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

normalized structure factors belonging to structure and model, respectively. The fraction of the scattering power of the n-atom model *versus* the N-atom structure is given by

$$\eta^2 \equiv \eta_c^2/\eta_o^2. \tag{1.2}$$

For point atoms with equal scattering power, $\eta_c^2 = n$ and $\eta_o^2 = N$.

The usefulness of an R_2-based criterion in a statistical decision procedure can be studied if at least the first two moments of the probability density function $P(R_2)$ are available. In previous papers of this series (Van Havere & Lenstra 1983*a*,*b*,*c*), we developed formulas valid for finite data sets in the space groups $P1$ and $P\bar{1}$, from which these moments can be calculated for various types of models, ranging from completely correct to totally incorrect including all in-between situations. One of the conclusions was that the iterative automated procedure which has the largest chance of success adds one new atom per cycle to a known partial model of the asymmetric unit. Introduction of incorrect atoms has a serious detrimental effect. Thus to prevent the acceptation of an incorrect atom the R_2 check has to be performed many times, which accumulates to a large fraction of the total computing time needed. An acceleration of the reliability check may be sought in a reduction of the total number of reflections involved in the computations. For example, one may eliminate all reflections for which $E_o^2 \leq a$, or alternatively eliminate a similar number of randomly chosen reflections. In this paper we explore the consequences of these two alternatives with respect to R_2-based reliability checks.

The use of conditional probabilities allowed us to handle finite data sets and to arrive at realistic, theoretical estimates of $\sigma(R_2)$. In this aspect the analysis presented here differs from previous ones (Petit, Lenstra & Van Loock, 1981; Petit & Lenstra, 1982). Now, the influence of the threshold a on $\langle R_2 \rangle$ as well as on $\sigma(R_2)$ is taken explicitly into account. This strengthens and enlarges considerably the conclusions drawn by Petit *et al.*

2. General expressions for the moments of $P(R_2)$

It is necessary to recall some of the nomenclature and results developed earlier in this series. A tentative n-atom model in which n_g atoms are correctly positioned and n_f ($n_g + n_f = n$) atoms are badly misplaced is denoted by $\{g,f\}$. The quantities to be investigated are $\langle R_2;\mathscr{E}_o \rangle$ and $\sigma^2(R_2;\mathscr{E}_o)$, that is the value of R_2 averaged over all models with fixed n_g and n_f under the constraint of the set of observed E values ($\mathscr{E}_o$) and the spread of $P(R_2)$ under the same conditions.

Expressions for these quantities are to be found in Van Havere & Lenstra (1983*c*) as equations (3.1) and (4.1) for $\langle R_2;\mathscr{E}_o \rangle$ in the space groups $P1$ and $P\bar{1}$, respectively. Similarly, equations (3.2) and (4.2) give $\sigma^2(R_2;\mathscr{E}_o)$ in $P1$ and $P\bar{1}$, respectively. In a particular structure average and spread can thus easily be enumerated for all models $\{g,f\}$. The impact of a threshold a simply follows from the elimination of all $E_o^2 \leq a$ from the summations.

If, however, we want to generalize the picture, that is without reference to a particular actual structure, we must construct an average structure. This is done by first replacing in equations (3.1), (3.2), (4.1), (4.2) of Van Havere & Lenstra (1983*c*) each term $\sum_H E_o^n$ by $\mathscr{H}\langle E_o^n \rangle$. $\mathscr{H}$ represents the number of reflections in the data set before the threshold is applied. Next, the threshold is introduced by replacing $\sum_H E_o^n$; $E_o^2 \geq a$ by $\mathscr{H}_a\langle E_o^n \rangle_a$, where the subscript a refers to the threshold value. We now have to evaluate

$$\langle E_o^n \rangle_a = \int_{a^{1/2}}^{\infty} E_o^n P(E_o) \mathrm{d}E_o \Big/ \int_{a^{1/2}}^{\infty} P(E_o) \mathrm{d}E_o$$

$$\mathscr{H}_a = \mathscr{H} \int_{a^{1/2}}^{\infty} P(E_o) \mathrm{d}E_o. \tag{2.1}$$

