic size associated in opposite orientations, such that the reciprocal lattices of the two orientations coincide perfectly but wrongly (Friedel, 1926). Because the individual domains are larger than the coherence length of the X-rays, there is no interference between the scattered waves originating from each domain. Instead, the intensities from each domain simply add up. As a consequence, the resulting intensity of each reflection is a weighted mean of the intensities of two non-equivalent reflections,

$$J(\mathbf{h}_1) = (1-\alpha)I(\mathbf{h}_1) + \alpha I(\mathbf{h}_2) \qquad (1a)$$

$$J(\mathbf{h}_2) = \alpha I(\mathbf{h}_1) + (1-\alpha)I(\mathbf{h}_2), \qquad (1b)$$

where $\mathbf{h}_1$ and $\mathbf{h}_2$ are the Miller indices of two reflections related by the twinning, J is the observed intensity, I is the exact intensity that would be measured in the absence of twinning and α is the fraction of twinning, *i.e.* the volume ratio of the two crystals.[1] *A priori*, all situations can be encountered, ranging from a negligible contribution from one of the two orientations ($\alpha \simeq 0$), to equal importance of both of them ($\alpha = 0.5$). In the latter situation, each reflection is exactly the average of two independent reflections, and an additional symmetry is present in the diffraction pattern. In an intermediate situation, *i.e.* for values of α significantly different from 0.5, the previous linear system of equation characterized by the matrix

$$T(\alpha) = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix} \qquad (2)$$

has a non-null determinant

$$D(\alpha) = (1-\alpha)^2 - \alpha^2 = 1 - 2\alpha \qquad (3)$$

and is readily inverted to yield the set of Is from the observed Js. In practice, the problem is not to perform this trivial 'untwinning' operation when possible, but rather to detect the twinning, which may not readily be apparent on visual inspection of the crystals. Furthermore, in the case of perfect twinning, even though 'untwinning' cannot be achieved, there is a strong need for its detection, since otherwise a wrong space group will be used and, potentially, time could be wasted in unsuccessful efforts.

To this end, different statistical tests have been proposed. In the case of partial twinning, α can be determined following one of the methods described or referenced in Yeates (1997). Alternatively, an analytical method based upon maximization of cross-correlation coefficients between all available data sets is proposed in this paper. In the case of perfect twinning, however, such methods are useless, since a real twofold symmetry without twinning could produce exactly the same situation. Another kind of test must then be relied upon, such as that described by Stanley (1972). This test, reminiscent of the classical test for the detection of centrosymmetry (Wilson, 1949), is based upon the observed statistics of $\langle F^2\rangle/\langle F\rangle^2$ or, analogously, of $\langle I^2\rangle/\langle I\rangle^2$ for acentric reflections. Values of $\langle F^2\rangle/\langle F\rangle^2$ of 1.13 and 1.27 or $\langle I^2\rangle/\langle I\rangle^2$ of 1.5 and 2 correspond to twinned and untwinned crystals, respectively.

During work on crystals of the HIV-1 genomic RNA dimerization-initiation site (DIS), crystals with trigonal symmetry were obtained which contain two molecules in the asymmetric unit (Yusupov *et al.*, 1999). Occasionally, hexagonal crystals with identical cell parameters were obtained. Incorrectly, it was first thought that a slight reorganization of the molecules would move the twofold non-crystallographic axis of symmetry in the trigonal space group exactly onto the threefold axis, thus transforming the trigonal crystal into a hexagonal crystal. Consistent with this interpretation was the presence of strong peaks for $\kappa = 180°$ in the self-rotation function of the trigonal crystals, at positions corresponding to the additional twofold axis for the hexagonal space group. Furthermore, when perfect twinning was correctly considered as a possible explanation, the test based on the value of $\langle F^2\rangle/\langle F\rangle^2$ did not clearly show the expected theoretical value for twinning for the hexagonal form: an ambiguous value was found at low and medium resolution and the characteristic value for the absence of twinning was even found at high resolution (Fig. 1). However, comparison with the values obtained for an untwinned trigonal data set showed that there was a tendency to obtain values which were too high (Fig. 1), which strongly suggested that the ambiguous values obtained for the hexagonal crystal could be the result of this tendency. It was then realised that improper accounting for experimental errors explained the observed tendency. In the following, we first examine how to deal with experimental errors in order to improve this test.

2. Improvement of the statistical test for the detection of twinning

Let us call R_X the ratio $\langle X^2\rangle/\langle X\rangle^2$, where X can be F or I. It is seen that R_X is strongly related to the quantity $\mathrm{var}(X) = \langle X^2\rangle - \langle X\rangle^2$, the variance of X. We immediately obtain

$$R_X = 1 + \mathrm{var}(X)/\langle X\rangle^2. \qquad (4)$$

Clearly, $\mathrm{var}(X)$ is the combination of two statistically independent terms. The first of these terms, $\mathrm{var}_\chi(X)$, results from the true 'crystallographic' variance originating from the distribution of the electron density within the unit cell. The

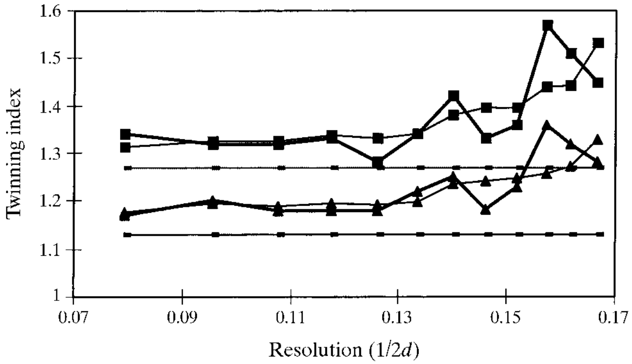

Figure 1
Representation of the twinning index $R_F = \langle F^2\rangle/\langle F\rangle^2$ in shells of resolution for two experimental structures. Thick lines, experimental values of R_F. Thin lines, theoretical values after linear least-squares fit to the experimental values (see text for explanations and values). It should be emphasized that these theoretical values are not the values obtained after substraction of the effect of experimental errors, but rather are theoretical values taking into account the additional term from experimental errors. Therefore, they should be superimposed onto the experimental terms (thick lines) as correctly seen in this figure. Upper curves (squares), experimental and theoretical values of R_F for a 'twinning-free' data set. Lower curves (triangles), experimental and theoretical values of R_F for a perfectly twinned data set showing incorrect hexagonal symmetry instead of trigonal symmetry. The two horizontal lines at ordinates of 1.13 and 1.27 show the theoretical values of R_F for perfectly twinned and untwinned data, respectively.

