Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 13 December 2006
Accepted 13 April 2007

The statistics of the highest *E* value

Grzegorz Chojnowski[a,b,c] and Matthias Bochtler[a,b]*

[a]International Institute of Molecular and Cell Biology, ul. Trojdena 4, 02-109 Warsaw, Poland, [b]Max-Planck-Institute for Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01309 Dresden, Germany, and [c]Department of Biophysics, Institute of Experimental Physics, Warsaw University, Zwirki i Wigury 93, 02-089 Warsaw, Poland. Correspondence e-mail: mbochtler@iimcb.gov.pl

In a previous publication, the Gumbel–Fisher–Tippett (GFT) extreme-value analysis has been applied to investigate the statistics of the intensity of the strongest reflection in a thin resolution shell. Here, a similar approach is applied to study the distribution, expectation value and standard deviation of the highest normalized structure-factor amplitude (*E* value). As before, acentric and centric reflections are treated separately, a random arrangement of scattering atoms is assumed, and *E*-value correlations are neglected. Under these assumptions, it is deduced that the highest *E* value is GFT distributed to a good approximation. Moreover, it is shown that the root of the expectation value of the highest 'normalized' intensity is not only an upper limit for the expectation value of the highest *E* value but also a very good estimate. Qualitatively, this can be attributed to the sharpness of the distribution of the highest *E* value. Although the formulas were derived with various simplifying assumptions and approximations, they turn out to be useful also for real small-molecule and protein crystal structures, for both thin and thick resolution shells. The only limitation is that low-resolution data (below 2.5 Å) have to be excluded from the analysis. These results have implications for the identification of outliers in experimental diffraction data.

1. Introduction

The distributions of the intensities and structure factors of 'typical' reflections in the X-ray diffraction pattern of a crystal are well known. They describe the distribution of normalized reflection intensities or normalized structure factors (unsigned *E* values) across reciprocal space (excluding some reflections in special positions). Alternatively, they describe the distributions of these values for a pre-picked reflection if the distribution of scattering atoms in the asymmetric unit is randomly varied. The statistics for acentric and for centric reflections are different. For a 'typical' reflection, the formulas for the *E*-value distributions are

$$f_a(x) = 2x\exp(-x^2), \qquad F_a(x) = 1 - \exp(-x^2), \quad (1)$$

$$f_c(x) = \left(\frac{2}{\pi}\right)^{1/2}\exp\left(-\frac{x^2}{2}\right), \qquad F_c(x) = \mathrm{erf}(x/2^{1/2}), \quad (2)$$

if x denotes the E value, f and F denote the non-cumulative and cumulative distributions and subscripts 'a' and 'c' distinguish acentric and centric reflections.

In a previous publication (Bochtler & Chojnowski, 2007), we have studied the intensity statistics of a very 'atypical' reflection, which was selected among n other unique reflections for having the strongest intensity (excluding reflections in special positions that are systematically strong). Using standard Gumbel–Fisher–Tippett (GFT) extreme-value analysis, we have derived analytical formulas that relate the expectation value μ and standard deviation σ of the highest reflection intensity J in a thin shell to the number of unique reflections n, from which it was selected. For convenience, reflection intensities were measured in units of the average reflection intensity in the thin resolution shell, which is equivalent to the transition from intensities to normalized intensities. With this convention and a separate treatment of acentric (subscript a) and centric (subscript c) reflections, the following formulas were obtained for the highest reflection intensity J:

$$\mu(J_a) = \ln n_a + \gamma, \quad \sigma(J_a) = \pi/6^{1/2}, \quad (3)$$

$$\mu(J_c) = 2(\ln n_c + \gamma) - \ln(\pi \ln n_c),$$
$$\sigma(J_c) = 2\pi/6^{1/2} - \pi/(6^{1/2}\ln n_c). \quad (4)$$

In (3) and (4), $\pi \simeq 3.142$ is Ludolf's number and $\gamma \simeq 0.577$ is the Euler–Mascheroni constant.

In the same work, we further showed that the highest reflection intensity was GFT-distributed and, based on this result, derived confidence intervals for the highest reflection intensity J:

$$P(\mu - 1.31\sigma \leq J \leq \mu + 1.87\sigma) = 90\% \quad (5)$$

$$P(\mu - 1.75\sigma \leq J \leq \mu + 3.68\sigma) = 99\%. \quad (6)$$

In these formulas, μ and σ refer to the expectation value and standard deviation of the highest reflection intensity but the same formulas would also be applicable to any other GFT-distributed variable. The boundaries of the confidence intervals were chosen to make the probabilities for large and small outliers equal. In the previous work, we limited ourselves to thin resolution shells to avoid complications from the systematic dependence of the average reflection intensity on resolution. The limitation of the thin resolution shells and the requirement for a sufficient number of reflections in the shell forced us to consider only protein crystals, which typically have large unit cells (Bochtler & Chojnowski, 2007).

In this work, we follow the approach of our prior work on the highest normalized reflection intensity to study the statistics of the highest E value of n unique reflections. In the *Theory* section of this work, we start from the well known E-value distributions for 'typical' reflections and pretend that the E values of 'typical' reflections are statistically independent. With this assumption, we can apply GFT extreme-value theory to the problem of the highest E value, demonstrate that this value is GFT-distributed and derive separate analytical expressions for its expectation value μ and standard deviation σ in the acentric and centric cases. As the analytical results are based on simplifying assumptions and approximations, their merits are checked by extensive numerical tests. Technical details of these tests are summarized in the section *Materials and methods*. In the section *Tests with simulated data*, we show that the analytical formulas are good predictors of the highest E value for crystal structures with randomly distributed scatterers, and in the section *Tests with real data*, we demonstrate that this applies also to real crystal structures if very low resolution data are excluded from the analysis. In contrast to our prior work on the highest normalized reflection intensity, we do not limit ourselves to thin resolution shells, and we perform numerical tests for small-molecule crystal structures from the Cambridge Structural Database (CSD) in addition to the tests for protein crystal structures from the Protein Data Bank (PDB). In the *Discussion* section, we present a short-cut to the leading order estimates in this work, and we demonstrate that the full expressions are consistent with our prior results on the highest normalized structure factor. Appendix *A* contains detailed calculations that were omitted from the *Theory* section for clarity.