2.1. *Space group* $P1$

For structures in $P1$ containing large numbers of atoms, Wilson (1949) has derived

$$P(E_o) = 2E_o \exp(-E_o^2). \tag{2.1.1}$$

Van Havere & Lenstra (1983*a*) showed that the asymptotical Wilson distribution is sufficiently accurate in most practical situations. Substitution of (2.1.1) in (2.1) gives

$$\langle E_o^\mu \rangle_a = \frac{\int_{a^{1/2}}^{\infty} E_o^{\mu+1} \exp(-E_o^2)\, \mathrm{d}E_o}{\int_{a^{1/2}}^{\infty} E_o \exp(-E_o^2)\, \mathrm{d}E_o}. \tag{2.1.2}$$

The denominator renormalizes the truncated E_o distribution so that $\langle E_o^0 \rangle_a = 1$. Realizing that (Magnus, Oberhettinger & Soni, 1966)

$$\Gamma(\xi,\alpha) \equiv \int_{\alpha}^{\infty} t^{\xi-1} \exp(-t)\, \mathrm{d}t, \tag{2.1.3}$$

where $\Gamma(\xi,\alpha)$ is an incomplete gamma function, we can write

$$\langle E_o^\mu \rangle_a = \frac{\Gamma(\mu/2+1,a)}{\Gamma(1,a)}. \tag{2.1.4}$$

Since we need here only the even moments, $\mu = 2n$ (see however Appendix A for the odd moments), we use the identity

$$\Gamma(n+1,a) = n! \exp(-a)\, e_n(a) \tag{2.1.5}$$

with $e_n(a) = \sum_{i=0}^{n} a^i/i!$, a truncated exponential series (Magnus, Oberhettinger & Soni, 1966). Combination of (2.1.4) and (2.1.5) gives

$$\langle E_o^{2n} \rangle = n!\, e_n(a)$$
$$\langle E_o^2 \rangle = 1 + a$$
$$\langle E_o^4 \rangle = 2 + 2a + a^2$$
$$\langle E_o^6 \rangle = 6 + 6a + 3a^2 + a^3. \qquad (2.1.6)$$

The results are the generalization of formulas given by Petit, Lenstra & Van Loock (1981).

2.2. *Space group* $P\bar{1}$

For structures in $P\bar{1}$ containing large numbers of atoms, Wilson has derived

$$P(E_o) = (2/\pi)^{1/2} \exp(-E_o^2/2). \qquad (2.2.1)$$

It was shown (Van Havere & Lenstra, 1983*a*) that this distribution is sufficiently accurate in most practical situations. We substitute (2.2.1) in (2.1), take $t = E_o^2/2$, use identity (2.1.3) and obtain

$$\langle E_o^{\mu} \rangle_a = 2^{\mu/2} \frac{\Gamma[(\mu + 1)/2, a/2]}{\Gamma(1/2, a/2)}. \qquad (2.2.2)$$

For the even moments, $\mu = 2n$ (see Appendix *B* for the odd moments), we can use the recursion relation (Magnus, Oberhettinger & Soni, 1966)

$$\Gamma(\xi,\alpha) = (\xi - 1)\Gamma(\xi - 1,\alpha) + \alpha^{\xi-1} \exp(-\alpha), \qquad (2.2.3)$$

which allows us to prove that

$$\langle E_o^{2n} \rangle_a = \frac{2^n \Gamma(n + \tfrac{1}{2})}{\Gamma(\tfrac{1}{2})} + \frac{2^n \exp(-a/2)}{\Gamma(1/2, a/2)} \sum_{i=1}^{n} \frac{\Gamma(n + \tfrac{1}{2})}{\Gamma(i + \tfrac{1}{2})} \left(\frac{a}{2}\right)^{i-1/2}. \qquad (2.2.4)$$

