derivatives where the heavy atoms of the derivatives occupy different positions. The distribution yields a reliable estimate (0 or π) for the invariant in the favorable case that the variance of the distribution is small. An example shows the improvement in estimates of the three-phase structure invariants which results from the ability now to exploit simultaneously the diffraction data from a triple of isomorphous structures, at least in the special case of a native protein and two heavy-atom derivatives in which the heavy atoms of the derivatives are located in different positions in the unit cell. Particularly noteworthy is the ease of unique origin and enantiomorph specification in direct-methods applications to all three structures.

It would be premature to assess, at this point, the role that the distributions will play in actual macromolecular structure determinations, or to compare the present technique with the standard multiple isomorphous replacement technique. As mentioned earlier, several questions remain to be answered, principally concerning the effects of errors in the diffraction data and of imperfect isomorphism. These questions are the subject of a present study and the results will be presented at a later date.

It should be stated in conclusion that, in view of the available evidence, the integrated direct methods–isomorphous replacement probability distributions constitute a sound theoretical basis for macromolecular phase determination.

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (SF), a Queen's University Research Award (SF); grant No. CHE-8203930 from the National Science Foundation (CMW and HH) and a grant from the James H. Cummings Foundation (CMW and HH).

References

Bernstein, F. C., Koetzler, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Crick, F. H. C. & Magdoff, B. (1956). *Acta Cryst.* **9**, 901–908.

Fortier, S., Weeks, C. & Hauptman, H. (1984). *Acta Cryst.* A**40**, 544–548.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.

Hauptman, H., Potter, S. & Weeks, C. (1982). *Acta Cryst.* A**38**, 294–300.

Timkovich, R. & Dickerson, R. E. (1973). *J. Mol. Biol.* **79**, 39–56.

Timkovich, R. & Dickerson, R. E. (1976). *J. Biol. Chem.* **251**, 4033–4046.

Acta Cryst. (1984). A**40**, 651-660

Exact Random-Walk Models in Crystallographic Statistics. I. Space Groups $P\bar{1}$ and $P1$

By Uri Shmueli

Department of Chemistry, Tel Aviv University, 69 978 *Tel Aviv, Israel*

George H. Weiss and James E. Kiefer

National Institutes of Health, Bethesda, Maryland 20205, *USA*

and Arthur J. C. Wilson

*Crystallographic Data Centre, University Chemical Laboratory, Cambridge CB*2 1*EW, England*

(*Received* 31 *October* 1983; *accepted* 5 *June* 1984)

Abstract

Probability density functions that are exact solutions to classical random-walk problems have been adapted to represent distributions of the magnitude of the normalized structure factor, for the space groups $P\bar{1}$ and $P1$. The functions are given by readily summable Fourier and Fourier–Bessel series, and account explicitly for the atomic composition of the asymmetric unit. These new probability density functions have been extensively tested by comparison with simulated histograms of $|E|$, for a wide range of atomic compositions. The most heterogeneous compositions examined are $C_{14}U$ and $C_{29}U$, for $P\bar{1}$ and $P1$, respectively. Very good agreement between the simulated and theoretical distributions has been obtained in all these tests, over the entire (useful) range $0<|E|<3$. A distribution of $|E|$ values, recalculated from published data on a triclinic platinum complex with chloroorganic ligands, has also been

0108-7673/84/060651-10$01.50

compared with the new probability functions and excellent agreement with the (expected) $P\bar{1}$ theoretical distribution has been obtained. The discrepancy between the recalculated distribution and the $P1$ theoretical rules out the latter space group both by visual comparison and quantitative discrepancy criteria. It is concluded that probability density functions are definitely preferable to moments in attempting to resolve a space-group ambiguity. Measures of discrepancy to be used in such statistical tests are proposed and discussed in some detail.

Introduction

Generalizations of probability density functions (p.d.f.'s) of the normalized structure factor and related statistics have so far been based on expansions of asymptotic distributions in terms of appropriate orthogonal polynomials (see, for example, Shmueli & Wilson, 1981, 1983). The coefficients of such expansions depend on factors such as symmetry, chemical composition and presence of dispersive scatterers, not allowed for by the asymptotic p.d.f.'s (Wilson, 1949; Shmueli, 1979). These generalized distributions are given by truncated expansions and are, therefore, approximate; improving their accuracy means adding more expansion terms of rapidly increasing complexity. The existing expansions can cope with problems that are due to the presence of outstandingly heavy atoms in all crystallographic space groups (Wilson, 1978; Shmueli & Kaldor, 1981, 1983; Shmueli, 1982*a*). However, for very heterogeneous asymmetric units in the space groups $P\bar{1}$ and $P1$, the discrepancies between the actual and the existing approximate generalized distributions are significantly large and may, possibly, hinder a successful resolution of a space-group ambiguity. Thus, either an extension of these generalized p.d.f.'s, or an alternative (exact) approach, is needed.

Exact p.d.f.'s of the structure factor, based on the solution of the random-walk problem (query: Pearson, 1905; solution: Kluyver, 1906) and applicable to $P\bar{1}$ and $P1$, were first introduced into the crystallographic literature by Hauptman & Karle (1952). However, Kluyver's solution can hardly be used as it stands, owing to its complexity, and generalization of its expansions to higher symmetries proved to be difficult, so the random-walk approach was abandoned in favour of truncated expansions (Karle & Hauptman, 1953; Hauptman & Karle, 1953*a*) of the types mentioned above.

The feasibility of a re-introduction of this conceptually attractive, but seemingly difficult approach, into crystallographic statistics was indicated to us by several recent achievements in random-walk methods [for a review see, for example, Weiss (1983)]. A problem, almost analogous to ours, related to statistics of combined sinusoidal waves (Rayleigh, 1880), has been investigated by Barakat (1974), in connection with his work on intensity distributions in laser speckle. Barakat observed that the amplitudes of the waves considered are bounded, and thus the p.d.f. of their combination (sum) can be represented by Fourier and Fourier–Bessel series, which are readily summable. This should be the case with the structure factor as well, since the maximum amplitudes of the 'waves' it combines are just the atomic scattering factors at a given temperature. Barakat's results have been discussed and further developed by Weiss & Kiefer (1983), who represented them in general forms that are applicable to one- and two-dimensional random-walk problems. Their expressions are valid for random walks with unequal step sizes. The latter authors (Weiss & Kiefer, 1983; Kiefer & Weiss, 1983) have also investigated steepest-descents approximations to random-walk p.d.f.'s (Daniels, 1954) and have shown them to compare favourably with other relevant approximate methods. Slightly earlier, Wilson (1983) showed that results due to Cramér (1938) could be applied to the equal-atom case; the expression obtained is equivalent to that from steepest descents.