[1] Anyone unaware of these problems is eagerly encouraged to read the illuminating review by Yeates (1997). For an even more recent account, see Yeates & Fam (1999).

other term, $\mathrm{var}_\varepsilon(X)$, is simply the result of experimental errors and is obviously unrelated to the first term. Therefore, both terms add up to produce the observed variance, $\mathrm{var}(X)$, and (4) should be written as

$$\begin{aligned} R_X &= 1 + [\mathrm{var}_\chi(X) + \mathrm{var}_\varepsilon(X)]/\langle X\rangle^2 \\ &= R_{X\mathrm{ideal}} + \mathrm{var}_\varepsilon(X)/\langle X\rangle^2. \end{aligned} \tag{5}$$

Doubtless, the statistical test supposed to detect a possible twinning deals only with $\mathrm{var}_\chi(X)$ and, therefore, the experimental variance can only interfere with the final result by increasing the observed value relative to the ideal value $R_{X\mathrm{ideal}}$. This is immediately apparent from the graphs of $\langle R_F\rangle$ in resolution shells for the DIS crystals (Fig. 1). The increase of the observed value with resolution is simply the result of $\langle F\rangle^2$ decreasing with resolution. In order to recover the ideal value, $\mathrm{var}_\varepsilon(X)$ must be made more explicit. The latter can be tentatively expressed by either of the following expressions:

$$\mathrm{var}_\varepsilon(X) = \mathrm{var}_0 + \varepsilon\langle X\rangle^2 \quad \text{or} \quad \mathrm{var}_\varepsilon(X) = \mathrm{var}_0 + \varepsilon\langle X^2\rangle, \tag{6}$$

with var_0 and ε being the constants to be determined. These are only *ad hoc* expressions which should not be scrutinized too closely. They can be particularly wrong for the weakest intensities, which often have large σ values. Therefore, excluding (if necessary) the highest resolution range for this study, and using the first expression for $\mathrm{var}_\varepsilon(X)$, together with (5) and the definition $R_X = \langle X^2\rangle/\langle X\rangle^2$, we obtain

$$\begin{aligned} R_X &= R_{X\mathrm{ideal}} + (\mathrm{var}_0 + \varepsilon\langle X\rangle^2)/\langle X\rangle^2 \\ &= R_{X\mathrm{ideal}} + \varepsilon + \mathrm{var}_0/\langle X\rangle^2, \end{aligned} \tag{7}$$

a form which is amenable to linear least-squares determination of $R_{X\mathrm{ideal}}$, with R_X and $\langle X\rangle^{-2}$ being the observed function and variable, respectively. There remains an ambiguity, as ε is not easily determined and the quantity $R_{X\mathrm{ideal}} + \varepsilon$ is obtained, rather than $R_{X\mathrm{ideal}}$. *A priori*, ε is a small term (at most a few percent, otherwise the data are of questionable quality) which should not significantly change the result. Furthermore, the spread of the data (see Fig. 1), which results in a high correlation between the obtained values of $R_{X\mathrm{ideal}} + \varepsilon$ on the one hand and of var_0 on the other, probably renders illusory the correction owing to ε. Thus, we feel that ε can be neglected and that the obtained value for $R_{X\mathrm{ideal}}$ is, therefore, rather slightly overestimated than underestimated. We are well aware, however, that only an investigation of several real cases would allow the drawing of firm conclusions on this.

This treatment has been introduced into the program *LOCHVAT* (Dumas, 1994*a*,*b*) to check systematically native and derivative data sets prior to using the different functions of the program, namely scaling derivative to native data, determining lack of isomorphism and heavy-atom searching. The results were satisfactory as judged by the two cases shown in Fig. 1, illustrating the excellent fit between the observed values for R_F and the calculated ones from (7) taking into consideration the influence of experimental errors. The first of these cases, corresponding to the possibly hexagonal crystal form which gave a quite misleading uncorrected average value $\langle R_F\rangle$ = 1.23, was in fact found to be perfectly twinned, as judged by the values var_0 = 9.7 (e.s.d. = 3) and $R_{F\mathrm{ideal}}$ = 1.124 (e.s.d. = 0.035), the latter being extremely close to 1.13, the theoretical value for twinning. (Fig. 1, lower curves). The corresponding values for the second data set were also quite clear: var_0 = 9.2 (e.s.d. = 2.6) and $R_{F\mathrm{ideal}}$ = 1.272 (e.s.d. = 0.036), the latter value being in perfect agreement with 1.27, the theoretical value for no twinning. It was also found that one data set (not shown in Fig. 1) from a partially twinned ruthenium derivative (α = 25.2%) correctly gave an intermediate value $R_{F\mathrm{ideal}}$ = 1.169 (e.s.d. = 0.031) with var_0 = 2.0 (e.s.d. = 2.9) before untwinning and gave, as expected, the excellent value $R_{F\mathrm{ideal}}$ = 1.28 (e.s.d. = 0.021) with var_0 = 1.9 (e.s.d. = 1.8) after untwinning was performed.