This equation can be simplified, using the identities (Magnus, Oberhettinger & Soni, 1966)

$$\Gamma(1/2, a/2) = \pi^{1/2} \operatorname{erfc}[(a/2)^{1/2}] \qquad (2.2.5)$$

$$\Gamma(n + \tfrac{1}{2}) = 2^{1-2n} \pi^{1/2} \Gamma(2n)/\Gamma(n) \qquad (2.2.6)$$

$$\Gamma(\tfrac{1}{2}) = \pi^{1/2} \qquad (2.2.7)$$

$$\Gamma(n + 1) = n!, \qquad (2.2.8)$$

to

$$\langle E_o^{2n} \rangle_a = 2^{1-n} \frac{(2n - 1)!}{(n - 1)!} + \frac{2^{-n} \exp(-a/2)}{\pi^{1/2} \operatorname{erfc}[(a/2)^{1/2}]} \times \sum_{i=1}^{n} 4^i \frac{(i - 1)!}{(n - 1)!} \frac{(2n - 1)!}{(2i - 1)!} \left(\frac{a}{2}\right)^{i-1/2}$$
$$\text{for } n = 1,2,3,\ldots. \qquad (2.2.9)$$

Equation (2.2.9) is a generalization of results by Petit, Lenstra & Van Loock (1981).

3. Discussion

Substitution of (2.1.6) into (3.1), (3.2) (Van Havere & Lenstra, 1983*c*) or substitution of (2.2.9) into (4.1), (4.2) (Van Havere & Lenstra, 1983*c*) allows one to evaluate the impact of the elimination of intensities $E_o^2 \leq a$ on the behaviour of $\langle R_2 \rangle$ and $\sigma^2(R_2)$ without reference to a specific structure. Since the results for $P1$ and $P\bar{1}$ differ only in a numerical way we restrict the discussion to $P1$ regarding the consequences on automated structure determination strategies.

Figs. 1 and 2 show the behaviour of $\langle R_2 \rangle$ as a function of model size and threshold a for two extreme situations, *viz* completely correct models $\{g,0\}$ and totally incorrect models $\{0,f\}$. We show the extreme situations because a four-dimensional graph would be necessary to depict the behaviour of the general in-between models $\{g,f\}$.

From Figs. 1 and 2 it is clear that the variation due to a threshold a is smaller for $\{g,0\}$ than for $\{0,f\}$ with $g = f$. This difference becomes understandable if one realizes that for correct models observed and calculated E values are correlated (*i.e.* small E_o values are more likely to be associated with small E_c values *etc.*). The

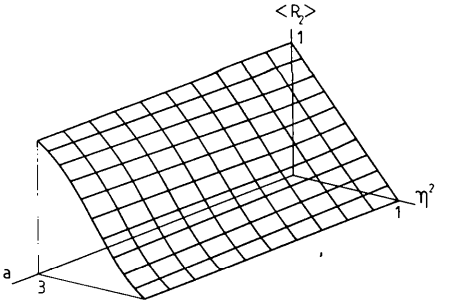

Fig. 1. Variation of $\langle R_2 \rangle$ as a function of model size and threshold a for models $\{g,0\}$ in $P1$.

relative differences $E_o^2 - \eta^2 E_c^2$ tend to be more evenly distributed over the data set and thus truncation of the set introduces only small changes in $\langle R_2 \rangle$. For incorrect models such a correlation does not exist resulting in the larger variation of $\langle R_2 \rangle$ with a.

The path of $\sigma(R_2)$ as a function of model size and threshold for situations $\{g,0\}$ and $\{0,f\}$ is depicted in Figs. 3 and 4, respectively, showing similar characteristics in variation. The figures also reveal another not-self-evident phenomenon: an initial decrease in $\sigma(R_2)$ with increasing values of a. The position of the minimum in $\sigma(R_2)$ for a particular model size depends on a. For correct models, the effect is very small (see Table 1), but becomes relatively large for incorrect models of large size. Unfortunately, however, the phenomenon cannot be used to increase the resolving power of R_2-based criteria by manipulating the data set through the threshold a. As a measure of the resolving power we take the quantity S, defined as

$$S = \frac{\langle R_2\{g,1\}\rangle - \langle R_2\{g+1,0\}\rangle}{3[\sigma(R_2\{g,1\}) + \sigma(R_2\{g+1,0\})]}. \quad (3.1)$$

This definition of resolving power matches the automation strategy in which iteratively new atoms are added to the known partial model. This strategy was found (Van Havere & Lenstra, 1983*c*) to be the one with the largest chance of success. As can be seen from Fig. 5 the introduction of a threshold always decreases S, and thus the usefulness of an R_2-based criterion.