Our results have implications for outlier rejection in experimentally determined diffraction data. In principle, the overall probability for erroneous rejections should be constant but, in the usual procedure that treats reflections individually, the overall probability for false rejections grows with the number of unique reflections in the diffraction pattern. Our formulas, which indicate a very weak dependence of the highest E value on the number of unique reflections, show that the error is small and indicate how to correct for it if desired.

2. Theory

2.1. Notations and conventions

Throughout this work, E, P, Q stand for E values and I, J, K for normalized intensities when these quantities are treated as (pseudo)random variables. The corresponding lower-case letters denote actual values, except that x is used instead of e to avoid confusion with Euler's number. E and I describe a 'typical' reflection, P and J the strongest reflection, Q and K the strongest reflection after rescaling to expectation value 0 and standard deviation 1. F, G, H denote cumulative probability distributions and f, g, h the corresponding non-cumulative distributions. The letters F and f describe the E-value distribution of a 'typical reflection'. G and g represent the cumulative and non-cumulative Gumbel distribution (Gumbel, 1958):

$$G(x) = \exp[-\exp(-x)] \quad (7)$$

$$g(x) = \exp(-x)\exp[-\exp(-x)]. \quad (8)$$

The term Gumbel distribution is reserved here for this special form of the GFT distribution with expectation value γ and standard deviation $\pi/6^{1/2}$. H and h stand for the cumulative and non-cumulative GFT distribution after rescaling to expectation value 0 and standard deviation 1:

$$H(x) = \exp\Big[-\exp\Big(-\frac{\pi}{6^{1/2}}x - \gamma\Big)\Big] \quad (9)$$

$$h(x) = \frac{\pi}{6^{1/2}}\exp\Big(-\frac{\pi}{6^{1/2}}x - \gamma\Big)\exp\Big[-\exp\Big(-\frac{\pi}{6^{1/2}}x - \gamma\Big)\Big]. \quad (10)$$

Subscripts c and a distinguish centric and acentric reflections where necessary. Throughout this work, n stands for the number of unique reflections from which the highest E value was selected.

2.2. Predictions

To proceed with the analytical treatment, we pretend that E values are statistically independent non-correlated variables. This simplifying assumption ignores the well known correlations between (complex) normalized structure factors which are the basis of direct methods (Cochran & Woolfson, 1955) and can be justified only by the success of the formulas that are derived from it. We note that the highest E value from reflections is below a threshold if the E values of all reflections are below the threshold and *vice versa*. With the assumption of statistical independence of the reflections, the cumulative distribution of the highest E value is therefore related to the cumulative distribution of the E value of a typical reflection by a simple power law with the number of unique reflections in the exponent.

$$\Pr(P \leq x) = [F(x)]^n \quad (11)$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\Pr(P \leq x) = n[F(x)]^{n-1}F'(x) \quad (12)$$

$$\mu(P) = \int_0^\infty \mathrm{d}x\, x \frac{\mathrm{d}}{\mathrm{d}x} \Pr(P \le x) = \int_0^\infty \mathrm{d}x\, xn[F(x)]^{n-1} F'(x)$$
$$= \int_0^\infty \mathrm{d}x\, xn[F(x)]^{n-1} f(x) \quad (13)$$

$$\mu(P^2) = \int_0^\infty \mathrm{d}x\, x^2 \frac{\mathrm{d}}{\mathrm{d}x} \Pr(P \le x) = \int_0^\infty \mathrm{d}x\, x^2 n[F(x)]^{n-1} F'(x)$$
$$= \int_0^\infty \mathrm{d}x\, x^2 n[F(x)]^{n-1} f(x) \quad (14)$$

$$\sigma(P) = [\mu(P^2) - \mu^2(P)]^{1/2} \quad (15)$$

The distribution of the largest E value depends only on the (exponential) tail of the distribution for a 'typical' E value. Therefore, the cumulative distribution of the E value of a typical reflection is conveniently written as

$$F(x) = 1 - \exp[-u(x)]. \quad (16)$$

In this expression, it is understood that the dependence of u on its argument x is polynomial and not exponential. As usual in GFT extreme-value theory, approximate expressions for equations (11) to (15) can be obtained by a Taylor expansion of u around $u^{-1}(\ln n)$. As the cumulative distributions F_a and F_c for acentric and centric reflections are different, the calculations have to be carried out separately for the two groups of reflections. The detailed calculations are presented in Appendix *A*, here we present only a summary of the results.

In the *acentric* case, Taylor expansion of u to the first order yields

$$\Pr(P_\mathrm{a} \le x) = G(2[\ln n_\mathrm{a}]^{1/2}\{x - [\ln n_\mathrm{a}]^{1/2}\}). \quad (17)$$

In words, equation (17) expresses that the highest E value is GFT distributed if the quadratic term in the expansion for u is neglected. If the quadratic term is taken into account, a different expression is obtained:

$$\Pr(P_\mathrm{a} \le x) = G(x^2 - \ln n_\mathrm{a}). \quad (18)$$

Note that equation (18) is an improvement over equation (17) only for some arguments. For $x \to -\infty$, (18) predicts $\Pr(P \le x) \to 1$, which is clearly incorrect, whereas (17) predicts $\Pr(P \le x) \to 0$, as it should be. Therefore, analytical expressions for $\mu(P_\mathrm{a})$ and $\sigma(P_\mathrm{a})$ were only obtained for the Taylor expansion up to the linear term. Detailed calculations that are presented in Appendix *A* lead to

$$\mu(P_\mathrm{a}) = (\ln n_\mathrm{a})^{1/2}\left(1 + \frac{\gamma}{2\ln n_\mathrm{a}}\right) \quad (19)$$

$$\sigma(P_\mathrm{a}) = \frac{\pi}{6^{1/2}} \frac{1}{2(\ln n_\mathrm{a})^{1/2}}. \quad (20)$$

In the *centric* case, Taylor expansion of u leads to the following expression for the highest E value:

$$\Pr(P \le x) = G(a_{n_\mathrm{c}}(x - b_{n_\mathrm{c}}) + \tfrac{1}{2}(x - b_{n_\mathrm{c}})^2). \quad (21)$$

