Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 15 June 2006
Accepted 20 December 2006

The highest reflection intensity in a resolution shell

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

[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

The Gumbel–Fisher–Tippett (GFT) extreme-value analysis is applied to evaluate the distribution, expectation value and standard deviation of the intensity J of the *strongest* reflection in a given resolution shell in the X-ray diffraction pattern of a crystal with many scattering atoms in the unit cell. For convenience, intensities are measured in units of the *average* reflection intensity in the resolution shell and, for simplicity, centric and acentric reflections are treated separately. For acentric reflections, the expectation value μ and standard deviation σ of J are $\mu = \ln n + \gamma$ and $\sigma = \pi/6^{1/2}$, where n is the number of crystallographically independent reflections in the resolution shell and $\gamma \approx 0.577$ is the Euler–Mascheroni constant. For centric reflections with expectation value 1 for the intensity, the corresponding expressions are $\mu = 2(\ln n + \gamma) - \ln(\pi \ln n)$ and $\sigma = 2\pi/6^{1/2} - \pi/(6^{1/2} \ln n)$. Extensive numerical simulations show that these formulas are excellent approximations for random atom configurations at all resolutions, and good approximations for real protein crystal structures in the resolution range between 2.5 and 1.0 Å.

1. Introduction

The distribution of *average* reflection amplitudes and intensities in the X-ray diffraction patterns of three-dimensional crystals has been analyzed extensively. The average reflection intensities decrease with resolution according to Wilson's law because the finite size of scattering atoms, static disorder and atomic mobility all attenuate high-resolution reflections more than low-resolution ones (Wilson, 1942). For wide enough resolution shells with sufficiently many reflections, the dependence of the average reflection intensity on resolution is smooth. For very thin shells with few reflections, pseudo-random fluctuations overlay the systematic dependence of the average reflection intensity but can be easily smoothed out, which is routinely done by programs such as *ECALC* that calculate normalized structure factors (Collaborative Computational Project, Number 4, 1994).

Within each resolution shell, the expectation value for most reflections equals the average reflection intensity in the resolution shell. A few reflections in special positions are expected to be stronger because symmetry enforces the constructive interference of X-rays (Stewart & Karle, 1976; Wilson, 1950). Here, we express all reflection intensities in a thin resolution shell in units of their expectation values, so that the expectation value for every reflection, including those in special positions, is 1 by definition. The distribution of reflection intensities within a thin resolution shell differs for acentric (subscript a) and centric (subscript c) reflections. The formulas for the cumulative distributions F_a and F_c and the non-cumulative distributions f_a and f_c of the normalized intensities i are well known (Wilson, 1949):

$$F_a(i) = \int_0^i f_a(x)\,dx = 1 - \exp(-i), \quad f_a(i) = \frac{dF_a}{di} = \exp(-i) \quad (1)$$

$$F_c(i) = \int_0^i f_c(x)\,dx = \mathrm{erf}[(i/2)^{1/2}],$$

$$f_c(i) = \frac{dF_c}{di} = \frac{1}{(2\pi)^{1/2}} \frac{1}{i^{1/2}} \exp(-i/2). \quad (2)$$

In this work, we are interested in the statistics of the strongest reflection in a given resolution shell. In other words, we consider a random arrangement of scattering atoms, find the strongest reflection and note its intensity. Then we repeat the procedure for a different random configuration of scattering atoms, again find the strongest reflection (which in general will have different indices) and again note the intensity. It is plausible that the histogram of largest intensities will converge to a limiting distribution as the procedure is reiterated. Here, we determine the analytical form of this distribution and derive approximate formulas for the expectation value and variance. In space groups with centric and acentric reflections, the two groups of reflections are treated separately because of their different underlying intensity distributions.

