Proc. Indian Acad. Sci. (Chem. Sci.), Vol. 92, Numbers 4 & 5, August & October 1983, pp. 335–339.

The probability of large structure amplitudes: The space group $P\bar{1}$

A J C WILSON

Crystallographic Data Centre, University Chemical Laboratory, Cambridge CB2 1EW, England

Abstract. Until recently, expressions derived for the probability distribution of the magnitudes of structure amplitudes have been valid only for the range from zero up to two or three times the average magnitude. Recent progress in obtaining results valid for larger structure amplitudes (up to unitary structure factors approaching unity) is reviewed; the methods employed include Fourier-series representations, steepest-descents approximations, and computer simulations. Except for the last, the results are valid only for the space group $P\bar{1}$, though those for the space groups belonging to the point group $2/m$ and possibly some other centro-symmetric groups of low symmetry are expected to behave similarly.

Keywords. Crystallographic statistics; large structure amplitudes; probability distribution; statistics.

1. Introduction

The calculation of the probability of a structure amplitude having a particular value is a special case of the random-walk problem. If correlations can be neglected (Wilson 1981), for the space group $P1$ the problem is exactly that of the random walk in two dimensions, and for $P\bar{1}$ it is that of the projection on a line of a random walk in two dimensions. Most attention has been given to walks that do not get far from the starting point—in crystallographic terms structure amplitudes that do not exceed two or three times their average magnitude—but for several years I have tried unsuccessfully (Wilson 1980b) to obtain an estimate of the probability of structure amplitudes with magnitudes approaching the maximum possible. This maximum possible is, of course, the sum of the atomic scattering factors of all the atoms in the unit cell:

$$|F|_{\max} \equiv \Phi = f_1 + f_2 + f_3 + \ldots + f_N. \tag{1}$$

In direct methods of structure determination the so-called unitary scattering factors and unitary structure amplitudes are often used; these are the ordinary ones divided by Φ, and are thus convenient also in discussing the probability of large structure amplitudes. The unitary atomic scattering factor of the jth atom is

$$u_j = f_j/\Phi, \tag{2}$$

and the unitary structure amplitude is

$$U = F/\Phi. \tag{3}$$

'Large' unitary structure amplitudes are thus somewhat less than unity. Derivations of the mathematical expressions for $p(U)\mathrm{d}U$, the probability that a structure amplitude will have a unitary value between U and $U + \mathrm{d}U$, are complicated and it is more useful to try to review the present state of knowledge than to work through any derivation. So far I have been able to locate useful analytic results only for structures in the space group

$P\bar{1}$; it seems likely that other low-symmetry centrosymmetric groups will behave very similarly.

2. Exact expression

For the pure mathematician there is no problem. Kluyver (1906) was able to give $p(U)$ as an integral:

$$p(U)\mathrm{d}U = (2\pi)^{-1} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{\frac{1}{2}N} J_0(2u_j x) \right] \cos(Ux)\mathrm{d}x\mathrm{d}U, \tag{4}$$

where J_0 is the zero-order Bessel function. Kluyver himself recognized, however, that this expression is of little value for actual computation in any range of U.

3. Series representations

For small U, the exact expression (4) reduces to the central-limit Gaussian expression (Wilson 1949), which can be made the basis of a series expansion in terms of Hermite polynomials (see, for example, Karle and Hauptman 1953; Rogers and Wilson 1953; Bertaut 1955; Foster and Hargreaves 1963; Shmueli and Wilson 1981, 1982). The expansion, in either the Edgeworth or the Gram-Charlier form, satisfies the Cramér criterion for convergence, but in practice there may be unacceptable negative probabilities in the region of present interest.

Barakat (1973) pointed out that $p(U)$ vanishes outside the range -1 to $+1$, so that it can be represented by an ordinary Fourier series. His actual application was to a three-dimensional random walk (with polymer chains in mind), but Weiss and Kiefer (1983) have modified it for the two-dimensional case. Their expression is

$$p(U)\mathrm{d}U = \frac{1}{2}\left\{1 + 2 \sum_{m=1}^{\infty} C_m \cos \pi m U\right\}\mathrm{d}U, \tag{5}$$

where the Fourier coefficients C_m are given by

$$C_m = \prod_{j=1}^{\frac{1}{2}N} J_0(2\pi m u_j). \tag{6}$$

The Fourier series has the great advantage that the coefficients are readily obtained for any required number of terms, whereas the calculation of the moments required for the Hermite coefficients rapidly becomes more tedious as the order of the polynomial increases. For most space groups only five-term Hermite expansions are available at present, whereas there is no definite limit to the number that can be used in the Fourier series. In most cases tried twenty terms seem to be ample, there is no significant change in the values of $p(U)$ when the number is increased to forty (Shmueli 1983, private communication). A series similar to equation (5), but with Bessel functions instead of cosines, can be developed for $P1$ (Weiss and Kiefer 1983, private communication).