In this expression,

$$b_{n_\mathrm{c}} \approx (2\ln n_\mathrm{c})^{1/2}\left(1 - \frac{\ln(\pi \ln n_\mathrm{c})}{4\ln n_\mathrm{c}}\right), \quad a_{n_\mathrm{c}} = b_{n_\mathrm{c}} + \frac{1}{(2\ln n_\mathrm{c})^{1/2}}. \quad (22)$$

As in the acentric case, the quadratic term is an improvement only for some arguments. If only the linear term is retained, further calculations and approximations that are presented in Appendix *A* lead to

$$\mu(P_\mathrm{c}) = (2\ln n_\mathrm{c})^{1/2}\left(1 - \frac{\ln(\pi \ln n_\mathrm{c}) - 2\gamma}{4\ln n_\mathrm{c}}\right) \quad (23)$$

$$\sigma(P_\mathrm{c}) = \frac{\pi}{6^{1/2}} \frac{1}{(2\ln n_\mathrm{c})^{1/2}}\left(1 + \frac{\ln(\pi \ln n_\mathrm{c}) - 2}{4\ln n_\mathrm{c}}\right). \quad (24)$$

Note that for fixed n the leading terms in (23) and (24) are a factor $2^{1/2}$ larger than the leading terms in equations (19) and (20) which are applicable in the acentric case. The factor $2^{1/2}$ is to be compared with the factor 2 in the corresponding expressions for the normalized intensities (Bochtler & Chojnowski, 2007).

3. Materials and methods

3.1. Numerical methods

Utility programs were implemented in the C++ language with extensive use of routines from the GNU scientific library (Galassi *et al.*, 2005) and the Clipper library (Cowtan, 2003). Random atom positions (compatible with the symmetry of the space group) were generated with the 'Mersenne twister' uniform random-number simulator of the GNU scientific library (Matsumoto & Nishimura, 1998). Constraints on inter-atom distances, such as bond lengths or exclusion volumes to avoid atom overlap, were not applied. Uncorrelated random data with a distribution according to equation (1) were generated by taking the square root of the sum of squares of two normally distributed variables with expectation value 0 and standard deviation $1/2^{1/2}$. Uncorrelated data with a distribution according to equation (2) were generated from a normally distributed variable with expectation value 0 and standard deviation 1 by taking the modulus. Numerical integrations were performed by a Gauss–Kronrod 21-point adaptive integration method. Infinite integrals were extended to a finite boundary, which was chosen sufficiently large so that the cut-off did not affect accuracy. Simulation errors were either calculated according to the standard formulas for the sample variance distribution (Eric W. Weisstein, *Sample Variance Distribution*, *WolframMathWorld – The Web's Most Extensive Mathematics Resource*. http://mathworld.wolfram.com/) or obtained by splitting the data into 10 separate bins. All graphs were prepared with the *GRACE* software (http://plasma-gate.weizmann.ac.il/Grace/).

3.2. Simulations, small unit cells

Simulations with small unit cells were run with parameters that are typical for small-molecule structures. We placed $0.068V$ [Å^3] C atoms with B factor 10 Å^2 in a unit cell of volume V, which is equivalent to an average mass density of

1.35 g cm^{-3}. Structure factors were calculated by the CCP4 program *SFALL* (Agarwal, 1978) and converted to E values by the *DREAR* program (Blessing *et al.*, 1998; Blessing & Smith, 1999), which uses overlapping resolution bins for normalization and is therefore optimal for handling diffraction data of small-molecule crystals, which typically have far fewer reflections than diffraction data of protein crystals.

3.3. Simulations, large unit cells

Simulations with large unit cells were run with parameters that are typical for protein crystals. We placed $0.0281V$ [Å^3] C atoms with B factor 0 in a unit cell of volume V, which results in a Matthews coefficient of 3 Å^3 Da^{-1}. Structure factors were calculated by the CCP4 program *SFALL* (Agarwal, 1978). A bulk solvent correction was not applied. The conversion of structure factors to E values was carried out with the *ECALC* program (Collaborative Computational Project, Number 4, 1994), which was designed to handle diffraction data from protein crystals.

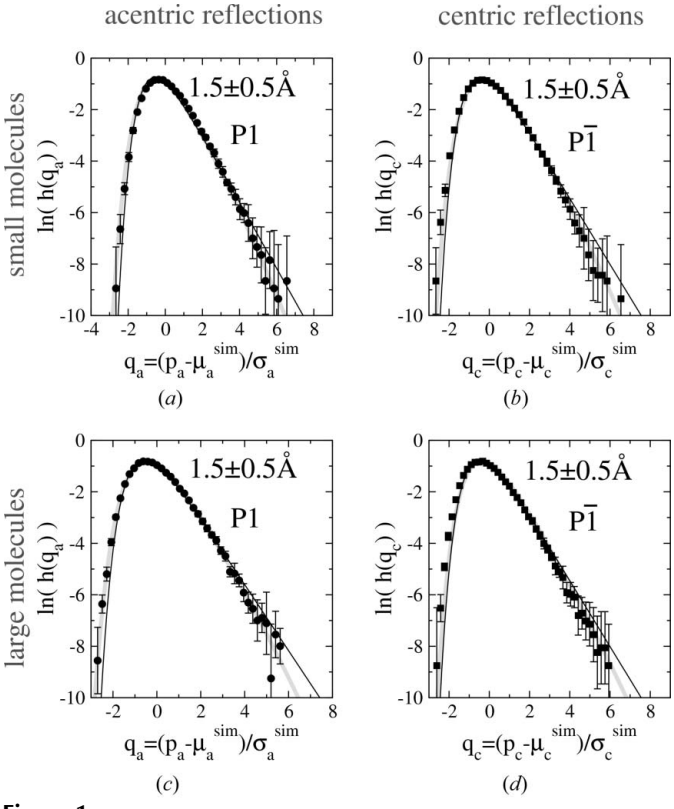

Figure 1
Distribution of the highest E value in the resolution shell 1.5 ± 0.5 Å for 100000 different configurations of scattering atoms in (*a*), (*c*) space group $P1$ and (*b*), (*d*) space group $P\bar{1}$. The top (*a*), (*b*) and bottom (*c*), (*d*) panels were obtained with parameters that are typical for small-molecule and protein crystal structures, respectively. In order to compare only the shape of the distributions, all distributions were analytically or numerically rescaled to expectation value 0 and standard deviation 1. The black dots show simulation results. The black line indicates the GFT distribution h of equation (10). This non-cumulative distribution is the appropriately rescaled derivative of the distribution of equation (17), which results from neglecting the quadratic term in the Taylor expansion of u. The grey line shows the prediction of equation (18), which takes the quadratic term in the Taylor expansion of u into account. Numerical integration was used to rescale the distribution in equation (18) to expectation value 0 and standard deviation 1.