A truncation of the data set *via* the threshold a is a selective way of reducing the number of reflections. It seemed of interest to investigate the impact on S when the number of reflections is reduced by an aselective mechanism. One has, for instance, a random selection mechanism from point-atom structures if one imposes a θ limit on the data set, where θ is the angle of

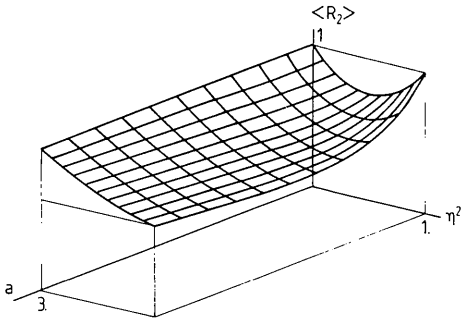

Fig. 2. Variation of $\langle R_2 \rangle$ as a function of model size and threshold a for models $\{0,f\}$ in $P1$.

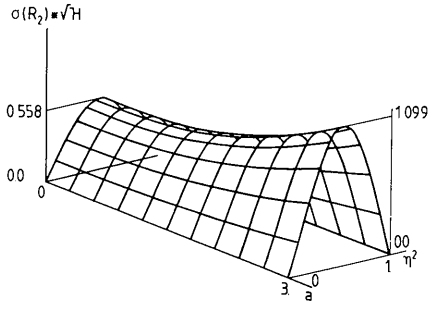

Fig. 3. Variation of $\sigma(R_2)$ as a function of model size and threshold a for models $\{g,0\}$ in $P1$.

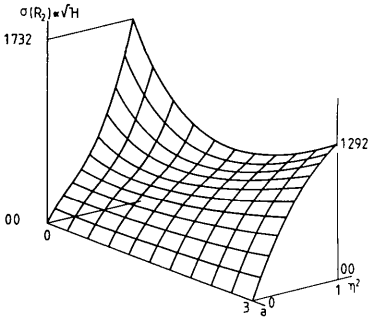

Fig. 4. Variation of $\sigma(R_2)$ as a function of model size and threshold a for models $\{0,f\}$ in $P1$.

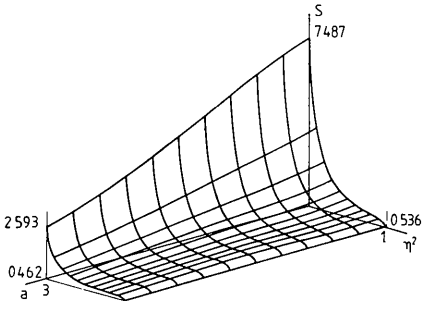

Fig. 5. Variation of resolving power S as a function of model size and threshold a in $P1$.

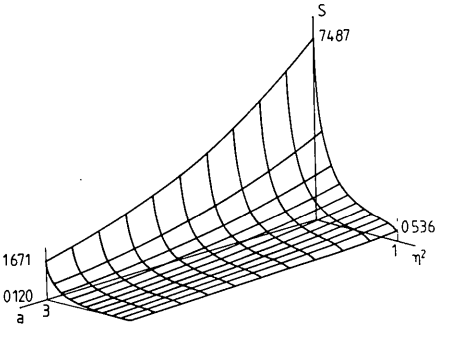

Fig. 6. Variation of S as a function of model size and number of reflections eliminated *via* the θ limit.

Table 1. *Minimum values of $\sigma^2(R_2)\mathscr{N}_a$ as a function of threshold a and model size*

Models are of type {g,0} in space group $P1$. For the sake of comparison the values of $\sigma^2(R_2)\mathscr{N}_a$ at $a = 0$ are given. $\mathscr{N}$ is the number of reflections in the data set.