3. A new form to express 'twinning' and 'untwinning' operations

The aim of this section is to propose an alternative parametrization to describe twinning. Although these considerations do not at first seem to have any other relevance than shedding light on some theoretical aspects, they in fact have interesting and unexpected consequences. The 'twinning' matrix $T(\alpha)$, given in (2), of the linear system (1*a*,*b*) can be expressed as the product of a scalar and of a matrix with unit determinant, namely

$$\begin{aligned} T(\alpha) &= \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix} \\ &= D(\alpha)^{1/2}\begin{bmatrix} (1-\alpha)/D(\alpha)^{1/2} & \alpha/D(\alpha)^{1/2} \\ \alpha/D(\alpha)^{1/2} & (1-\alpha)/D(\alpha)^{1/2} \end{bmatrix}, \end{aligned} \tag{8a}$$

with $D(\alpha)$ defined by (3). The matrix $U(\alpha)$ is thus of the form

$$\begin{aligned} U(\alpha) &= \begin{bmatrix} (1-\alpha)/D(\alpha)^{1/2} & \alpha/D(\alpha)^{1/2} \\ \alpha/D(\alpha)^{1/2} & (1-\alpha)/D(\alpha)^{1/2} \end{bmatrix} \\ &= \begin{pmatrix} a & b \\ b & a \end{pmatrix} \end{aligned} \tag{8b}$$

and, therefore, $\det[U(\alpha)] = a^2 - b^2 = 1$, from which it follows that we can make the change of variables defined by

$$\cosh\theta = [\exp(\theta) + \exp(-\theta)]/2 = (1-\alpha)/D(\alpha)^{1/2} \tag{9a}$$

$$\sinh\theta = [\exp(\theta) - \exp(-\theta)]/2 = \alpha/D(\alpha)^{1/2} \tag{9b}$$

by virtue of the identity $\cosh^2\theta - \sinh^2\theta = 1$. For the operation of 'twinning' represented by $T(\alpha)$, this new parametrization leads to

$$\exp(-2\theta) = 1 - 2\alpha = D(\alpha) \tag{10a}$$

and for the operation of 'untwinning' represented by $T^{-1}(\alpha)$ with determinant $1/D(\alpha)$, it leads to

$$\exp(-2\theta) = 1/(1-2\alpha) = 1/D(\alpha). \tag{10b}$$

With this new variable, the matrix with unit determinant in (8*b*) can be transformed into

$$U(\theta) = \begin{pmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{pmatrix} \tag{11}$$

and $T(\alpha)$ is replaced by $\Theta(\theta)$ such that

$$\Theta(\theta) = \exp(-\theta)U(\theta). \tag{12}$$

This form has the following interesting properties. Firstly, it is immediately verified that these matrices have a Lie group structure,[2] as $\Theta(\theta_1)\,\Theta(\theta_2) = \Theta(\theta_1 + \theta_2)$, from which it follows that $\Theta^{-1}(\theta) = \Theta(-\theta)$. Although, unlike α, it has no obvious physical meaning, the new variable θ thus gives sense to the intuitive (and loosely expressed) idea that 'twinning a data set by θ_1' and then again 'by θ_2', is identical to 'twinning it once by $\theta_1 + \theta_2$'. Secondly, contrary to the usual form which makes use of $T(\alpha)$ and $T^{-1}(\alpha)$, both twinning and 'untwinning' are expressed by the same matrix $\Theta(\theta)$, θ being positive for twinning and negative for 'untwinning'. Incidentally, it is apparent, as the determinant of $\Theta(\theta)$ is equal to $\exp(-2\theta)$, that twinning ($\theta > 0$) corresponds to 'shrinking' the data (owing to averaging) and 'untwinning' ($\theta < 0$) corresponds to 'expanding' the data, and the errors as well (Grainger, 1968; Fisher & Sweet, 1980)! The most interesting consequence of such parametrization will be apparent in §4.

4. Cross-determination of the twinning factors for several data sets

When partial twinning corrupts data, their use depends on a correct determination of the twinning factor α (or θ as defined in §3) in order to invert the set of linear equations (1*a*,*b*). There exist statistical methods to obtain an estimate of α by use of the intensities of the twinned data set independently of all other data sets (Britton, 1972; Fisher & Sweet, 1980; Rees, 1980; Yeates, 1988). However, one may want to refine such estimates, as any systematic error will negatively affect the result. In the case of phasing by the MIR method, for example, it is likely that the best method would probably be to consider the twinning factor as a refinable parameter in the program used for heavy-atom parameter refinement. In order to keep the number of free parameters to a minimum, which is always good practice, the following method can be proposed. It relies on the fact that a correctly 'untwinned' data set should be maximally correlated to any reference data set (*i.e.* considered 'twinning-free'),[3] For the sake of generality, the hypothesis that one data set truly is 'twinning-free' can be abandoned, as it is this which is to be verified, and there may therefore be considered to be a twinning factor to be determined for all the data sets. The correlation coefficient between two data sets, $I_i(\xi_i)$ and $I_j(\xi_j)$, is thus a function of the two parameters ξ_i and ξ_j necessary to invert the linear system (1*a*,*b*), with ξ denoting either the usual 'twinning ratio' α, or the newly proposed θ parameter. By definition,

$$C_{ij}(\xi_i, \xi_j) = \frac{\langle[I_i(\xi_i) - \langle I_i(\xi_i)\rangle][I_j(\xi_j) - \langle I_j(\xi_j)\rangle]\rangle}{[\langle I_i^2(\xi_i)\rangle - \langle I_i(\xi_i)\rangle^2]^{1/2}[\langle I_j^2(\xi_j)\rangle - \langle I_i(\xi_j)\rangle^2]^{1/2}}. \tag{13}$$

The use of a correlation coefficient is certainly a good choice, as this is a very robust tool and does not depend on the correct scaling of the compared data sets, contrary to the usual R factor. Since we seek to maximize $C_{ij}(\xi_i, \xi_j)$, which is positive in cases of practical interest, we can equivalently consider its square.

4.1. Using of the usual 'twinning ratio' ($\xi = \alpha$)

Somewhat lengthy, but elementary, algebraic calculations lead to the following expression:

$$C_{ij}^2(\alpha_i, \alpha_j) = \frac{\{A_{ij}[\alpha_i\alpha_j - (\alpha_i + \alpha_j)/2] + B_{ij}\}^2}{[A_{ii}\alpha_i(\alpha_i - 1) + B_{ii}][A_{jj}\alpha_j(\alpha_j - 1) + B_{jj}]} \tag{14}$$

with, explicitly,

$$A_{ij} = 2(R_{ij} - S_iS_j/N_{ij}), \tag{15a}$$

$$B_{ij} = P_{ij} - S_iS_j/(2N_{ij}), \tag{15b}$$

$$R_{ij} = \sum_h (J_{ih}^1 + J_{ih}^2)(J_{jh}^1 + J_{jh}^2), \tag{15c}$$

$$P_{ij} = \sum_h (J_{ih}^1 J_{jh}^1 + J_{ih}^2 J_{jh}^2), \tag{15d}$$

$$S_i = \sum_h (J_{ih}^1 + J_{ih}^2), \quad S_j = \sum_h (J_{jh}^1 + J_{jh}^2), \tag{15e}$$

$$Q_{ij} = 4B_{ij} - A_{ij} = 2\sum_h (J_{ih}^1 - J_{ih}^2)(J_{jh}^1 - J_{jh}^2), \tag{15f}$$

the J_{ih}^1 and J_{ih}^2 terms being the observed hth pair of twinning-related intensities (hence the exponents '1' and '2') of the ith data set and $2N_{ij}$ being the number of common reflections between the ith and jth data sets (*i.e.* N_{ij} terms J_{ih}^1 are compared with N_{ij} terms J_{jh}^1 and likewise for J_{ih}^2 and J_{jh}^2). For the terms A_{ii}, B_{ii}, A_{jj} and B_{jj}, N_{ii} and N_{jj} should both be considered equal to N_{ij}.