3.4. Real crystals, small molecules

We downloaded all structures from the CSD (Allen, 2002) (version 5.27, November 2005) that had R factor better than 0.05, were free of disorder or errors according to the annotation, and had an estimated C—C bond standard deviation below 0.005 Å. Structures that were determined by powder diffraction were excluded from the set. Using these criteria, 35387 structures were obtained. No effort was made to remove duplicate or highly similar structures from the set. As temperature factors are not deposited in the CSD, we arbitrarily set them to 10 Å^2 for all atoms.

3.5. Real crystals, protein molecules

Structures that had been solved at 1.5 Å resolution or better were downloaded from the PDB (Berman *et al.*, 2000) (release date 18 April 2006). Duplicates or near duplicates (cut-off 90% identity) and nucleic acid structures were removed from the set. We also removed all structures from the set that had pseudo-origin peaks in the Patterson map that reached 40% or more of the height of the origin peak (PDB identities 1dy5, 1ob6, 1xy1, 1m2d, 1vrz, 2bfi, 1w5u, 1m1n, 1hqj, 1m70, 1k6f, 1t6u, 1av2, 2f46, 1pp0, 1i88, 2a8y, 1o6v, 1p4o, 1wzb or 1.7% of all structures in the set) and three very regular structures (1t8z, 1jl0, 1k5c).

4. Tests with simulated data

The theory relies on several assumptions and approximations. In a first series of tests, we assessed the merit of the approximations with simulated crystal structures with random atom configurations.

4.1. Distribution of the highest E value

As a first step, we focused on the distribution of the highest E value and compared the predictions from the expansion to the linear and quadratic terms with each other and with simulations for 100000 random atom configurations in space groups $P1$ and $P\bar{1}$. The simulations were run with parameters that are typical for small-molecule structures (unit-cell volume 1300 Å^3, 0.068 non-H atoms per Å^3) and then separately with parameters that are typical for protein crystal structures (unit-cell volume 140000 Å^3, 0.0281 non-H atoms per Å^3). In order to compare the shapes of the distributions independently of the scale parameters that determine expectation value and standard deviation, all distributions were rescaled to expectation value 0 and standard deviation 1. The results for the small and large unit cells are presented in Figs. 1(*a*), (*b*) and 1(*c*), (*d*), respectively. The conclusion is that in all cases the GFT distribution approximates very well the real frequencies of the highest E values. The quadratic term in the expansion of u has only a minor effect, except perhaps for large arguments where it slightly improves the agreement between predictions and simulation results (Fig. 1).

4.2. Expectation value and standard deviation of the highest *E* value

Simulations were run in several space groups, but only the (representative) calculations in space group $P2_12_12_1$ are presented here. The atom density was kept fixed and the unit-cell size was varied. As the volume of the reciprocal unit cell is inversely proportional to the volume of the direct-space unit cell, a change in unit-cell size alters the number of scattering atoms and also the number of reflections in each thin resolution shell. For each unit-cell size, 10000 random atom configurations were generated. A comparison of the simulation results with the predictions reveals generally good agreement, both for the simulation with small-molecule parameters and for the simulation with protein parameters. In spite of the good overall agreement, there are small systematic deviations, which could be due either to neglected correlations in the diffraction data or to other approximations that were required to obtain analytically tractable expressions. To distinguish between the two possibilities, we ran additional simulations, in which we replaced the diffraction data that were calculated by the Fourier transform with identically distributed but uncorrelated random numbers (see *Materials and methods*). Moreover, expressions (13) and (14) were integrated numerically. The numerical integration results rely only on the validity of the underlying *E*-value distributions for typical reflections and on the assumption that *E*-value correlations can be ignored. We expected that the simulation results with uncorrelated 'pseudo' *E* values and the numerical integration results should be fully consistent, but might differ both from the simulations with random atom configurations and from the results of the analytical treatment. This is indeed the case and sheds light on the causes of residual discrepancies between the analytical formulas and the derivations.

In the simulations with small-molecule parameters, both the expectation values and standard deviations come out slightly lower from the simulations than from the analytical treatment. In the case of the expectation values, the simulations with 'pseudo' *E* values and the results of numerical integration agree almost perfectly with the analytical treatment, suggesting that the remaining discrepancy from the simulation with random atom conformations is either due to shortcomings of the underlying distribution for the *E* values of 'typical' reflections or to the neglected correlations in the diffraction data. In the case of the standard deviations, the situation is more complex because the simulations with 'pseudo' *E* values from the random number generator and numerical integration yield results that are intermediate between the simulation results with random atom configurations and the predictions. For centric reflections, the leading term in equation (24) predicts the standard deviation of the highest *E* value better than the full expression with the correction factor in the brackets. The reasons for this are not clear, particularly because the analogous correction term in