Our approach to the statistics of the strongest reflection is based on the Gumbel–Fisher–Trippett (GFT) theory of extreme values (Gnedenko, 1943; Gumbel, 1958; Sivapalan & Bloeschl, 1998; Kotz & Nadarajah, 2001). This theory states that the distribution of the largest quantity of n statistically independent random variables converges towards one of three universal distributions, the GFT distribution, the Fréchet distribution or the Weibull distribution. In the context of crystallographic reflection intensities, the relevant distribution is the GFT distribution, which can always be rescaled to the standard Gumbel form

$$g(x) = \exp(-x)\exp[-\exp(-x)], \tag{3}$$

with the corresponding cumulative distribution:

$$G(x) = \exp[-\exp(-x)]. \tag{4}$$

The expectation value and standard deviation of $g(x)$ can be expressed in terms of the Euler–Mascheroni constant $\gamma \simeq 0.577$ and $\pi \simeq 3.14159$ as

$$\mu = \gamma, \quad \sigma = \pi/6^{1/2}. \tag{5}$$

Here, we explore the predictions of GFT theory for the statistics of the strongest reflection separately for centric and acentric reflections. We compare the results with extensive numerical simulations for crystal structures with randomly placed scattering atoms and for real crystal structures taken from the Protein Data Bank (PDB, http://www.rcsb.org) (Berman *et al.*, 2000). The numerical results show that the analytical formulas are excellent approximations for random atom configurations at all resolutions, and good approximations for real protein crystal structures in the resolution range between 2.5 and 1.0 Å.

2. Materials and methods

2.1. Numerical methods

Utility programs were implemented in the C++ language with extensive use of routines from the GNU scientific library (Galassi *et al.*, 2005). 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). Random data distributed according to (1) were generated from a normally distributed random-number series with expectation value 0 and standard deviation 1 by squaring. Random data with a distribution according to (2) were obtained by adding the squares of two normally distributed variables with expectation value 0 and standard deviation $1/2^{1/2}$. Numerical integrations to evaluate equations (7) and (8) 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.

2.2. Simulations

Utility programs were written in the C++ language and combined with software of the CCP4 suite for protein crystallography (Collaborative Computational Project, Number 4, 1994) and the Clipper libraries (Cowtan, 2003). In simulations with random atom configurations, we placed $0.0281V$ [Å^3] C atoms in a unit cell of volume V, which is equivalent to the number of non-H protein atoms in a protein crystal with 50% solvent content or a Matthews coefficient of 2.5 Å^3 Da^{-1} (Matthews, 1968). Owing to the mass differences between C, N and O atoms, this corresponds to a density of 1 Da (3 Å^3)$^{-1}$ or a Matthews coefficient of 3 Å^3 Da^{-1}. Structure factors and normalized structure factors were calculated with the CCP4 programs *SFALL* (Agarwal, 1978) and *ECALC* (Collaborative Computational Project, Number 4, 1994), respectively. Intensities in units of their expectation values were determined by squaring the moduli of the E values or by dividing squared structure factors by the shell averages, as indicated. In the former procedure, systematically strong reflections in special positions were automatically treated correctly. In the latter procedure, the required corrections were applied by our own programs. Errors of all simulation results were estimated according to the standard formulas for the sample variance distribution (*MathWorld: The Web's Most Extensive Mathematical Resource*, http://mathworld.wolfram.com/).

2.3. Real crystals

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, 1 m2d, 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). All graphs were prepared with the *GRACE* software (http://plasma-gate.weizmann.ac.il/Grace/).

3. Predictions

3.1. Notations and conventions

Throughout this work, I, J, K stand for intensities treated as (pseudo)random variables, and i, j, k are used when actual values are meant. I and i stand for the intensity of a pre-picked reflection (in units of the expectation value or shell average), J and j for the intensity of the strongest reflection in a thin resolution shell (in the same units), and K and k for the intensity of 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 are reserved for the intensity distribution of any pre-picked reflection and are well known [equations (1) and (2)]. g and G denote the non-cumulative and cumulative Gumbel distributions according to equations (3) and (4). Note that the term

Gumbel distribution is reserved for this special case and that rescaled and shifted versions of the distribution are referred to as Gumbel–Fisher–Tippett (GFT) distributions. Here, we only need the special case with expectation value 0 and variance 1, and denote it by H and h for the cumulative and non-cumulative forms. As usual, μ and σ stand for the expectation value and the standard deviation. Throughout, Pr abbreviates probability, n denotes the number of unique reflections, from which the strongest reflection is selected, and N stands for the number of scattering atoms in the unit cell. Subscripts a and c distinguish between acentric and centric reflections.