Such a simple analytical expression (14), yielding all cross-correlation coefficients C_{ij}, may give the impression that all twinning ratios can be uniquely determined by imposing the condition that all partial derivatives $\partial C_{ij}/\partial\alpha_i$ and $\partial C_{ij}/\partial\alpha_j$ be zero. This is wrong, however, as shown by Fig. 2(*a*), which illustrates a test concerning two synthetic data sets with known twinning ratios: a degenerate elongated peak, rather than a single peak, is obtained. This will be fully explained later. However, if one of the two twinning ratios is considered to be known, say α_i, and solving $\partial \mathrm{C}_{ij}/\partial\alpha_j = 0$ for α_j, we obtain a simple second-degree algebraic equation whose (always real) solutions are

$$\alpha_j = (A_{ij}\alpha_i - 2B_{ij})/[A_{ij}(2\alpha_i - 1)], \tag{16a}$$

$$\alpha_j = [A_{ij}Q_{jj}\alpha_i + 2(A_{jj}B_{ij} - A_{ij}B_{jj})]/A_{jj}Q_{ij}. \tag{16b}$$

Solution (16*a*) is uninteresting as it can be verified that it corresponds to $C_{ij}(\alpha_i, \alpha_j) = 0$ (also an extremum of the square of the correlation coefficient), whilst solution (16*b*) gives the value of α_j which maximizes $C_{ij}(\alpha_i, \alpha_j)$. The value of this

[2] It may be noticed that the infinitesimal generator of the group formed by the $\Theta(\theta)$s is simply $T(\theta)$. This means that $[T(\theta/n)]^n \rightarrow \Theta(\theta)$ for $n \rightarrow \infty$.

[3] Although this very intuitive statement does not seem to need long proof, it is examined in some detail in *Appendix A*.

maximum, as a function of α_i, is explicitly obtained by combining (14) and (16b)[4]

$$C_{ij}^{\max}(\alpha_i) = \big(\{2A_{jj}B_{ij}(A_{ij} - 2B_{ij}) - A_{ij}^2[B_{jj} + (A_{jj} - 4B_{jj})\alpha_i(1-\alpha_i)]\}/ \{A_{jj}(A_{jj} - 4B_{jj})[B_{ii} - A_{ii}\alpha_i(1-\alpha_i)]\}\big)^{1/2}. \tag{17}$$

Interestingly, (16b) gives the equation of the straight line corresponding to the the locus of the maxima of correlation in Fig. 2(a) and (17) confirms that there is no isolated maximum along this line (Fig. 2c), contrary to the clear maximum along the line at a constant value of α_i (Fig. 2b; note that $i = 1$ and $j = 2$ in Fig. 2).

If, therefore, several data sets are considered, the knowledge of one twinning ratio leads in turn to accurate cross-determination of all others. Therefore, considering again the refinement of twinning ratios within the scope of the MIR method, only one should be considered as a free parameter, since all others are uniquely determined from (16b). Such a parameter would doubtlessly be easily refined, as it would have a considerable influence on all intensities of all data sets.

4.2. Use of the new parametrization (ξ = θ)

By simply replacing the twinning ratio α in (14) by its θ equivalent as given in (10b) (since untwinning is required), we obtain

$$C_{ij}^2(\theta_i, \theta_j) = \{[A_{ij}\exp(\theta_i + \theta_j) + Q_{ij}\exp(-\theta_i - \theta_j)]^2\} \div \{[A_{ii}\exp(2\theta_i) + Q_{ii}\exp(-2\theta_i)] \times [A_{jj}\exp(2\theta_j) + Q_{jj}\exp(-2\theta_j)]\}, \tag{18}$$

with the same definitions of the A_{ij} and Q_{ij} terms as in (15a–f). Although this form does not seem to be anything other than the result of a change of variable, it yields very unexpected results. Contrary to the previous situation, the graph of $C_{ij}(\theta_i, \theta_j)$ (Fig. 3a) strongly indicates that a solution for both θ_i and θ_j could be found, as a clear feature (a narrowing of the iso-correlation lines) is visible around the solution. Quite interestingly, the line of maximum correlation given by (16b) (the dashed line in Fig. 2a) is transformed equally into a straight line in Fig. 3(a), as (16b) is transformed into

$$\theta_j = \theta_i + \log[(A_{ij}Q_{jj}/A_{jj}Q_{ij})^{1/2}]. \tag{19}$$

The fact that this line has unit slope leads to optimization of the search for a solution by maximizing the correlation coefficient at a constant value of $\theta_S = \theta_i + \theta_j$. By considering θ_i and θ_S as independent variables, one obtains θ_i as a function of θ_S,

$$\theta_i = \tfrac{1}{2}[\theta_S + \tfrac{1}{4}\log(A_{jj}Q_{ii}/A_{ii}Q_{jj})], \tag{20a}$$

which is equivalent to

$$\theta_i - \theta_j = \tfrac{1}{4}\log(A_{jj}Q_{ii}/A_{ii}Q_{jj}) = \tfrac{1}{4}\log(Q_{ii}/A_{ii}) - \tfrac{1}{4}\log(Q_{jj}/A_{jj}). \tag{20b}$$