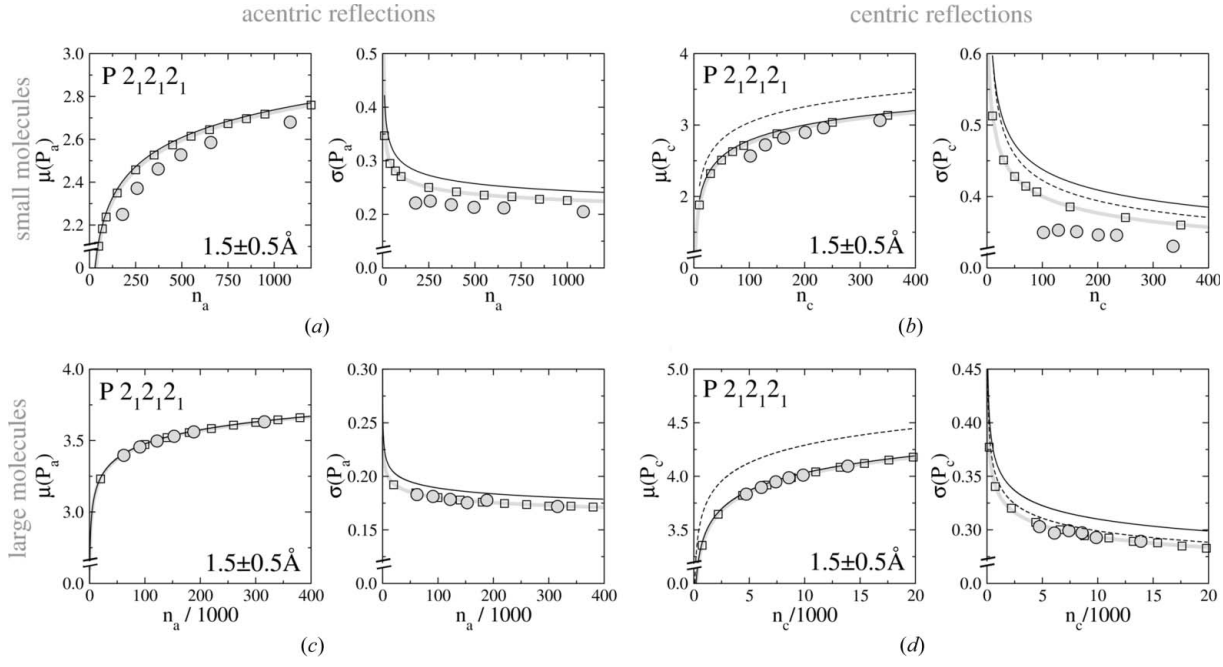

Figure 2
Expectation value μ and standard deviation σ of the highest E value P for $P2_12_12_1$ crystal structures with random atom conformations for (*a*), (*c*) the acentric case and (*b*), (*d*) the centric case. The top (*a*), (*b*) and bottom (*c*), (*d*) panels were obtained with parameters that are typical for small-molecule and protein crystal structures, respectively. n is the number of unique reflections from which the highest E value was selected. Simulation results with random atom conformations are indicated by circles. For comparison, simulation results with uncorrelated random numbers that are distributed according to the E-value distribution for a 'typical reflection' [equations (1) and (2)] are indicated by open boxes. Grey lines indicate the values for μ and σ that result from numerical integration according to equations (13) and (14) [with cumulative distributions F taken from equations (1) and (2)]. The black lines present the predictions of the analytical formulas. These are equations (19) and (20) in the acentric case and equations (23) and (24) in the centric case. The dashed lines in (*b*) and (*d*) show the predictions of equations (23) and (24) without the correction factors in brackets.

equation (23) for the expectation value of the highest centric E value strongly improves the agreement with the simulations (Figs. 2*a* and 2*b*).

In the simulation parameters that are typical for protein crystals (Figs. 2*c* and 2*d*), qualitatively similar results were obtained, but the discrepancies are generally smaller. Within the error of our simulations, we cannot detect any discrepancy between the predictions and simulation results for the expectation value of the highest E value, for both acentric and centric reflections, provided the full expression in equation (23) including the correction term in brackets is used for the predictions. Standard deviations come out essentially identical from simulations with random atom configurations, with 'pseudo' E values and from numerical integration. The predictions for the standard deviation from the analytical formulas are slightly too large, but the discrepancies are generally smaller than with small-molecule parameters. As before, omission of the correction term in brackets in equation (23) improves the agreement between predictions and simulations (Figs. 2*c* and 2*d*).

5. Tests with real data

5.1. Expectation value and standard deviation of the highest *E* value

In contrast to simulated data, which can be generated in any desired quantity, there is typically only one structure with exactly n unique reflections in a given resolution shell. Therefore, it was necessary to cluster real structures into bins with similar n. Bins for n values were chosen as a compromise between the conflicting requirements for large bins to collect sufficient statistics for $\mu(P)$ and $\sigma(P)$ and for small bins to keep the spread of n low. Throughout, calculated diffraction data were used instead of experimental data. In the case of the small-molecule structures, this was necessary because experimental structure factors are not deposited in the CSD. Temperature factors, which are also not deposited in the CSD, were arbitrarily set to 10 Å^2 for all atoms. In the case of protein structures from the PDB, experimental diffraction data were available in some cases but were not used, to exclude the influence of measurement errors. The results of this analysis are presented in Fig. 3 and show that the non-random features of real crystal structures have no effect or only a minor effect on the statistics of the highest E value for the resolution shell 1.5 ± 0.5 Å. In the acentric and centric cases, the predictions for the expectation value are excellent, but the standard deviations slightly underestimate the actual values for real data. As predicted from our prior work on the highest normalized intensity, predictions break down if very low resolution data are included (Fig. 3).

5.2. Confidence interval for the highest *E* value

As the highest E value is to a very good approximation GFT distributed, the results for its expectation value and standard deviation can be rephrased in terms of confidence intervals for the highest E value. The confidence interval can be uniquely