The present paper introduces exact expressions for the probability density function of the magnitude of the normalized structure factor $|E|$, calculable to any accuracy for an arbitrary atomic composition of the asymmetric unit, for the space groups $P\bar{1}$ and $P1$. These new expressions are tested by comparing them with (i) simulated distributions corresponding to hypothetical structures of various degrees of atomic heterogeneity and (ii) a distribution of $|E|$ which has been recalculated from published atomic parameters of a triclinic complex of platinum with a chloroorganic moiety (Faggiani, Lippert & Lock, 1980). The statistical significance of the differences between experimental and theoretical distributions is examined in some detail, and quantitative discrepancy critieria, of importance in the resolution of space-group ambiguities, are proposed.

Probability density functions of $|E|$ in $P\bar{1}$ and $P1$

The original problem of the two-dimensional random walk (Pearson, 1905; Kluyver, 1906) assumes a sequence of n contiguous steps, of known lengths but random relative orientations, and requires the probability density function of the distance between the start and end points of this walk, usually called the end-to-end distance.

This general problem in statistics finds its counterparts in many branches of science. In crystallography, it was first identified by Hauptman & Karle (1952) with the problem of finding the probability density function of the magnitude of the structure factor, $|F|$. These authors introduced the vector polygon representation of the structure-factor equation

$$F(\mathbf{h})=\sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\,.\,\mathbf{r}_j), \tag{1}$$

where each of the atomic contributions to $F(\mathbf{h})$ is represented as a vector in the complex plane, $F(\mathbf{h})$ is the resultant, and its magnitude, $|F(\mathbf{h})|$, is comparable to the distance between the ends of the open polygon formed by these vectors (Hauptman & Karle, 1952).

Two assumptions must hold for the above identification to be valid: (i) the atomic contributions to $F(\mathbf{h})$ are independent, and (ii) the atomic phase angles, $2\pi\mathbf{h}\,.\,\mathbf{r}_j$, are uniformly distributed in the $[0, 2\pi]$ range. The second assumption is satisfied, in practice, if all the atoms are located in general positions and a large set of reflection data (*i.e.* reciprocal-lattice vectors $\mathbf{h}$) is considered. It should, however, be pointed out that the necessary uniformity of atomic phase angles may be described as arising from (i) a fixed structure and uniformly distributed reciprocal-lattice vectors, or (ii) a fixed $\mathbf{h}$ and atoms uniformly distributed throughout the unit cell. The p.d.f. of the structure factor is the same for the above two situations only if the set of data is infinite (Hauptman & Karle, 1953*b*; Giacovazzo, 1977). Wilson (1981) has discussed the possibility of allowing for non-independence of atomic positions in expansions of the Hermite–Laguerre type.

Space group $P\bar{1}$

The appropriate random-walk p.d.f. is that of the end-to-end distance projected onto an arbitrary direction, and the real axis may conveniently be chosen for this purpose. The projected end-to-end distance of a classical two-dimensional random walk is given by $|R'|$, where

$$R'=\sum_{j=1}^{n} L_j \cos\theta_j, \tag{2}$$

n is the number of steps, L_j is the length of the jth step, θ_j is the angle it forms with the real axis, and the probability that R' lies between R' and $R'+\mathrm{d}R'$ can be represented by the Fourier series

$$P(R')\,\mathrm{d}R'=\frac{1}{2L_T}\left\{1+2\sum_{m=1}^{\infty} C_m \cos(\pi m R'/L_T)\right\}\mathrm{d}R', \tag{3}$$

where

$$C_m==\prod_{j=1}^{n} J_0(\pi m L_j/L_T), \tag{4}$$

J_0 is a Bessel function of zero order and $L_T=\sum_{j=1}^{n} L_j$ is the maximum extension of the walk (Barakat, 1974; Weiss & Kiefer, 1984).

The corresponding expression for the p.d.f. of a centrosymmetric structure factor, in the space group $P\bar{1}$, can now be readily obtained by identifying the number of steps with one half the number of atoms per unit cell, the step length with twice the atomic scattering factor, and the maximum extension of the walk with the sum of the scattering factors. The sum of the projected pairs of steps is the structure factor

$$F(\mathbf{h})=2\sum_{j=1}^{N/2} f_j \cos(2\pi\mathbf{h}\,.\,\mathbf{r}_j) \tag{5}$$

itself. Introducing the above replacements into (3) and (4), we obtain

$$P(F)=\frac{1}{2S_1}\left\{1+2\sum_{m=1}^{\infty} C_m^{(\bar{1})}\cos(\pi m F/S_1)\right\}, \tag{6}$$

where

$$C_m^{(\bar{1})}=\prod_{j=1}^{N/2} J_0(2\pi m f_j/S_1) \tag{7}$$

and

$$S_1=\sum_{j=1}^{N} f_j \tag{8}$$

for the p.d.f. of (5). Since $P(F)=P(-F)$, as expected, we obtain the p.d.f. of the magnitude of F by doubling the right hand side of (6).