At first sight, (19) and (20b) should yield identical results. This is not the case, however, as (19) results from searching for a maximum at a constant value of θ_i and (20b) from searching at constant θ_S. It is seen that the following solutions

$$\theta_i = \tfrac{1}{4}\log(Q_{ii}/A_{ii}) \text{ and } \theta_j = \tfrac{1}{4}\log(Q_{jj}/A_{jj}) \tag{21a}$$

or, expressed with the usual twinning ratio variables,

$$\alpha_i = \tfrac{1}{2}[1 - (Q_{ii}/A_{ii})^{1/2}] \text{ and } \alpha_j = \tfrac{1}{2}[1 - (Q_{jj}/A_{jj})^{1/2}], \tag{21b}$$

are consistent with (20b). Obviously, this consistency does not imply that these are the true solutions and, indeed, they are not. However, many numerical tests have shown that these simple expressions are remarkably close to the true solutions. Furthermore, the solution for θ_i is independent of the jth data set and *vice versa*! This is striking, as these solutions essentially result from comparisons between the two data sets. Expressing (21a) in more detail, we obtain

$$\theta_i = \tfrac{1}{2}\log\{[\langle(J_i^1 - J_i^2)^2\rangle]/[\mathrm{var}(J_i^1 + J_i^2)]\}^{1/2}$$

and

$$\theta_j = \tfrac{1}{2}\log\{[\langle(J_j^1 - J_j^2)^2\rangle]/[\mathrm{var}(J_j^1 + J_j^2)]\}^{1/2}. \tag{22}$$

The interpretation of this result is clear: the denominator $\mathrm{var}(J_i^1 + J_i^2)$, the statistical variance of $(J_i^1 + J_i^2)$, is independent of the twinning level as seen from (1a,b), while the numerator $\langle(J_i^1 - J_i^2)^2\rangle$ depends on it. This result is quite similar to that obtained by Yeates (1988) based upon the statistical behaviour of the index $H_i = |J_i^1 - J_i^2|/(J_i^1 + J_i^2)$, which is strongly related to the quantity appearing in (22). Yeates obtained

$$\alpha_i = \tfrac{1}{2}(1 - 2\langle H_i\rangle), \tag{23}$$

a result identical in its form to (21b). It should be mentioned that the result here obtained is free of any statistical hypothesis about intensities.

These results have been verified with many test data obtained as explained in Fig. 2. In all cases, even with large level of 'experimental' errors affecting the calculated intensities of a twinned crystal, the levels of twinning obtained from (21a,b), as well as from (23), were extremely close to the values used for calculating the intensities (see Fig. 3 for one particular, but significant, case). It should be emphasized, however, that these calculated intensities were 'perfect' as far as the statistical independence between the subsets of terms I^1 and I^2 is concerned. This probably means that these tests are representative of real cases not affected by non-crystallographic symmetry which may lead to correlation between the two subsets of terms (the kind of correlation detected by a self-rotation function!). Effectively, for the experimental cases described in the next section which are known to be affected by non-crystallographic symmetry, the values obtained for α from (21b) were too high by an additional 10–12%, while the estimates using Yeates' method from (23) were more accurate [although slightly too low, as small negative values for α (as low as −4%) were obtained for the supposedly twinning-free data sets]. More examples are

[4] These calculations, as well as those in §4.2, were performed using *Mathematica* 3.0 (Wolfram Research).

required to decide whether some 'average' of (21*b*) and (23) would yield better estimates.

5. Application to an experimental case

The new cross-correlation-based method for the determination of the level of twinning will now be tested on an experimental case. It should be emphasized that the test is based only on the determination of α (or θ) and not on its refinement within a MIR phasing program; the latter would require a great deal of work beyond the scope of this paper. However, the effect of lack of isomorphism will be examined. The experimental case concerns the aforementioned ruthenium derivative of the DIS crystals (Yusupov *et al.*, 1999). This derivative data set was in fact a second ruthenium data set which had been collected with a crystal which had been transferred into low magnesium concentration solution prior to ruthenium soaking. Such change in magnesium concentration (from 100 m*M* in the usual conditions to 2 m*M*) produced a noticeable lack of isomorphism arising from a 1 Å change in the *c* parameter and, for that reason, a new native data set was also collected at low magnesium concentration. This new pair of native and ruthenium derivative data sets, despite a high level of isomorphism, did not show any of the known sites from the previous data sets at high magnesium concentration until the ruthenium derivative was detected to be twinned (see the last example in §2). The new method to determine α (in fact α_2, the native and ruthenium data sets corresponding to $i = 1$ and $j = 2$, respectively) was then applied by setting $\alpha_1 = 0$ in (16*b*) and using all common data between 15–3 Å. The resulting value was $\alpha_2 = 25.2\%$, corresponding to a maximum value of the correlation coefficient of 96.9% in comparison with 92.7% for $\alpha_2 = 0\%$. Yeates' method (Yeates, 1988), making use of the acentric reflections of the ruthenium data set only, gave $\alpha = 23.5\%$ in good agreement with our value. The correctness of this treatment (and of the ruthenium positions) was clearly assessed by the fact that, after untwinning was performed with $\alpha = 25.2\%$, the major ruthenium site appeared clearly as the first peak in the correlation function used by the program *LOCHVAT* (Dumas, 1994*a*,*b*).

It is quite interesting to consider the effect of lack of isomorphism (LOI) by using the native data set at 100 m*M* magnesium instead of the isomorphous native data set at 2 m*M* magnesium. With the same resolution range for the