3.2. Distribution of the highest reflection intensity

Analytical formulas for the intensities of the strongest acentric and centric reflections in a thin resolution shell can be derived under the assumption that statistical interdependencies between reflection intensities can be neglected. This might appear as a severe approximation, because intensity correlations are expected even for random atom configurations. For normalized structure factors, it has been shown that the importance of correlations decreases with the number of scattering centers (Cochran & Woolfson, 1955; Woolfson, 1987), which explains the success of direct methods in the field of small-molecule crystallography and the much smaller role of these methods in macromolecular crystallography. We are not aware of similar correlation analyses for normalized intensities, but it is at least plausible (and confirmed by our unpublished calculations) that also for normalized intensities the importance of correlations goes down with the number of scattering atoms. Ultimately, this approximation is justified by the success of the predictions that are based on it.

If the reflection intensities are treated as statistically independent, the cumulative distribution of the intensity J of the strongest of n reflections can be expressed simply as the product of the n cumulative distributions F of the individual reflection intensities, which are conveniently written as distributions with an exponential tail $F(x) = 1 - \exp[-u(x)]$:

$$\Pr(J \leq x) = [F(x)]^n = \{1 - \exp[-u(x)]\}^n. \quad (6)$$

Then, by definition,

$$\mu(J) = \int_0^\infty \mathrm{d}x\, x \frac{\mathrm{d}}{\mathrm{d}x} \Pr(J \leq x)$$

$$= \int_0^\infty \mathrm{d}x\, xn\{1 - \exp[-u(x)]\}^{n-1} \exp[-u(x)]u'(x) \quad (7)$$

$$\mu(J^2) = \int_0^\infty \mathrm{d}x\, x^2 \frac{\mathrm{d}}{\mathrm{d}x} \Pr(J \leq x)$$

$$= \int_0^\infty \mathrm{d}x\, x^2 n\{1 - \exp[-u(x)]\}^{n-1} \exp[-u(x)]u'(x) \quad (8)$$

$$\sigma(J) = [\mu(J^2) - \mu^2(J)]^{1/2}. \quad (9)$$

These expressions can only be evaluated numerically, but fortunately GFT extreme-value analysis allows significant simplification. Note that the probability for the strongest reflection intensity to be smaller than $u^{-1}(\ln n)$, where u^{-1} denotes the inverse of u, $u(u^{-1}(x)) = u^{-1}(u(x)) = x$, can be approximated as

$$\Pr(J \leq u^{-1}(\ln n)) = [F(u^{-1}(\ln n))]^n = \left[1 - \frac{1}{n}\right]^n \approx \frac{1}{e} \quad \text{for large } n. \quad (10)$$

Therefore, the value of the cumulative distribution is intermediate between 0 and 1 for the argument around $u^{-1}(\ln n)$, which suggests that the non-cumulative distribution peaks near this value. This is confirmed by a more detailed analysis and suggests interpolation of u around $u^{-1}(\ln n)$.

$$u(x) \approx \ln n + a_n(x - b_n) \quad (11)$$

$$a_n = u'(u^{-1}(\ln n)), \quad b_n = u^{-1}(\ln n). \quad (12)$$

Plugging this into equation (6) yields

$$\Pr(J \leq x) \approx \left(1 - \frac{\exp[-a_n(x - b_n)]}{n}\right)^n. \quad (13)$$