The transition to $|E|$, the magnitude of the normalized structure factor, is readily achieved by noting that

$$P(|F|)=P(|E|)\bigg/\left(\sum_{j=1}^{N} f_j^2\right)^{1/2} \tag{9}$$

(see, for example, Shmueli & Wilson, 1981). The resulting p.d.f. of $|E|$ can then be written as

$$P(|E|)=\alpha\left\{1+2\sum_{m=1}^{\infty} C_m^{(\bar{1})}\cos(\pi m\alpha|E|)\right\},\qquad 0<|E|<|E|_{\max}, \tag{10}$$

where the coefficients $C_m^{(\bar{1})}$ are defined as in (7), and

$$\alpha=\left(\sum_{n=1}^{N} f_j^2\right)^{1/2}\bigg/\left(\sum_{j=1}^{N} f_j\right)\equiv\Sigma^{1/2}/S_1. \tag{11}$$

For the equal-atom case, (10) simplifies to

$$P(|E|)=\frac{1}{N^{1/2}}\left\{1+2\sum_{m=1}^{\infty}\left[J_0\left(\frac{2\pi m}{N}\right)\right]^{N/2}\times\cos(\pi m\alpha|E|)\right\}. \tag{12}$$

Equation (11) will serve for the evaluation of the distribution when the number of atoms is small enough to become a disturbing factor. However, for $N=30$, $P(|E|)$ already deviates from the Wilson-type Gaussian p.d.f., $(2/\pi)^{1/2}\exp(-E^2/2)$, by less than 2%, throughout the $[0, 3]$ range of $|E|$. The convergence of the Fourier series in (10) has been examined and it appears that 40 terms are sufficient for a hypothetical asymmetric unit containing 14 carbon atoms and one uranium. In general, the number of terms required increases with increasing heterogeneity and decreasing numbers of atoms in the asymmetric unit. However, (10) is so readily evaluated that the number of terms should not present any difficulties.

*Space group P*1

The model density is now the p.d.f. of the end-to-end distance itself, as can be seen from the foregoing considerations. This can be treated as a joint density of the real and imaginary parts of the structure factor, where the phase is subsequently integrated out, but the resulting double Fourier series is less convenient to handle than the 'radial' distribution derived by Barakat (1974) and discussed by Kiefer & Weiss (1983). Denoting the end-to-end distance by

$$|R| = \left|\sum_{j=1}^{n} L_j \exp(i\theta_j)\right|, \qquad (13)$$

where the symbols have the same meaning as in (2), the probability density function of $|R|$ is given by

$$P(|R|) = \frac{2|R|}{L_T^2} \sum_{m=1}^{\infty} D_m J_0(\gamma_m |R|/L_T), \quad 0<|R|<L_T, \qquad (14)$$

where γ_m are successive zeros of the Bessel function $J_0(x)$(see, for example, Abramowitz & Stegun, 1972), the expansion coefficients in (14) have the form

$$D_m = \frac{1}{J_1^2(\gamma_m)} \prod_{j=1}^{n} J_0(\gamma_m L_j/L_T), \qquad (15)$$

where $J_1(x)$ is a Bessel function of order one and L_T is the maximum extension of the walk, defined as in (3).

The translation of (14) to the p.d.f. of $|E|$ for the space group $P1$ proceeds in an analogous manner to that described above. We obtain

$$P(|E|) = 2\alpha^2|E| \sum_{m=1}^{\infty} D_m^{(1)} J_0(\gamma_m \alpha |E|), \quad 0<|E|<|E|_{\max}, \qquad (16)$$

where

$$D_m^{(1)} = \frac{1}{J_1^2(\gamma_m)} \prod_{j=1}^{N} J_0(\gamma_m f_j/S_1) \qquad (17)$$

and the quantities S_1 and α have the same meaning as in (11).

Equations (10) and (16) depend explicitly on the atomic composition of the asymmetric unit and thus constitute formal solutions to the problem of effects of atomic heterogeneity on intensity statistics in the above two space groups. The fact that these p.d.f.'s are given as infinite series does not diminish their practical value, since both series converge rapidly. We have examined the numerical effects of the precision that was used by computing the Bessel functions and their zeroes (*cf.* Appendix *A*), and found the usual polynomial approximations to Bessel functions of orders zero and one (see, for example, Abramowitz & Stegun, 1972) adequate for most practical purposes.

Expressions for exact cumulants and moments of the projection p.d.f., corresponding to (3), have been derived by Kiefer & Weiss (1983) and we have compared them with those of the five-term distribution for $P\bar{1}$, derived by Shmueli (1982*a*). The comparison was subject to the same replacements and identifications used in translating the random-walk p.d.f. to that of the normalized structure factor. A complete agreement was obtained for the first five even moments of $|E|$, which have been expanded in detail starting from the closed expressions furnished by both methods. The agreement is, of course, not surprising since the 2nth moment of either of the p.d.f.'s given as expansions in terms of orthogonal polynomials requires only the first n terms of the expansion (Shmueli & Wilson, 1981).

Simulated distributions

The performance of the above probability density functions has been tested by comparing them with simulated distributions of $|E|$, for some problematic atomic compositions. Such simulations have been described in some detail by Shmueli (1982*b*), and are recalled below.

The uniform distribution of the atomic phase angles (see above) is simulated by replacing the scalar products $\mathbf{h}.\mathbf{r}_j$, in (1), with computer-generated random numbers, uniform in the [0, 1] range. Of course, this is justified only if the atoms (and especially the heavy scatterers) do not occupy special positions, and the set of data to be considered is large. Atomic scattering factors are replaced with quantities proportional to atomic numbers, since the normalized structure factor depends on the ratios $f_j/(\sum_{k=1}^{N} f_k^2)^{1/2}$ and the latter do not depend strongly on the Bragg angle. Thus, for example, the normalized structure factor for a unit cell containing 29 carbons and one uranium can be written, for the present purpose, as

$$E = \sum_{j=1}^{30} a_j \exp(i\theta_j)/\langle|F'|^2\rangle^{1/2}, \qquad (18)$$

where a_j equals 1 or $15\frac{1}{3}$, according as j corresponds to carbon or uranium, respectively, and θ_j is uniform in the $[0, 2\pi]$ range; the denominator in (18) can be computed as the root mean square of the magnitude of the numerator, over the simulated sample, or approximated by $\sum_k a_k^2$ [in analogy with Wilson's (1942) Σ].

The present experiments are based on samples of 3000 $|E|$'s each, and the histograms are constructed in the $[0, 3]$ range of $|E|$, the counts being recorded for thirty channels of equal widths in the above range. The normalization is performed with simulated rather than estimated sigma's, in order to provide another check on the simulation.