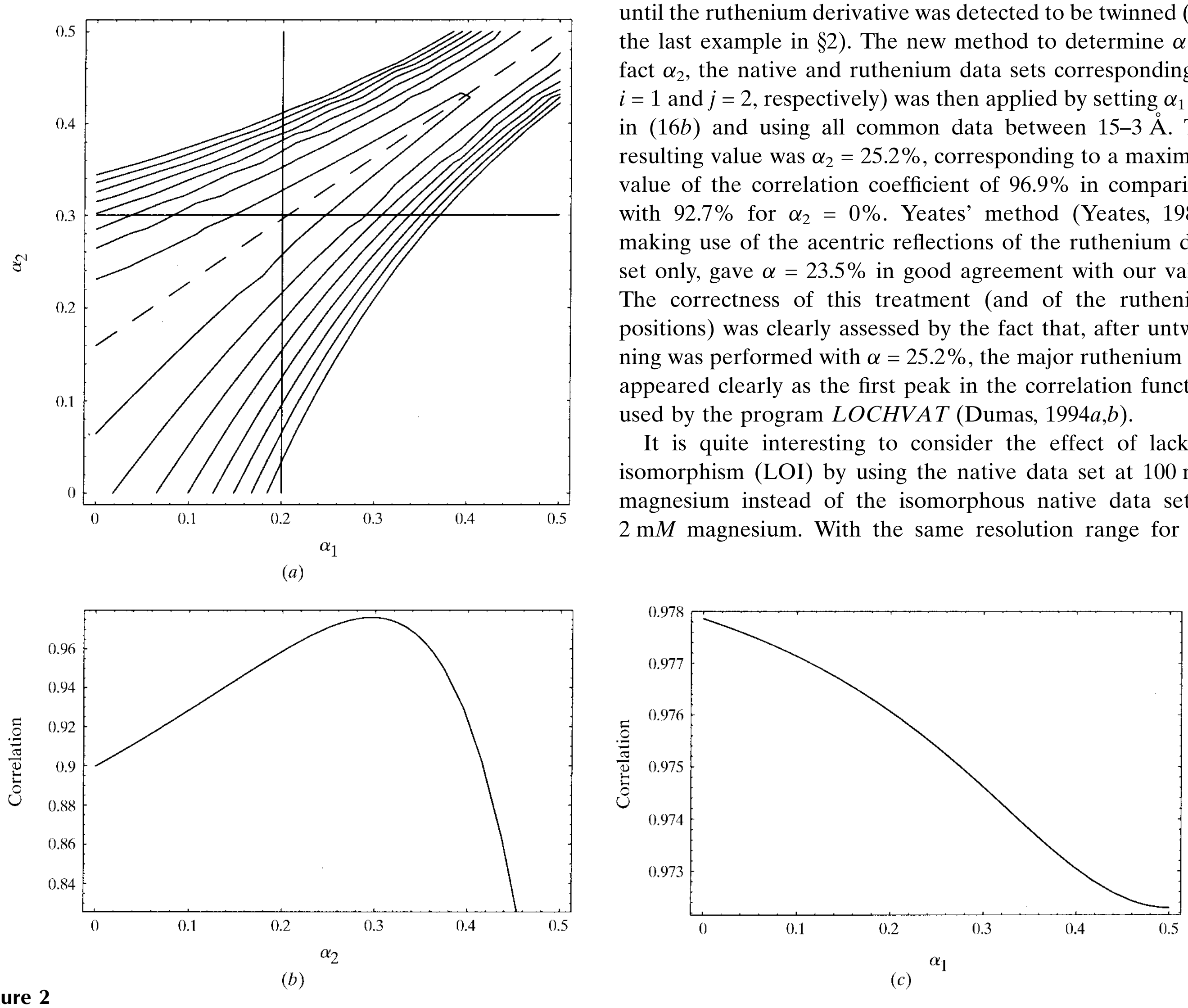

Figure 2
(*a*) Theoretical calculations of $C_{12}(\alpha_1, \alpha_2)$ from (14) for two synthetic data sets. Each data set comprised 1500 reflections. The first data set was calculated as random numbers following Wilson statistics (with parameter Σ taken arbitrarily equal to 1 without loss of generality) and the second was calculated as the sum of the previous distribution and of an independent distribution with $\Sigma = 0.2$ (aimed at representing, for example, the contribution of heavy atoms used for MIR phasing). The two data sets were further modified for twinning with the twinning ratios $\alpha_1 = 0.2$ and $\alpha_2 = 0.3$, marked on the graph as vertical and horizontal lines, respectively. Finally, these two data sets were further modified by the addition of 'experimental' random errors following Gaussian distribution with zero mean and variance equal to 0.05 (*i.e.* 5% of the average intensity). The contours are from 0.9 (minimum contoured value) and are equally spaced by 0.0125. The dashed line corresponds to the theoretical line of maximum of correlation as given by (16*b*). Note that the crossing point of the horizontal and vertical lines lies almost exactly on this dashed line. Calculations and plots were performed with the software *Mathematica* 3.0 (Wolfram Research). (*b*) Correlation profile from (14) along the vertical line $\alpha_1 = 0.2$ of (*a*). The curve shows a clear maximum at $\alpha_1 = 0.296$, very close to the expected value $\alpha_1 = 0.3$. (*c*) Correlation profile, as given by (17), along the ridge of maximum correlation of $C_{12}(\alpha_1, \alpha_2)$ shown in (*a*). Note that the profile is almost flat along this line and that no maximum appears for the correct value $\alpha_1 = 0.2$.

common reflections (15–3 Å), we obtained significantly lower values: α_2 = 21.46% for the twinning ratio and 83.3% for the correlation coefficient. The increased LOI immediately explains the decrease in the correlation coefficient, but also explains the decrease in the twinning ratio. This can be rationalized by reasoning on a hypothetical continuous increase of LOI: if it increased without limits, the correlation coefficient between the two 'twinning-free' data sets would tend to vanish and thus α_2 would become 0. Thus, it can be understood that a lower value is found for a moderate increase in LOI.[5] Interestingly, the values obtained for the twinning ratio α_2 with various resolution ranges for the common reflections (from 15–5.5 Å to 15–3 Å) are extremely stable for both cases. We obtained $\langle\alpha_2\rangle$ = 0.252 with an e.s.d. of 0.0014 and $\langle\alpha_2\rangle$ = 0.214 with an e.s.d. of 0.0035, for the more isomorphous and less isomorphous pairs of data sets, respectively.