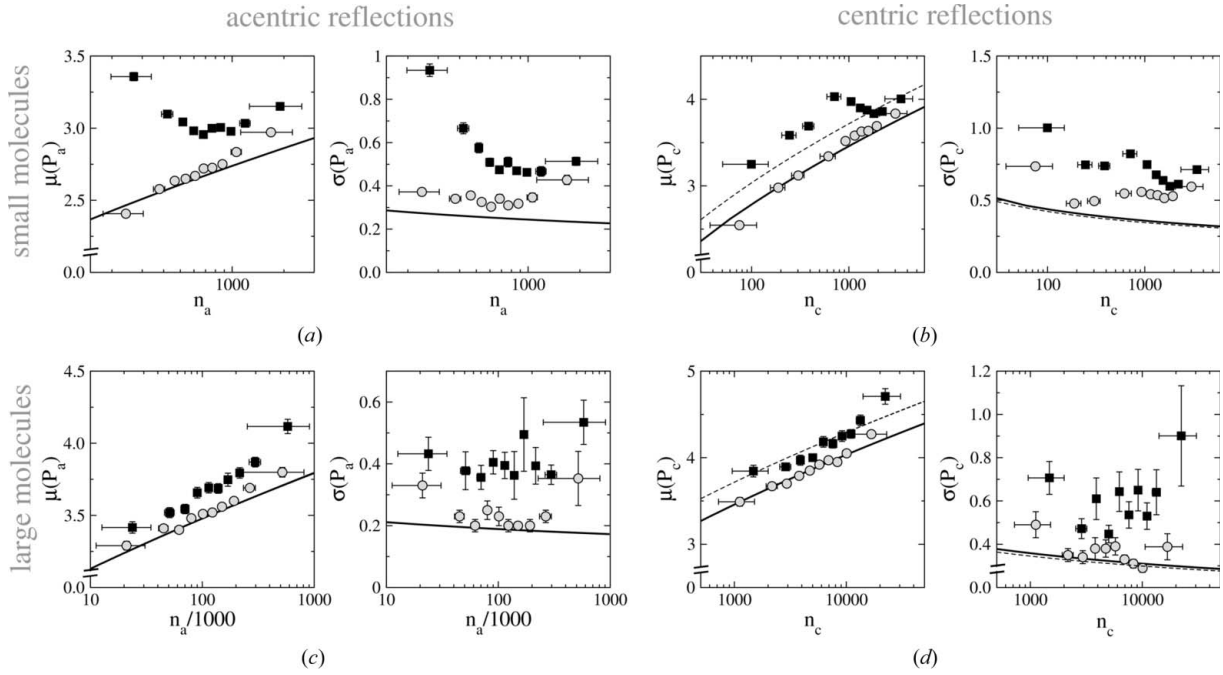

Figure 3
Expectation value μ and standard deviation σ of the highest E value for real crystal structures from (*a*), (*b*) the CSD and (*c*), (*d*) the PDB and for (*a*), (*c*) the acentric case and (*b*), (*d*) the centric case. Bins for n values were chosen as a compromise between the conflicting requirements for large bins to collect sufficient statistics for $\mu(P)$ and $\sigma(P)$ and for small bins to keep the spread of n low. Circles indicate results for the 1.5 ± 0.5 Å shell. Black squares are for the resolution range from 12.0 to 1.0 Å. The black lines present the predictions of the analytical formulas. These are equations (19) and (20) in the acentric case and equations (23) and (24) in the centric case. The dashed lines in (*b*) and (*d*) show the predictions of equations (23) and (24) without the correction factors in brackets.

defined by the additional requirement that the number of small and large outliers should be equal. For the 90 and 99% confidence intervals, this leads directly to equations (5) and (6), which were derived in our previous work on normalized intensities. A qualitative test for the 90% confidence interval is presented in the scatter plots of Fig. 4, which show the actual values of the highest E value and the theoretical 90 and 99% confidence intervals. Qualitatively, the agreement seems good but, quantitatively, there are too many low and high outliers. Their actual number is quantified for the 99% confidence interval for the 1.5 ± 0.5 Å shell and for thinner shells centered at the same resolution and is presented in the quantitative outlier statistics. Except for very thin shells, we find more outliers than the predicted 5%, probably in part because the predictions for $\sigma(P)$ underestimate the spread of P for real crystal structures (Fig. 4).

6. Discussion

6.1. A leading order estimate for $\mu(P)$

The derivations which have been presented so far rely heavily on extreme-value statistics. However, the key results, the estimates for $\mu(P_a)$ and $\mu(P_c)$ in equations (19) and (23) contain the Euler–Mascheroni constant γ, which is typical for extreme value statistics, only in the $O(1/[\ln n^{1/2}])$ terms. This suggested that the leading-order estimates for $\mu(P_a)$ and $\mu(P_c)$ should be derivable without reference to the mathematics of extreme-value statistics. It is plausible that the number of reflections that are larger than the expectation value of the highest reflection should be $O(1)$. For simplicity, take this quantity to be 1 and consider the consequences according to equations (1) and (2).

$$n_a \exp[-\mu^2(P_a)] = 1, \quad n_c\left\{1 - \operatorname{erf}\left[\frac{\mu(P_c)}{2^{1/2}}\right]\right\} = 1. \qquad (25)$$

The equations can be used to solve for $\mu(P_a)$ and $\mu(P_c)$, in the centric case using the expansion of the error function for large arguments. The result is

$$\mu(P_a) = (\ln n_a)^{1/2}, \quad \mu(P_c) = (2\ln n_c)^{1/2}\left[1 - \frac{\ln(\pi \ln n_c)}{4\ln n_c} + \ldots\right]. \qquad (26)$$

The expressions for $\mu(P_a)$ and for $\mu(P_c)$ require only corrections of order $O(1/[\ln n]^{1/2})$ to match the more exact expressions (19) and (23).

6.2. A better estimate for $\mu(P)$

Improved estimates of $\mu(P)$ can be derived from our previous results about the statistics of the highest normalized intensity. From $J = P^2$ and $\mu^2(Z) \leq \mu(Z^2)$ for any random variable, it follows that $\mu^2(P) \leq \mu(J)$. Therefore, we initially expected that the expectation values for the strongest normalized intensities, $\mu(J_a)$ and $\mu(J_c)$, would merely act as upper bounds for the squares of $\mu(P_a)$ and $\mu(P_c)$, respectively. However, from the expressions for $\mu(J_a)$ and for $\mu(J_c)$ in equations (3) and (4), one readily obtains

$$[\mu(J_a)]^{1/2} = (\ln n_a)^{1/2}\left(1 + \frac{\gamma}{\ln n_a}\right)^{1/2} = (\ln n_a)^{1/2}\left(1 + \frac{\gamma}{2\ln n_a} + \ldots\right) \qquad (27)$$

$$[\mu(J_c)]^{1/2} = (2\ln n_c)^{1/2}\left(1 - \frac{\ln(\pi \ln n_c) - 2\gamma}{2\ln n_c}\right)^{1/2} = (2\ln n_c)^{1/2}\left(1 - \frac{\ln(\pi \ln n_c) - 2\gamma}{4\ln n_c} + \ldots\right). \qquad (28)$$