Fig. 1(a) and (b) show the results of such simulations, and their comparison with theoretical probability density funitions: the 'exact' ones, *i.e.* those computed from (10) and (16), and the appropriate centric and acentric asymptotic p.d.f.'s based on the central limit theorem approach (Wilson, 1949). The hypothetical atomic compositions chosen are $C_{14}U$ and $C_{29}U$ asymmetric units of the space groups $P\bar{1}$ and $P1$ corresponding to Fig. 1(a) and (b), respectively. Fig. 1(b) also contains a p.d.f. of $|E|$, based on the p.d.f. of $|F|$ given by Sim (1958) for a single heavy atom, and a number of light ones, in the unit cell of space group $P1$. The p.d.f.'s are scaled to the histograms as explained by Shmueli (1982b).

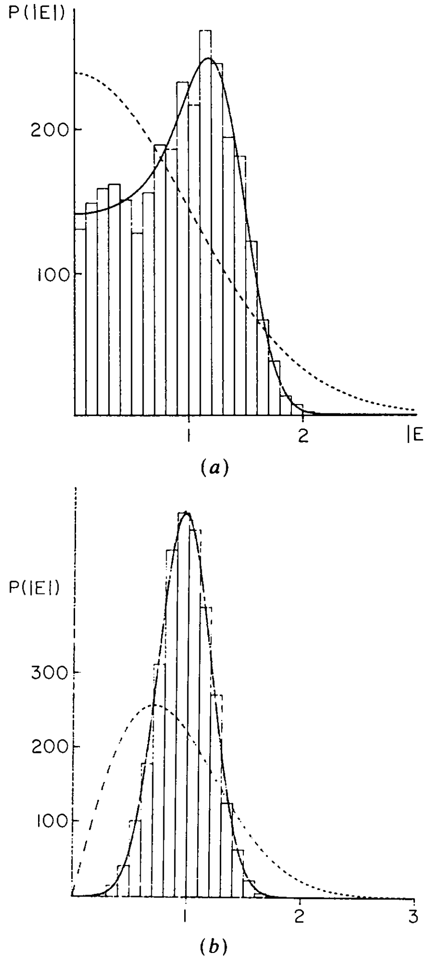

Fig. 1. Simulated and theoretical distributions of $|E|$. The theoretical p.d.f.'s given in each figure are scaled to the histogram. The solid lines denote the random-walk p.d.f.'s and the (well resolved) dashed lines correspond to the central-limit asymptotic p.d.f.'s. The height of each rectangle equals the number of $|E|$ values which lie within the corresponding histogram channel. (a) A $C_{14}U$ asymmetric unit in $P\bar{1}$, (b) a $C_{29}U$ asymmetric unit in $P1$; a dashed curve, which is nearly indistinguishable from the solid one in b, corresponds to Sim's (1958) p.d.f. for a single heavy atom in $P1$.

The agreement of the Fourier (10) and the Fourier–Bessel (16) expansions for the p.d.f.'s of $|E|$ with the simulated histograms of this quantity, is very good throughout the range of $|E|$ considered. The inability of the asymptotic p.d.f.'s to account for these distributions is uue to the breakdown of the central limit theorem for such highly heterogeneous sums of random variables. Neither can the existing expansions of the p.d.f.'s in terms of orthogonal polynomials cope with such heterogeneities throughout the useful range of $|E|$, for $P\bar{1}$ and $P1$ (Shmueli, 1982a, b). On the other hand, the remarkable agreement of Sim's (1958) p.d.f. with the histogram in Fig. 1(b) [discrepancy indices: $\chi^2 = 7{\cdot}93$, $k = 14$, $R = 0{\cdot}035$ (see below)] is because the Wilson (1949) distribution is applied, in his derivation, to the light-atom part of the structure (Sim, 1958) only. Similar equations are given for two equal heavy atoms in triclinic space groups (Srinivasan & Parthasarathy, 1976), as integrals that have to be numerically evaluated.

Many similar simulations, for other assumed compositions, lead to comparable visual and quantitative (see below) agreement with (10) and (16).

A real example

Several distributions of $|E|$, which were recalculated from published and well refined structural parameters, have been previously compared with generalized expansions for the cumulative distribution functions of $|E|$ (Shmueli, 1982a). One of the structures that were tested in the latter study, a complex of platinum with a chloroorganic moiety crystallizing in the space group $P\bar{1}$ ($C_6Cl_2N_4O_4Pt$, Faggiani, Lippert & Lock, 1980), has been chosen for the present comparison.

The magnitudes of the normalized structure factors E have been obtained with the aid of program *NORMAL* of the *MULTAN*80 system (Main, Fiske, Hull, Lessinger, Germain, Declercq & Woolfson, 1980) for 2871 recalculated independent structure factors in the copper sphere. This version of *NORMAL* has also been modified to produce a histogram of $|E|$ and the remaining output required by subsequent statistical routines (*cf.* Shmueli, 1982a, where such modifications were outlined for *NORMAL* of the *MULTAN*78 system).

The composition-dependent terms in (10) and (16), *i.e.* the unitary scattering factors $f_j/(\sum_{k=1}^{N} f_k)$ [*cf.* (7)] and α [*cf.* (11)], have been obtained as weighted averages over the overlapping shells used by *NORMAL* in the construction of the Wilson plot (*cf.* Shmueli, 1982a). Thus, for example, the mean value of α from (11) is computed as

$$\bar{\alpha} = \sum_{i=1}^{N_B} n_i(\Sigma^{1/2}/S_1)_i / \sum_{i=1}^{N_B} n_i, \qquad (19)$$

where n_i is the number of reflections in the ith shell and N_B is the number of shells or number of points in the Wilson plot.

The distribution (histogram) of the recalculated $|E|$ values and the theoretical probability density functions (10) and (16), scaled to the histogram, are shown in Fig. 2. Both theoretical p.d.f.'s refer to the observed unit-cell contents of the structure considered.

It is seen from Fig. 2 that the centrosymmetric p.d.f. (10) agrees remarkably well with the histogram of $|E|$ throughout the range, while the agreement with the non-centrosymmetric Fourier–Bessel series (16) is much worse, even from a visual inspection of the graphs alone. The space group $P\bar{1}$ is thus correctly indicated.