4. Cramér's limit theorem

The central-limit theorem is not the only limit theorem, and results due to Cramér

(1938) can be applied in the equiatomic centrosymmetric case (Wilson 1983). The result is

$$P(U)\mathrm{d}U = \left\{\frac{N}{4\pi\sigma^2}\right\}^{\frac{1}{2}} \{I_0(h)\}^{\frac{1}{2}N} \exp(-\tfrac{1}{2}NhU)\mathrm{d}U, \tag{7}$$

where h is the solution of

$$I_1(h)/I_0(h) = U, \tag{8}$$

the I's are modified Bessel functions, and

$$\sigma^2 = 1 - U/h - U^2. \tag{9}$$

For small U equation (7) reduces to the familiar Gaussian, and for large U to

$$p(U)\mathrm{d}U = \left\{\frac{N}{4\pi}\right\}^{\frac{1}{2}} \left\{\frac{e}{\pi}\right\}^{\frac{1}{4}N} (1-U)^{\frac{1}{4}N-1}\mathrm{d}U. \tag{10}$$

5. Steepest descents

Weiss and Kiefer (1983) use a more general approach, the method of steepest descents (Daniels 1954), and obtain equivalent expressions for the unequal-atom case:

$$p(U)\mathrm{d}U = \left\{\frac{N}{4\pi\sigma^2}\right\}^{\frac{1}{2}} \left\{\prod_{j=1}^{\frac{1}{2}N} I_0(Nu_jh)\right\} \exp(-\tfrac{1}{2}NhU)\mathrm{d}U, \tag{11}$$

where h is the solution of the equation

$$\sum_{j=1}^{\frac{1}{2}N} 2u_jI_1(Nu_jh)/I_0(Nu_jh) = U, \tag{12}$$

and

$$\sigma^2 = 2N \sum_{j=1}^{\frac{1}{2}N} u_j^2\{1 - I_1(Nu_jh)/Nu_jhI_0(Nu_jh) - [I_1(Nu_jh)/I_0(Nu_jh)]^2\} \tag{13}$$

$$= 2N \sum_{j=1}^{\frac{1}{2}N} u_j^2 - U/h - 2N \sum_{j=1}^{\frac{1}{2}N} [u_jI_1(Nu_jh)/I_0(Nu_jh)]^2. \tag{14}$$

Equations (11)–(14) reduce in an obvious fashion to (7)–(9) in the equal-atom case, in which all $u_j = N^{-1}$.

6. Resultants near maximum possible

An unequal-atom analogue of (10) is readily obtained from (11), but Weiss and Kiefer (1983) adopt a different approach that leads to a better result. In effect, they expand the Bessel functions in the exact expression (4) into asymptotic series, multiply out, and integrate term by term. The result is a series in ascending powers of $(1-U)$, of which the first term is

$$p(U)\mathrm{d}U = \left[(2\pi)^{\frac{1}{4}N}\,\Gamma(\tfrac{1}{4}N) \prod_{j=1}^{\frac{1}{2}N} (2u_j)^{\frac{1}{2}}\right]^{-1} (1-U)^{\frac{1}{4}N-1}\mathrm{d}U. \tag{15}$$

Weiss and Kiefer give two more terms explicitly and recurrence relations for higher terms.

Wilson (unpublished) has developed a calculation that suggests that $p(U)$ will behave like a power of $(1-U)$ for large U for any centrosymmetric space group.

7. Simulation

Shmueli (1982) has described a method of computer simulation that gives an 'experimental' distribution for any postulated structure, or rather for any postulated composition; the effects of correlation of the real atomic positions are not taken into account. None of the theoretical distributions described above take such correlations into account, so a comparison of the simulated distributions with the theoretical is comparing like with like, whereas a comparison with distributions actually observed for real crystals involves a further complication, the full effect of which is unknown (Wilson 1981). Briefly, the simulation method consists in calculating a suitable number of structure amplitudes (3000 is typical), with the usual atomic scattering factors but with hx, ky and lz or their combinations replaced by computer-generated 'random' numbers. The resulting $|F|$'s are sorted by size into suitable groups to give a histogram; examples are given by Shmueli (1982, 1983). The method is applicable to any space group.