These expressions match the expressions for $\mu(P_a)$ and $\mu(P_c)$, also for the terms of $O(1/[\ln n]^{1/2})$, which are missing from the simpler estimates of equation (26). Qualitatively, $[\mu(J_a)]^{1/2}$ and $[\mu(J_c)]^{1/2}$ are excellent approximations for $\mu(P_a)$ and $\mu(P_c)$ because the distributions of P_a and P_c peak sharply around their expectation values. Quantitatively, $\mu(P^2) = \mu^2(P) + \sigma^2(P)$ and therefore

$$\mu(P) = \frac{[\mu(J)]^{1/2}}{\left\{1 + \left[\frac{\sigma(P)}{\mu(P)}\right]^2\right\}^{1/2}} = [\mu(J)]^{1/2}\left[1 + O\left(\frac{1}{(\ln n)^2}\right)\right]. \qquad (29)$$

6.3. Estimating $\sigma(P)$

The leading-order estimates for $\sigma(P_a)$ and for $\sigma(P_c)$ can also be derived by an alternative route, which is easier to remember but lacks a rigorous justification. As E values are defined here as positive quantities (unsigned E values), the reflection with the highest E value p is also the reflection with the highest normalized intensity j.

$$j_a = p_a^2, \quad \mathrm{d}j_a = 2p_a \mathrm{d}p_a, \quad j_c = p_c^2, \quad \mathrm{d}j_c = 2p_c \mathrm{d}p_c. \qquad (30)$$

The next step is to identify j's and p's with expectation values and the differentials with standard deviations. So

$$j_a = \mu(J_a), \quad p_a = \mu(P_a), \quad \mathrm{d}j_a = \sigma(J_a), \quad \mathrm{d}p_a = \sigma(P_a), \qquad (31)$$

$$j_a = \mu(J_c), \quad p_c = \mu(P_c), \quad \mathrm{d}j_c = \sigma(J_c), \quad \mathrm{d}p_c = \sigma(P_c). \qquad (32)$$

Using only leading-order estimates from equations (3), (4), (19) and (23), we readily find that

$$\sigma(P_a) = \frac{\pi}{6^{1/2}}\frac{1}{2(\ln n_a)^{1/2}}, \quad \sigma(P_c) = \frac{\pi}{6^{1/2}}\frac{1}{(2\ln n_c)^{1/2}}, \qquad (33)$$

which agrees in leading order with the more accurate estimates of equations (20) and (24).

APPENDIX A

We begin with the Taylor expansion of u near the argument $u^{-1}(\ln n)$, where $u(u^{-1}(x)) = u^{-1}(u(x)) = x$,

$$u(x) = \ln n + a_n(x - b_n) + \tfrac{1}{2}A_n(x - b_n)^2 + \ldots. \qquad (34)$$

In this expansion,

$$b_n = u^{-1}(\ln n), \quad a_n = u'(b_n), \quad A_n = u''(b_n). \qquad (35)$$

Therefore,

$$\Pr(P \le x) = [F(x)]^n = \left(1 - \frac{1}{n}\left\{\exp[-a_n(x-b_n) - \tfrac{1}{2}A_n(x-b_n)^2]\right\}\right)^n, \quad (36)$$

which in turn can be approximated as

$$\Pr(P \le x) = G\left(a_n(x-b_n) + \tfrac{1}{2}A_n(x-b_n)^2\right), \quad (37)$$

where G is the well known Gumbel distribution of equation (7). It follows that the highest E value is Gumbel distributed, provided that the quadratic term can be neglected. Standard GFT theory tells us that a Gumbel-distributed variable has expectation value γ and standard deviation $\pi/6^{1/2}$. From this, it follows straightforwardly that

$$\mu(P) = b_n + \frac{\gamma}{a_n}, \quad \sigma(P) = \frac{\pi}{6^{1/2}}\frac{1}{a_n} \quad (38)$$

if the quadratic term in the argument of the Gumbel function can be neglected. A full evaluation of equations (37) and (38) requires an explicit expression for $u(x)$. This function depends on the E-value distribution for a 'typical' reflection and is therefore different for acentric and centric reflections.

A1. Acentric reflections

For *acentric* reflections (Giacovazzo *et al.*, 2002),

$$F_a(x) = 1 - \exp(-x^2) = 1 - \exp[-u_a(x)] \quad (39)$$

immediately implies that

$$u_a(x) = x^2. \quad (40)$$

Using the definitions in (35), it follows that

$$b_{n_a} = (\ln n_a)^{1/2} \quad (41)$$

$$a_{n_a} = 2(\ln n_a)^{1/2} \quad (42)$$

$$A_{n_a} = 2. \quad (43)$$

Inserting this into equation (37) yields equations (17) and (18) depending on whether or not the quadratic term is neglected. Combining equations (41) and (42) with equation (38) leads to equations (19) and (20).

A2. Centric reflections

For *centric* reflections (Giacovazzo *et al.*, 2002),

$$F_c(x) = \mathrm{erf}(x/2^{1/2}) = 1 - \exp[-u_c(x)]. \quad (44)$$

To solve for $u_c(x)$, it is necessary to expand the error function for large arguments.

$$\mathrm{erf}(x) = 1 - \frac{\exp(-x^2)}{x\pi^{1/2}}\left(1 - \frac{1}{2x^2} + \ldots\right) \quad (45)$$

$$u_c(x) = \frac{x^2}{2} + \ln x + \ln\left(\frac{\pi}{2}\right)^{1/2} + \frac{1}{x^2} + \ldots. \quad (46)$$

The calculation of a_n, b_n and A_n is more complicated in the centric case than in the acentric case because equation (46) cannot be inverted straightforwardly. If only the leading term is taken into account, one obtains $u_c^{-1}(\ln n_c) \approx (2\ln n_c)^{1/2}$ as a