Fig. 2 illustrates a successful application of exact random-walk distributions to the heavy-atom problem in intensity statistics, but it also displays the problem of space-group ambiguity to its full extent. It is seen that the non-centrosymmetric p.d.f. deviates from the centrosymmetric one mainly in the regions of small $|E|$ values and those around the peaks of the distributions. The inclusion of 'zeros' and 'unobserved' reflections in such tests is therefore imperative. The non-centrosymmetric p.d.f. in Fig. 2 is appreciably broader than that in Fig. 1*b*, because there are two platinum atoms in a 34-atom cell of $P\bar{1}$ (excluding H atoms). Also, the modes of the theoretical $P\bar{1}$ and $P1$ distributions are fairly close to each other and the overlap of the distributions is considerable. A possible consequence of this situation is the limited value of the moments of $|E|$ as sole discriminators between the two possible space groups, in such extreme cases. Thus, the values of $\langle|E|^4\rangle$, in the present example, are 1·69 (from the recalculated distribution of $|E|$ for the published structure), 1·72 (for $P\bar{1}$) and 1·57 (for $P1$), the latter two values being obtained by the method of Shmueli & Wilson (1981). This is clearly a narrow margin, even though the indication is correct. The cumulative distribution function provides a much better discrimination between the space groups in question, but it is not readily interpretable in terms of the fine details of the distribution, which may be needed in non-trivial situations. The most meaningful statistical test still remains the direct comparison of the experimental distribution with the possible probability density functions.

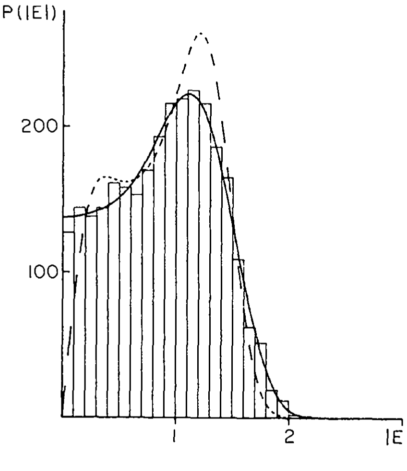

Fig. 2. A histogram of $|E|$ for a solved triclinic platinum complex and its comparison with centrosymmetric and non-centrosymmetric triclinic random-walk p.d.f.'s. The solid curve corresponds to the centrosymmetric Fourier Series (10) and the dashed one to the non-centrosymmetric Fourier–Bessel series (16). Both theoretical p.d.f.'s have been scaled to the histogram. The height of each rectangle equals the number of $|E|$ values that lie in the corresponding histogram channel. For explanations see text.

The need for discrepancy critieria that may also provide statistical measures of confidence, with which a space group is indeed preferably indicated, is apparent and such criteria are treated in the next section.

Significance of differences between distributions

Consider a histogram, which has been constructed by sorting N observations among k (not necessarily equal) channels, and let $p(x)$ be a given theoretical p.d.f. to be compared with the histogram. In order to compare the observed and theoretical densities, we require the actual and expected numbers of observations falling in each channel of the histogram, as well as appropriate measures of discrepancy between these two sets of numbers.

Let n_i and m_i be the actual and the expected numbers of observations falling in the ith channel of the histogram. The discrepancy between these sets of numbers can be conveniently estimated by using a residual such as the familiar R, defined by

$$R = \left[\frac{\sum_i w_i(n_i - m_i)^2}{\sum_i w_i n_i^2}\right]^{1/2}, \qquad (20)$$

where w_i is a weight, or by evaluating

$$\chi^2 = \sum_i (n_i - m_i)^2 / m_i, \qquad (21)$$

which is a widely employed measure in statistical practice.

In order to evaluate the m_i, we observe that the probability of an observation falling in the ith channel is given by

$$\alpha_i = \int_{x_{i-1}}^{x_i} p(x)\,\mathrm{d}x, \qquad (22)$$

where $[x_{i-1}, x_i]$ is the range of the variable x (in our case $|E|$) spanned by the ith channel. If the width of the channel is small, (22) can often be approximated by

$$\alpha_i = p[(x_{i-1}+x_i)/2]\Delta x_i, \tag{23}$$

where $\Delta x_i = x_i - x_{i-1}$ is the width of the ith channel. The expected number of observations falling in this channel is thus

$$m_i = N\alpha_i, \tag{24}$$

where $N = \sum_{i=1}^{k} n_i$ is the total number of the observations, and the m_i can be simply evaluated by computing the theoretical p.d.f. at the midpoints of the channels. The discrepancy measures R and χ^2 are very often informative as they stand, especially when a given experimental distribution is compared with two or more possible theoretical ones (*cf.* Table 1). However, they also have statistical interpretations. Especially well known are the tables of percentage points of the χ^2 distribution [*e.g.* Sachs, 1982] which permit one to obtain the probability that a value of χ^2, for a given number of degrees of freedom [in our case, the effective number of channels (see below)], corresponds to a good fit between the theoretical and experimental distributions – at a given confidence level. An alternative, albeit more qualitative, procedure is an examination of the expected values of R and χ^2, given in this paper.

As shown in Appendix *B*, the expected value and the variance of χ^2 are given by

$$\langle\chi^2\rangle = k-1 \tag{25}$$

and

$$\sigma^2(\chi^2) = 2(k-1), \tag{26}$$

respectively, where k is the actual number of terms in the summation for χ^2 [*cf.* (21)]; this may equal, or be smaller than, the number of channels in the histogram (see below). Hence, if an observed value of χ^2 much exceeds

$$\langle\chi^2\rangle + 2\sigma(\chi^2) = (k-1) + 2[2(k-1)]^{1/2} \tag{27}$$

it is likely* that the distribution of n (the histogram) is significantly different from that of the given p.d.f. with which the histogram is compared. For $k = 25$, which is typical for the number of effective channels in the examples cited in this paper, this limit is about 38.

Similarly, it is shown in Appendix *B* that the expected value and variance of R^2 are given by

$$\langle R^2\rangle = \frac{1}{N}\left[\left(\sum_i \alpha_i^2\right)^{-1} - 1\right] \tag{28}$$

and

$$\sigma^2(R^2) = 2\sum_i \sigma_i^4 / (\sum_i m_i^2)^2, \tag{29}$$

respectively, where

$$\sigma_i^2 = N\alpha_i(1-\alpha_i) \tag{30}$$

is the variance of the (binomial) distribution of n_i (*cf.* Appendix *B*), or very crudely

$$\sigma^2(R^2) \simeq 2\langle R^2\rangle / N \tag{31}$$

on using (30) and (B9). Equations (28), (29) and (31) are based on assumed unit weights in (20).