Model $N = 10$	Minimum in $\sigma^2(R_2)\mathscr{N}_a$	at $a =$	$\sigma^2(R_2)\mathscr{N}_a$ at $a = 0$
{0,0}	0·00000	0·000	0·00000
{1,0}	0·14968	0·080	0·14973
{2,0}	0·29730	0·141	0·29772
{3,0}	0·42235	0·187	0·42355
{4,0}	0·51072	0·225	0·51286
{5,0}	0·55267	0·256	0·55551
{6,0}	0·54208	0·282	0·54510
{7,0}	0·47654	0·306	0·47965
{8,0}	0·35796	0·305	0·35943
{9,0}	0·19382	0·270	0·19420
{10,0}	0·00000	0·000	0·00000

diffraction. Values for $\langle R_2 \rangle$ and $\sigma(R_2)$ under the condition of a θ limit follow from the previous expressions by simply inserting the appropriate number of reflections $\mathscr{N}$ and putting $a = 0$.

In order to compare the impact of a on S (S_a) with the impact of the θ limit (S_θ) we eliminated the same number of reflections by both procedures. The results, presented in Figs. 5 and 6, respectively, clearly show that for all model sizes $S_a \geq S_\theta$. Thus, a random elimination of E_o values gives rise to a larger loss of resolving power than does an elimination of the same number of small E_o values.

WVH is grateful to the Belgian organization IWONL for financial support. The authors wish to thank Dr G. H. Petit and Mr J. F. Van Loock for stimulating discussions. The help of Professor H. J. Geise in the preparation of the manuscript is gratefully acknowledged.

APPENDIX *A*

The odd moments for space group $P1$ are obtained by putting $\mu = 2n + 1$, $n = 0, 1, 2, \ldots$, in equation (2.1.4):

$$\langle E_o^{2n+1} \rangle_a = \frac{\Gamma(n + \frac{3}{2}, a)}{\Gamma(1, a)}. \qquad (A.1)$$

This can be written, using equation (2.2.3), as

$$\langle E_o^{2n+1} \rangle_a = \frac{\Gamma(n + \frac{3}{2})\Gamma(\frac{1}{2}, a)}{\Gamma(\frac{1}{2})\Gamma(1, a)} + \frac{e^{-a}}{\Gamma(1, a)} \sum_{i=1}^{n+1} \frac{\Gamma(n + \frac{3}{2})}{\Gamma(i + \frac{1}{2})} a^{i-1/2}. \qquad (A.2)$$

Using identities (2.1.5) and (2.2.5–2.2.8) we get

$$\langle E_o^{2n+1} \rangle_a = 2^{-2(n+1/2)} \pi^{1/2} e^a \operatorname{erfc}(a^{1/2}) \frac{(2n + 1)!}{n!} + 2^{-2(n+1)} \frac{(2n + 1)!}{n!} \times \sum_{i=1}^{n+1} 4^i \frac{(i - 1)!}{(2i - 1)!} a^{i-1/2}. \qquad (A.3)$$

APPENDIX *B*

The odd moments for space group $P\bar{1}$ are obtained by putting $\mu = 2n + 1$, $n = 1, 2, \ldots$, in equation (2.2.2).

$$\langle E_o^{2n+1} \rangle_a = 2^{n+1/2} \frac{\Gamma(n + 1, a/2)}{\Gamma(1/2, a/2)}. \qquad (B.1)$$

Using equations (2.1.5) and (2.2.5) we can write this as

$$\langle E_o^{2n+1} \rangle_a = \frac{2^{n+1/2} n!\, e^{-a/2} e_n(a/2)}{\pi^{1/2} \operatorname{erfc}[(a/2)^{1/2}]}. \qquad (B.2)$$

References

MAGNUS, W., OBERHETTINGER, F. & SONI, R. P. (1966). *Formulas and Theorems for Special Functions of Mathematical Physics*. New York: Springer-Verlag.

PETIT, G. H., LENSTRA, A. T. H. & VAN LOOCK, J. F. (1981). *Acta Cryst.* A**37**, 353–360.

PETIT, G. H. & LENSTRA, A. T. H. (1982). *Acta Cryst.* A**38**, 67–70.

VAN HAVERE, W. K. L. & LENSTRA, A. T. H. (1983*a*). *Acta Cryst.* A**39**, 553–562.

VAN HAVERE, W. K. L. & LENSTRA, A. T. H. (1983*b*). *Acta Cryst.* A**39**, 562–565.

VAN HAVERE, W. K. L. & LENSTRA, A. T. H. (1983*c*). *Acta Cryst.* A**39**, 847–853.

WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.