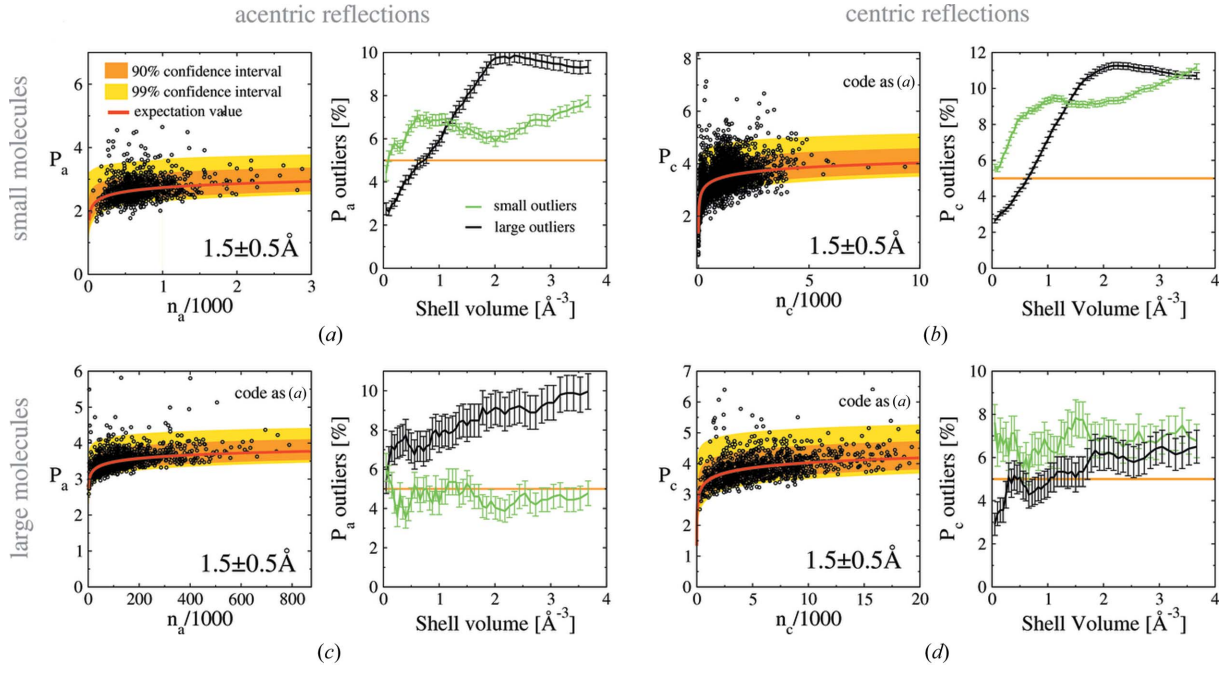

Figure 4
Confidence intervals for the highest E value P for real crystal structures from (*a*), (*b*) the CSD (10% of the structures were randomly picked) and (*c*), (*d*) the PDB and for (*a*), (*c*) the acentric case and (*b*), (*d*) the centric case. The scatter plots were calculated for the resolution range 1.5 ± 0.5 Å. Red lines indicate the expectation values of P according to equations (19) and (23). The orange and yellow regions mark the predicted 90 and 99% confidence intervals. According to the predictions, there should be 5% low and 5% high outliers that do not fall into the orange 90% confidence interval. The actual percentage of outliers was quantified for the 1.5 ± 0.5 Å shell and subsequently also for thinner shells centered at 1.5 Å resolution. In the diagrams that show the percentage of outliers *versus* shell volume, green lines represent low P outliers and black lines represent high P outliers.

first approximation. To improve upon this crude approximation, one can set $u_c^{-1}(\ln n_c) = (2 \ln n_c)^{1/2}(1 + \delta_c)$ and solve for the small correction δ. If only the leading term is retained, the correction is

$$\delta_c = -\frac{\ln(\pi \ln n_c)}{4 \ln n_c}. \quad (47)$$

Therefore,

$$b_{n_c} \approx (2 \ln n_c)^{1/2}\left(1 - \frac{\ln(\pi \ln n_c)}{4 \ln n_c}\right) \quad (48)$$

$$a_{n_c} \approx (2 \ln n_c)^{1/2}\left(1 - \frac{\ln(\pi \ln n_c) - 2}{4 \ln n_c}\right) \quad (49)$$

$$A_{n_c} \approx 1. \quad (50)$$

The combination of equation (37) with equations (48), (49) and (50) yields equations (21) and (22). The combination of equations (48) and (49) with equation (38) leads to equations (23) and (24). Note that, in equation (47) and in the brackets of equations (48), (49), (21), (22), (23) and (24), terms of the type $[\ln(\pi \ln n_c)/2 \ln n_c]^2$ have been neglected. For typical n_c, these terms have values comparable to the terms $1/\ln n_c$. Therefore, inclusion of the latter terms in the expressions for centric reflections is to some extent conventional.

This work was supported by the Polish Ministry of Scientific Research and Information Technology grant to MB (MNiI, decision KO89/PO4/2004). MB thanks the European Molecular Biology Organization (EMBO) and HHMI for a Young Investigator award. Author contributions: MB derived the formulas and wrote the manuscript. GCh did the numerical work.

References

Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Blessing, R. H., Guo, D. Y. & Langs, D. A. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 47–71. Dordrecht: Kluwer Academic Publishers.
Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
Bochtler, M. & Chojnowski, G. (2007). *Acta Cryst.* A**63**, 146–155.
Cochran, W. & Woolfson, M. M. (1955). *Acta Cryst.* **8**, 1–12.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (2003). *IUCr Commission on Crystallographic Computing Newsletter*, No. 2, pp. 4–9, http://www.iucr.org/iucr-top/comm/ccom/newsletters/2003jul/index.html.
Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M. & Rossi, F. (2005). *Gnu Scientific Library: Reference Manual.* Bristol: Network Theory Limited.
Giacovazzo, C., Monaco, H. L., Artioli, G., Viterbo, D., Ferraris, G., Gilli, G., Zanotti, G. & Catti, M. (2002). *Fundamentals of Crystallography*. Oxford University Press.
Gumbel, E. J. (1958). *Statistics of Extremes.* New York: Columbia University Press.
Matsumoto, M. & Nishimura, T. (1998). *ACM Trans. Modeling Comput. Simul.* **8**, 3–30.