Considering the probabilistic aspect of the discrepancies, the theoretical p.d.f. of the structure factor applies to all the configurations that correspond to a given space-group symmetry and atomic composition, and is thus of the 'fixed **h** and random **r**' type, while the true p.d.f. of the observations is related to a certain (albeit unknown) fixed structure and a finite set of diffraction data. Such two p.d.f.'s are, in principle, different (*e.g.* Giaçovazzo, 1977) and this should be reflected in the discrepancy between the theoretical and experimental distributions. This is another aspect of the finite sampling, to which the discrepancies are due in part.

Results and some practical considerations

The measures of discrepancy discussed above have been applied to the results presented in Figs. 1 and 2 as well as to other distributions, which are not displayed. Table 1 summarizes the values of R, χ^2 as well as the effective number of channels used in the computation of the latter.

It appears that the χ^2 criterion is a rather sensitive one, and its value for a correct distribution is usually smaller than the effective number of channels used. Similarly, the value of R for a correct distribution also tends to be lower than the corresponding expected value that follows from (*B*9). These expected values of χ^2 and R thus appear to be useful indicators of discrepancies that are associated with a good fit.

Table 1 shows that the presence of two heavy atoms in the asymmetric unit greatly decreases the effect of atomic heterogeneity on the distributions, as compared with one heavy-atom only. This is in agreement with predictions and observations reported elsewhere Shmueli, 1982*b*), and also explains the numerous successful applications of Wilson (1949) statistics to situations where it would not be expected to work.

Thinly populated channels pose no particular problem with R^2; they simply add small increments to both numerator and denominator, without greatly affecting the ratio. With χ^2, however, small values of m_i give rise to considerable uncertainties, since the effect of a particular difference $n_i - m_i$ is greatly

* If the distribution were normal a range of two standard deviations (more precisely 1·96 standard deviations) on either side of the mean value would correspond to 95% confidence limits.

Table 1. *Discrepancy measures for comparison of simulated and recalculated distributions with asymptotic and exact p.d.f.'s*

For simulations, the assumed composition of the asymmetric unit is C_mX_p and the indicator of heterogeneity is denoted by $\rho = Z_X/Z_C$, where Z is the atomic number, *e.g.* $\rho = 15\frac{1}{3}$ means that the heavy atom is uranium. The number of atoms in the asymmetric unit is taken as 15 and 30 for the space groups $P\bar{1}$ and $P1$, respectively. Discrepancy measures: the subscripts on R and χ^2 are 10, 16, $\bar{1}$ or 1, according as the distribution is compared with (10), (16), the Wilson (1949) centric p.d.f. or the Wilson (1949) acentric p.d.f., respectively. The effective number of channels which participate in the calulation of χ^2 is denoted by k_{10}, k_{16}, $k_{\bar{1}}$ or k_1, where the subscripts have the same meaning as for R and χ^2 above.

Simulated $P\bar{1}$ distributions

m	p	ρ	χ^2_{10}	k_{10}	$\chi^2_{\bar{1}}$	$k_{\bar{1}}$	R_{10}	$R_{\bar{1}}$
14	1	5	18·5	26	89·2	29	0·058	0·141
14	1	10	16·7	22	549·2	29	0·068	0·365
14	1	$15\frac{1}{3}$	20·2	20	1028·7	29	0·080	0·465
13	2	5	24·3	27	61·9	29	0·082	0·104
13	2	10	19·5	24	165·6	29	0·071	0·156
13	2	$15\frac{1}{3}$	22·1	23	258·3	29	0·077	0·197

Simulated $P1$ distributions

m	p	ρ	χ^2_{16}	k_{16}	χ^2_1	k_1	R_{16}	R_1
29	1	5	18·8	23	96·3	25	0·071	0·151
29	1	10	13·1	18	871·8	25	0·054	0·415
29	1	$15\frac{1}{3}$	8·8	14	2167·8	25	0·029	0·596
28	2	5	11·1	23	62·5	25	0·050	0·109
28	2	10	17·0	20	347·3	25	0·061	0·275
28	2	$15\frac{1}{3}$	16·3	18	691·6	25	0·063	0·387

Distribution recalculated for the solved chloroplatinate (Fig. 2)

$\chi^2_{10} = 9.58$	$k_{10} = 20$	$R_{10} = 0{\cdot}047$	(test for $P\bar{1}$)
$\chi^2_{16} = 405{\cdot}0$	$k_{16} = 19$	$R_{16} = 0{\cdot}183$	(test for $P1$)

inflated when m_i is small. This effect is well recognized and most statistical texts recommend aggregating adjacent thinly populated channels until an acceptable aggregate value of m_i is attained. Cramér (1945) suggests a minimum m_i of ten; earlier books suggest a higher value, later books (*e.g.* Sachs, 1982) only four. The value of the m_i threshold used in Table 1 is five. It should be noted that the correct procedure is to aggregate channels until the set minimum is achieved, and not to omit channels with small m_i.

A comment on the effect of experimental errors is in order. When an experimental p.d.f. is being compared with a theoretical one, the difference between n_i and m_i will depend not only on the sampling fluctuations discussed above, and on the difference between the true p.d.f. of the observations and the theoretical one, but also on the experimental errors in the observed p.d.f. The experimental errors may be of two kinds: (i) systematic errors, which shift the center of gravity of the p.d.f., and (ii) random, which blur its details but do not alter its mean. An example of type (i) is uncorrected extinction, which shifts the higher intensities systematically to lower values; an example of type (ii) is statistical fluctuations in counting rates, which lead to shifts that may result in higher or lower values of the intensities.

A quantitative discussion of these effects will be presented elsewhere.

Concluding remarks

The present paper contains the first proposal of exact probability density functions of the magnitude of the normalized structure factor that depend explicitly on the atomic composition of the asymmetric unit and can be readily computed to any desired accuracy. The present simulated (Fig. 1, Table 1) and recalculated (Fig. 2) distributions, for moderately as well as highly heterogeneous asymmetric units, agree remarkably well with the appropriate theoretical p.d.f.'s.