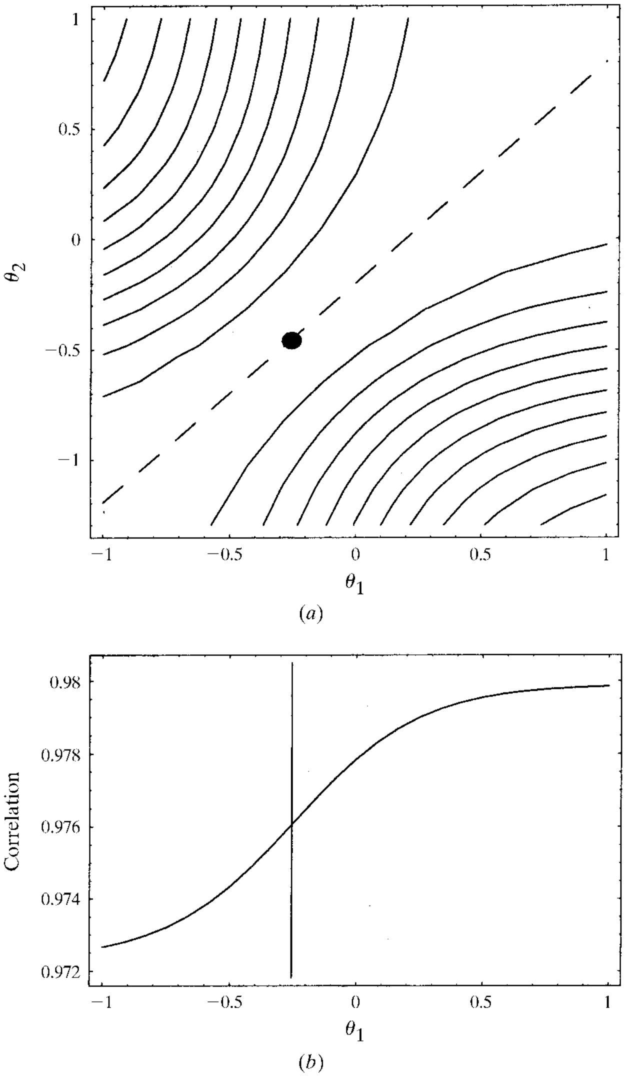

Figure 3
(*a*) Same case as for Fig. 2, but using the parameter θ instead of the usual twinning ratio α. A distinct narrowing of iso-correlation lines around the correct values θ_1 = −0.2554 and θ_2 = −0.4581 (marked with a spot), corresponding to α_1 = 0.200 and α_2 = 0.300, respectively, can be observed. The average values and their e.s.d.s, obtained from (21*a*,*b*) and (23) with 20 independent calculated data sets, are θ_1 = −0.254 (0.013) and θ_2 = −0.454 (0.013), α_1 = 0.199 (0.008) and α_2 = 0.298 (0.005) for (20*a*,*b*), and α_1 = 0.200 (0.004) and α_2 = 0.326 (0.003) for (23). All values but one are in perfect agreement with the exact values (as judged from the e.s.d.). (*b*) Correlation profile along the dashed line in (*a*). Contrary to the usual parametrization with α (Fig. 2*c*), in this case there is a distinct feature along this line, namely an inflection point for the correct value of θ_1 (corresponding to the vertical line).

6. Conclusions

This work has dealt with two distinct aspects of the twinning problem. The first aspect concerns an improvement of a statistical test aimed at detecting twinning *a priori*, that is to say without even knowing which pairs of reflections are potentially related by twinning. The improvement is based on explicit consideration of experimental errors. The second aspect concerns determination of the importance of twinning by maximization of the correlation coefficient between intensities of two data sets. Probably the most interesting aspect of the proposed method is that it should allow the accurate determination of the level of twinning of *all* data sets used for phasing with the MIR method, by considering *only one* of them as an additional refinable parameter. For the moment, however, this is only hypothetical, as this suggestion awaits practical implementation. Finally, a new parametrization of twinning by hemihedry is proposed as an alternative to the usual parametrization representing the volume ratio α of the two twinned crystals. This new parameter θ has the interesting property of allowing representation of both twinning and 'untwinning' operations by the same matrix $\Theta(\theta)$, θ being positive for twinning and negative for 'untwinning'. It also yields an unexpected result in the cross-correlation-based method for the determination of twinning level. At variance with what happens when using α, the latter method yields, for each data set, an estimate of the θ value which is independent of all other data sets. Whether this new estimate of the twinning level has real practical application in addition to its theoretical interest is not yet clear.

APPENDIX *A*

We will consider here in some detail the intuitive statement following which one can 'untwin' a data set by maximizing its correlation with a reference data set. Our goal is not to prove this fairly obvious statement, but rather to test the effect of different levels of difference between the two data sets being compared. In particular, we consider the effect of large differences between two data sets (not merely experimental

[5] In practical cases, for which one would be obliged to use a non-isomorphous pair of data sets, one may be concerned by this discrepancy between values obtained at different levels of LOI. It is quite possible, however, that this too low value for the twinning ratio is an optimum for the purpose of MIR phasing, since the method in use leads to the best correlation between native and derivative data sets. More work is necessary to test this optimistic statement.

errors) and make use of the θ variable. From (17), $C_{12}^2(\theta)$, the square of the correlation coefficient between the reference data set I_1 and the twinned data set J_2 (*i.e.* I_2 'twinned by $\theta_2 = \theta$') is given by

$$C_{12}^2(\theta) = \frac{1}{\mathrm{var}(I_1)} \frac{[A\exp(\theta) + B\exp(-\theta)]^2}{C\exp(2\theta) + D\exp(-2\theta)}, \tag{24}$$

with $\mathrm{var}(I_1) = \langle I_1^2 \rangle - \langle I_1 \rangle^2$, the variance of the data set I_1 and the terms A, B, C and D being defined by comparison with (18). Examination of the graph of $C_{12}(\theta)$ for test cases (Fig. 4), even those with quite an important level of difference between I_1 and I_2, clearly shows that the correlation decreases in a Gaussian-like fashion from its maximum value $C_{12}^{\max}$ at θ very close to 0, to an asymptotic value close to $0.5C_{12}^{\max}$ [as can be seen from a careful examination of (24) with $|\theta| \to \infty$ and the definitions of A and C]. The differences between the two data sets (whatever their origin) easily explain a small displacement of the maximum of correlation relative to its exact position $\theta = 0$. It is therefore verified, both theoretically and numerically, that the maximum is 1 and lies exactly at $\theta = 0$ for two identical data sets, *i.e.* $I_1 \equiv I_2$ (Fig. 4, upper curve). The fact that the graphs of $C_{12}(\theta)$ are almost symmetric is a consequence of A and B on one hand, and C and D on the other hand, being almost equal (see legend of Fig. 4 for numerical values).