The simulation method seems to be the easiest and most flexible method yet devised for obtaining ideal distributions for small and moderate structure amplitudes, but it has one aesthetic and one practical disadvantage. Aesthetically, numerical calculations tied to a specific atomic composition and a specific symmetry are less pleasing than closed or series expressions in which the effect of changing atomic composition or symmetry can be 'seen'. Practically, because of the statistical nature of the simulation, there are random fluctuations in the heights of the bars of the histogram. Wilson (1983, unpublished) has estimated the extent of these; for 3000 simulations and 30 histogram divisions the residual R between the actual histogram and the ideal distributions should be about 0·07. To obtain reliable estimates of the probability of really large (greater than four times the average, say) structure amplitudes would require a very large number of simulations.

8. Accuracy

The Barakat Fourier series (5) seems to reproduce the exact ideal distribution to several decimal places, the limiting factor being probably the approximations used in the computer calculation of the Bessel functions; at present its use is limited to the space group $P\bar{1}$. Weiss and Kiefer (1983 and private communications) and Shmueli (1982 and private communications) have made various numerical comparisons of the Hermite-Gaussian and steepest-descents expressions with the Fourier results for various total numbers of atoms (10 to 30) and various ratios of atomic scattering factors. As would be expected, the non-Fourier analytic expressions work best for equal-atom structures, and are poor if there is one pair of atoms much heavier than the rest, when the distribution becomes bimodal. The expression (15) is very good for large U. For smaller U, the steepest-descents results are usually, but not always, better than the Gaussian-Hermite. Simulation results agree with the Fourier within the expected R.

The preceding remarks refer to ideal distributions and different methods of approximating them. The distributions observed experimentally for real crystals will show statistical fluctuations of the same type as in those obtained by simulation. The size of the fluctuations from this source (there are several other sources) is inversely proportional to the square root of the number of reflexions observed, but 3000 is fairly typical. Only differences between an observed and a calculated histogram appreciably exceeding an R of 0·07, or between an observed and a simulated histogram appreciably exceeding 0·1, could, with reasonable certainty, be attributed to non-statistical factors (Wilson 1980a). It is hoped to include a discussion of tests of the statistical significance of apparent differences between two distributions, either one observed and one calculated, or two calculated, in a forthcoming paper (Shmueli *et al* 1983).

Acknowledgements

It is with great pleasure that I have accepted the invitation to contribute to this volume in honour of Professor S. Ramaseshan.

The work was begun before the author's retirement from the chair of crystallography at the University of Birmingham, and a preliminary version was presented at a meeting of the British Crystallographic Association in March 1983. The revision for publication was greatly aided by conversations with Professor U Shmueli and Dr G H Weiss at the University of Tel Aviv in April 1983.

References

Barakat R 1973 *J. Phys.* **A6** 796
Bertaut E F 1955 *Acta Crystallogr.* **8** 823
Cramér H 1938 Sur un nouveau théoréme-limite de la theorie des probabilitiés No. 736 in the series *Actualités scientifiques et industrielles* (Paris: Hermann)
Daniels H E 1954 *Ann. Math. Stat.* **25** 631
Foster F and Hargreaves A 1963 *Acta Crystallogr.* **16** 1124
Karle J and Hauptman H 1953 *Acta Crystallogr.* **6** 131
Kluyver J C 1906 *Kon. Akad. Wet. Amsterdam* **8** 341
Rogers D and Wilson A J C 1953 *Acta Crystallogr.* **6** 439
Shmueli U 1982 pp 53–82 in *Crystallographic statistics* (eds) S Ramaseshan, M F Richardson and A J C Wilson (Bangalore: Indian Academy of Sciences)
Shmueli U 1983 submitted for publication
Shmueli U, Weiss G H and Wilson A J C 1983 *In preparation*
Shmueli U and Wilson A J C 1981 *Acta Crystallogr.* **A37** 342
Shmueli U and Wilson A J C 1982 in *Conformation in biology* (eds) R Srinivasan and R H Sarma (New York: Adenine Press) pp. 383–388
Weiss G H and Kiefer J E 1983 *J. Phys.* **A16** 489
Wilson A J C 1949 *Acta Crystallogr.* **2** 318
Wilson A J C 1980a *Acta Crystallogr.* **A36** 937
Wilson A J C 1980b *Technometrics* **22** 629
Wilson A J C 1981 *Acta Crystallogr.* **A37** 808
Wilson A J C 1983 *Acta Crystallogr.* **A39** 26