One of the important, but not often realized, points made in the present study is the visual superiority of probability density functions over cumulative distributions and moments in practical intensity statistics. When an observed distribution is compared with the possible centric and acentric p.d.f.'s, the asymmetric unit taken for the construction of the acentric p.d.f. is of necessity twice as large as that for the centric one, and the heterogeneity of the acentric model is considerably decreased. This situation leads, in the presence of outstandingly heavy atoms, to little apparent difference between the theoretical cumulative distributions for intermediate and high $|E|$ values. On the other hand, the difference between the corresponding p.d.f.'s is much more obvious [compare Fig. 1*b* of Shmueli (1982*a*) with Fig. 2 in this paper, which both refer to the same solved structure]. This is particularly important when there is an appreciable proportion of unobserved reflections and proper advantage cannot be taken of the marked difference between the theoretical p.d.f.'s at the low side of the distribution.

Although a graphical representation of the results of a statistical test is often the most convincing one, the measures of discrepancy dealt with above are very informative and may also be indispensable if the margin of discrimination between distributions is particularly narrow. Both χ^2 and R are conveniently computed, and the former has the advantage of being useful in estimating the probability of an identification being formally correct (*e.g.* Sachs, 1982).

We conclude with a comment on the 'exact' p.d.f.'s presented in this paper, compared with the expansions in terms of Hermite and Laguerre polynomials (Shmueli & Wilson, 1981, 1982; Shmueli, 1982*a*, *b*). At present, the orthogonal-polynomial expansions can be evaluated for any space group, given the necessary moments, while the accurate and simpler random-walk p.d.f.'s can be used only for $P\bar{1}$ and $P1$. Since the departures of experimental p.d.f.'s from the popular asymptotic ones (Wilson, 1949) are usually largest for low symmetries, it appears logical to try and replace the Hermite–Laguerre p.d.f.'s by exact statistics for these symmetries. The extension of the present study to the monoclinic system is in progress and will be reported at a later date.

APPENDIX A
Calculation of the zeros of Bessel functions

Since the number of roots of the equation $J_0(\gamma)=0$, required for evaluating (16) or similar Fourier–Bessel expansions, may be quite large, it seems desirable to summarize a convenient algorithm for their computation. In outline, the method consists of using an initial approximation to the zero given by Abramowitz & Stegun (1972), and refining the initial estimate by the Newton–Raphson method (*e.g.* Hamming, 1973). The lowest-order approximation for the nth root of $J_0(\gamma)=0$ is

$$\gamma_n=\beta+\frac{1}{8\beta}-\frac{124}{3(8\beta)^3}+\frac{120\,928}{15(8\beta)^5}-\frac{401\,743\,168}{105(8\beta)^7}+\ldots, \qquad (A1)$$

where $\beta=(n-1/4)\pi$. For $n>5$ the values given by (A1) have a relative error less than 10^{-11} so that no refinement is needed for the higher zeroes. Tables of zeros of Bessel functions (Table 9.5, Abramowitz & Stegun, 1972, and references quoted therein) can be used for checking out the above procedure.

APPENDIX B
Expected values and variances of χ^2 and R^2

Given a set of observations n_1, n_2, ..., n_k and a corresponding set of their expected values, it is most unlikely that the ith observation (or channel) n_i equal exactly its expected value m_i. The p.d.f. of any particular distribution $n_1,\ldots,n_k$ is given by the multinomial expression

$$p(n_1,\ldots,n_k)=\frac{N!\alpha_1^{n_1}\ldots\alpha_k^{n_k}}{n_1!\ldots n_k!} \qquad (B1)$$

(Cramér, 1945, pp. 318–319; Johnson & Kotz, 1969, pp. 281–291). Any particular n_i has a binomial distribution with parameters N and α_i [*cf.* (23) and (24)], so that its variance is

$$\sigma_i^2=N\alpha_i(1-\alpha_i). \qquad (B2)$$

However, the n_i are not independent variables, since $N=\sum_{i=1}^k n_i$, where k is the number of channels. The covariance of any pair is

$$\operatorname{cov}(n_i,n_j)=-N\alpha_i\alpha_j, \qquad (B3)$$

which is negligible only if the peak region of the p.d.f. spans a large number of channels.

The required statistics of χ^2 and R^2 are now readily evaluated. We have, using (B2), (24) and the condition: $\sum_{i=1}^k\alpha_i=1$

$$\langle\chi^2\rangle=\sum_i\frac{\langle(n_i-m_i)^2\rangle}{m_i} \qquad (B4)$$

$$=\sum_i\frac{\sigma_i^2}{m_i} \qquad (B5)$$

$$=\sum_i(1-\alpha_i) \qquad (B6)$$

$$=k-1 \qquad (B7)$$

and the variance of the χ^2 distribution is given (for large N) by (26) in the text. These statistics of χ^2 thus depend on the number of channels alone.

The calculations of $\langle R^2\rangle$ and $\sigma^2(R^2)$ are more complicated when the correct definition of R^2 [(20)] is used. However, noting that the use of n_i in the denominator of (20) is not significantly different from the use of m_i, and making the appropriate replacement, the derivations are greatly simplified. We thus have

$$\langle R^2\rangle=\sum_i\sigma_i^2/\sum_i m_i^2 \qquad (B8)$$

$$=N\sum_i\alpha_i(1-\alpha_i)/\sum_i m_i, \qquad (B9)$$

which readily leads to (28) in the text. The variance of R^2 is obtained in a similar manner, by first evaluating $\langle R^4\rangle$, from

$$\sigma^2(R^2)=\langle R^4\rangle-\langle R^2\rangle^2.$$

The correspondence between R^2 and χ^2 turns out to be rather close. In much crystallographic work weights proportional to $1/\sigma_i^2$ are used, or in this case $1/[m_i(1-\alpha_i)]$ [*cf.* (24) and (30)]. If the number of channels k is reasonably large this differs little from $1/m_i$, the factor required to convert the numerator of (20) into χ^2. The weighted value of R^2 is thus

$$R^2\simeq N^{-1}\chi^2. \qquad (B10)$$

It should, however, be remembered that R^2 depends on the actual value of k that was used for the construction of the histogram, while k for χ^2 is the effective number of channels, *i.e.* those channels for which the value of m_i exceeds some threshold, which was taken as 5 in this paper.