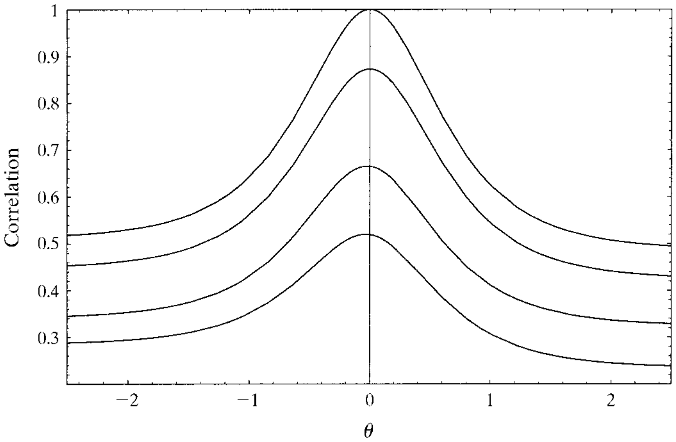

Figure 4
Graphs of $C_{12}(\theta)$ from (24) illustrating that twinning can only decrease the correlation coefficient between two data sets. The uppermost curve shows the influence of twinning on a given data set considered a reference data set. The three other curves, from top to bottom, correspond to data having increasing levels of difference with the reference data set. Each data set was calculated as the sum of two independent contributions following Wilson statistics, with parameters Σ and $\xi\Sigma$, respectively, ξ taking the values 0.0, 0.40, 0.75 and 1.0 from top to bottom ($\Sigma = 1$, without loss of generality). The corresponding values of A, B, C and D, are from top to bottom: A = 1.047, 1.031, 1.082, 1.014; B = 1.097, 1.088, 1.092, 1.147; C = 2.094, 2.333, 3.371, 4.079; D = 2.195, 2.463, 3.266, 4.309. It is obvious that, even for cases of large differences between the compared data sets (much larger than those occurring in the case of heavy-atom substitution, for which ξ rarely attains 0.2), the correlation maximum lies almost exactly at $\theta = 0$. Note that $|\theta|$ takes the maximum value of 2.5 on this graph, which corresponds to a large value $\alpha = 45.9\%$ of the 'twinning ratio'.

APPENDIX *B*

After this paper had been submitted, our attention was drawn to another correlation-based method for the determination of the level of twinning of a crystal (Taylor & Leslie, 1998). In contrast to the method presented here, this other method does not try to maximize a cross-correlation coefficient between different data sets but instead seeks to minimize the correlation coefficient between the two halves of 'untwinned' intensities, I_h^1 and I_h^2, related by the twinning operation. The basis of the method is very clear, as twinning evidently tends to increase this correlation (up to almost 1 in case of perfect twinning, as the sole source of differences remaining between the two halves is then the experimental errors). It is straightforward to use the method described in this paper in this new frame by replacing equation (13) by

$$C(\xi) = \frac{\big\langle [I^1(\xi) - \langle I^1(\xi)\rangle][I^2(\xi) - \langle I^2(\xi)\rangle] \big\rangle}{\{\langle [I^1(\xi)]^2 \rangle - \langle I^1(\xi)\rangle^2]^{1/2} \{\langle [I^2(\xi)]^2 \rangle - \langle I^2(\xi)\rangle^2\}^{1/2}}. \tag{25}$$

Again, after all calculations and simplifications coming from elementary hyperbolic trigonometry have been performed, we obtain using $\xi = \theta$

$$C(\theta) = \frac{V\sinh 2\theta + C_{12}\cosh 2\theta}{[1 - V^2 + (C_{12}\sinh 2\theta + V\cosh 2\theta)^2]^{1/2}}, \tag{26}$$

with C_{12} being equal to the correlation coefficient between the two subsets of experimental intensities J^1 and J^2, that is to say $C_{12} = C(0)$ as defined by (25), and V being defined by

$$V = \tfrac{1}{2}\{[\mathrm{var}(J^1)/\mathrm{var}(J^2)]^{1/2} + [\mathrm{var}(J^2)/\mathrm{var}(J^1)]^{1/2}\}. \tag{27}$$

One immediate simplification can be made as, to a second-order approximation, V is very close to 1, since there are no reasons for the two statistical variances $\mathrm{var}(J^1)$ and $\mathrm{var}(J^2)$ to differ significantly, and thus (27) leads to

$$C(\theta) \simeq (\sinh 2\theta + C_{12}\cosh 2\theta)/(C_{12}\sinh 2\theta + \cosh 2\theta), \tag{28a}$$

which can be transformed into the remarkably simple expression

$$C(\theta) \simeq \tanh 2(\theta + \kappa) \tag{27b}$$

with $\tanh(2\kappa) = C_{12}$. In order to make the correlation coefficient vanish, one obtains for θ

$$\theta = -\kappa.$$

This solution is very aesthetic, as it immediately makes apparent the direct link between the required level of 'untwinning' and the correlation between the two subsets of intensities J^1 and J^2 resulting from twinning. However, for that very reason, it should be noted that this solution can cause problems, because any non-crystallographic symmetry leading to correlation between the two subsets of intensities I^1 and I^2 would be wrongly interpreted as being a consequence of twinning.

References

Britton, D. (1972). *Acta Cryst.* A**28**, 296–297.
Dumas, P. (1994*a*). *Acta Cryst.* A**50**, 526–537.

Dumas, P. (1994*b*). *Acta Cryst.* A**50**, 537–546.
Fisher, R. G. & Sweet, R. M. (1980). *Acta Cryst.* A**36**, 755–760.
Friedel, G. (1926). *Leçons de Cristallographie*. Paris: Berger–Levrault.
Grainger, C. T. (1968). *Acta Cryst.* A**25**, 427–434.
Rees, D. C. (1980). *Acta Cryst.* A**36**, 578–581.
Stanley, E. (1972). *J. Appl. Cryst.* **5**, 191–194.
Taylor, H. O. & Leslie, A. G. W. (1998). *CCP4 Newslett.* **35**, 9.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Yeates, T. O. (1988). *Acta Cryst.* A**44**, 142–144.
Yeates, T. O. (1997). *Methods Enzymol.* **276**, 344–358.
Yeates, T. O. & Fam, B. C. (1999). *Structure*, **7**, R25–R29.
Yusupov, M., Walter, P., Marquet, R., Ehresmann, C., Ehresmann, B. & Dumas, P. (1999). *Acta Cryst.* D**55**, 281–284.