References

ABRAMOWITZ, M. & STEGUN, I. (1972). *Handbook of Mathematical Functions.* New York: Dover.

BARAKAT, R. (1974). *Opt. Acta*, **21**, 903–921.

CRAMÉR, H. (1938). *Sur un Nouveau Théorème – Limite de la Théorie des Probabilités.* No. 736 in the Series *Actualités Scientifiques et Industrielles.* Paris: Hermann.

CRAMÉR, H. (1945). *Mathematical Methods of Statistics.* Uppsala: Almqvist and Wiksells.

DANIELS, H. E. (1954). *Ann. Math. Stat.* **25**, 631–650.

FAGGIANI, R., LIPPERT, B. & LOCK, C. J. L. (1980). *Inorg. Chem.* **19**, 295–300.

GIACOVAZZO, C. (1977). *Acta Cryst.* **A33**, 50–54.

HAMMING, R. W. (1973). *Numerical Methods for Scientists and Engineers.* New York: McGraw-Hill.

HAUPTMAN, H. & KARLE, J. (1952). *Acta Cryst.* **5**, 48–59.

HAUPTMAN, H. & KARLE, J. (1953*a*). *Acta Cryst.* **6**, 136–141.

HAUTPMAN, H. & KARLE, J. (1953*b*). *Solution of the Phase Problem I. The Centrosymmetric Crystal.* ACA Monogr. No. 3. Pittsburgh: Polycrystal Book Service.

JOHNSON, N. L. & KOTZ, S. (1969). *Distributions in Statistics: Discrete Distributions.* Boston: Houghton Mifflin.

KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131–135.

KIEFER, J. E. & WEISS, G. H. (1984). *Proceedings on Random Walks and their Applications in the Physical and Biological Sciences*, edited by M. SHLESINGER & B. J. WEST. *Am. Inst. Phys. Conf. Proc.* **102**, 11–32.

KLUYVER, J. C. (1906). *K. Ned. Akad. Wet. Proc.* **8**, 341–350.

MAIN, P., FISKE, S. J., HULL, S. E., LESSINGER, L., GERMAIN, G., DECLERCQ, J. P. & WOOLFSON, M. M. (1980). *A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Univ. of York, England, and Louvain, Belgium.

PEARSON, K. (1905). *Nature (London)*, **72**, 294, 342.

RAYLEIGH, LORD (1880). *Philos. Mag.* **10**, 73–78.

SACHS, L. (1982). *Applied Statistics: A Handbook of Techniques.* New York: Springer.

SHMUELI, U. (1979). *Acta Cryst.* A**35**, 282–286.

SHMUELI, U. (1982*a*). *Acta Cryst.* A**38**, 362–371.

SHMUELI, U. (1982*b*). In *Crystallographic Statistics: Progress and Problems*, edited by S. RAMASESHAN, M. F. RICHARDSON & A. J. C. WILSON, pp. 53–82. Bangalore: Indian Academy of Sciences.

SHMUELI, U. & KALDOR, U. (1981). *Acta Cryst.* A**37**, 76–80.

SHMUELI, U. & KALDOR, U. (1983). *Acta Cryst.* A**39**, 619–621.

SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* A**37**, 342–353.

SHMUELI, U. & WILSON, A. J. C. (1982). In *Crystallographic Statistics: Progress and Problems*, edited by S. RAMASESHAN, M. F. RICHARDSON & A. J. C. WILSON, pp. 83-97. Bangalore: Indian Academy of Sciences.

SHMUELI, U. & WILSON, A. J. C. (1983). *Acta Cryst.* A**39**, 225–233.

SIM, G. A. (1958). *Acta Cryst.* **11**, 123–124.

SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography*, ch. 3. Oxford: Pergamon Press.

WEISS, G. H. (1983). *Am. Sci.* **71**, 65–71.

WEISS, G. H. & KIEFER, J. E. (1983). *J. Phys. A*, **16**, 489–495.

WILSON, A. J. C. (1942). *Nature (London)*, **150**, 151, 152.

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

WILSON, A. J. C. (1978). *Acta Cryst.* A**34**, 986–994.

WILSON, A. J. C. (1981). *Acta Cryst.* A**37**, 808–810.

WILSON, A. J. C. (1983). *Acta Cryst.* A**39**, 26–28.

Acta Cryst. (1984). A**40**, 660-662

The Determination of Absolute Structure. I. Some Experiences with the Rogers η Refinement

BY PETER G. JONES

Institut für Anorganische Chemie der Universität, Tammannstrasse 4, D-3400 Göttingen, Federal Republic of Germany

(*Received 4 April 1984; accepted 11 June 1984*)

Abstract

The η refinement of Rogers [*Acta Cryst.* (1981), A**37**, 734–741] has been applied to a wide range of noncentrosymmetric structures containing medium to strong anomalous scatterers; it has been shown to be an effective and robust method. The use of the general term 'absolute structure' (to signify a structure successfully distinguished from its inverse by, for example, analysis of anomalous scattering effects) is recommended.

Introduction

The absolute configuration/polar-axis direction (sometimes referred to as chirality/polarity) of a noncentrosymmetric crystal structure is often determined by least-squares refinement of both alternative models followed by a statistical comparison of R values using Hamilton's (1965) test. An attempt to provide a more reliable method was made by Rogers (1981), who suggested refining a parameter η as a factor multiplying all imaginary components f''_i of the anomalous dispersion terms of the atomic scattering factors; η should then adopt values of +1 or −1, corresponding to the correct or incorrect model, respectively. The least-squares estimate of the standard deviation of η may then be used as a measure of confidence, being assessed against the value 2 (the range of possible η values). Some criticisms of the method have been made by Flack (1983), who suggested the use of an alternative parameter x, derived from considerations of enantiomorphic twinning, to avoid certain technical problems of η refinement in cases where the structure is almost centrosymmetric. The purpose of this article is to present the results of some η refinements based on the experience of the author and colleagues in this institute.

All structures (see Table 1), except where otherwise stated, were measured with Mo $K\alpha$ radiation on a Stoe–Siemens four-circle diffractometer in profile-fitting mode (Clegg, 1981). The η refinement is part of the standard *SHELXTL* program system (Sheldrick, 1978).

0108-7673/84/060660-03$01.50