Using the well known identity

$$\lim_{n\to\infty}\left(1 - \frac{\alpha}{n}\right)^n = \exp(-\alpha) \quad (14)$$

and assuming that $a_n = u'(u^{-1}(\ln n))$ and $b_n = u^{-1}(\ln n)$ vary sufficiently slowly with n to be treated as constants, equation (13) can be simplified further to the usual forms in extreme-value analysis:

$$\Pr(J \leq x) \approx \exp\{-\exp[-a_n(x - b_n)]\} = G(a_n(x - b_n)) \quad (15)$$

$$\Pr(a_n(J - b_n) \leq x) \approx \exp[-\exp(-x)] = G(x). \quad (16)$$

In (15) and (16), $G(x)$ is the cumulative Gumbel distribution [equation (4)]. A more rigorous derivation of some of the steps from equation (6) to equations (15) and (16) can be found in more mathematically oriented treatments of extreme-value analysis (Gnedenko, 1943; Gumbel, 1958; Sivapalan & Bloeschl, 1998; Kotz & Nadarajah, 2001). In words, (15) and (16) express that the highest reflection intensity J in a thin resolution shell is GFT distributed if a large set of crystal structures that differ only in the (randomly chosen) positions of scattering atoms is considered. Note that this result applies separately to the strongest acentric reflection J_a and to the strongest centric reflection J_c.

3.3. Expectation value and standard deviation of the highest reflection intensity

Equation (15) is a rescaled and shifted form of the cumulative Gumbel distribution. As the expectation value μ and standard deviation σ of the Gumbel distribution are known from equation (5), the corresponding values for the distribu-

Table 1
Probabilities for large deviations from the expectation value according to the GFT distribution [calculated from equation (22)].

Note the asymmetry in the distribution and note that the values are independent of the values for μ and σ.

m	$\Pr(J \le \mu - m\sigma)$	$\Pr(J \ge \mu + m\sigma)$	$\Pr(J \le \mu - m\sigma \vee J \ge \mu + m\sigma)$
0	57%	43%	100%
1	13%	14%	27%
2	0.07%	4.2%	4.3%
3	$<10^{-9}$%	1.2%	1.2%
4	$<10^{-9}$%	0.33%	0.33%

Table 2
Confidence intervals around the expectation value according to the GFT distribution.

The boundaries of the interval were chosen so that high- and low-intensity outliers are equally probable. Numerical values were calculated from equation (21).

$\Pr(\mu - \alpha\sigma \le J \le \mu + \beta\sigma)$	α	β
90%	1.31	1.87
95%	1.47	2.42
99%	1.75	3.68
99.5%	1.85	4.22
99.9%	2.03	5.48

tion of the highest reflection intensity can be readily derived. Combining equations (5) and (15), one readily obtains

$$\mu(J) = \frac{1}{a_n}\gamma + b_n, \quad \sigma(J) = \frac{1}{a_n}\frac{\pi}{6^{1/2}}. \tag{17}$$

This equation relates the expectation value and standard deviation of the GFT distribution of the strongest reflection to the distribution for a typical reflection. As this distribution differs for acentric and centric reflections [see equations (1), (2)], equation (17) has to be evaluated separately for the two cases.

(*a*) *Acentric case*: The cumulative distribution of intensities of acentric reflections is $F_a(i) = 1 - \exp[-u_a(i)]$, where $u_a(i) = i$ and $u_a'(i) = 1$ according to equation (1). It follows straightforwardly from equation (12) that $a_{n,a} = 1$ and $b_{n,a} = \ln n_a$, which in combination with (17) leads to

$$\mu(J_a) = \ln n_a + \gamma, \quad \sigma(J_a) = \pi/6^{1/2}. \tag{18}$$

Note that for a typical thin shell, such as the 1.5 ± 0.01 Å shell discussed below, and for typical crystals in the PDB, the number of acentric reflections n_a in the shell is between a few hundred and a few thousand. Further note that $\mu(J_a)$ grows logarithmically with n_a, which is physically unreasonable for very large n_a (very high resolution) and a fixed number of scattering atoms N. The non-physical result is due to a breakdown of the formulas for the distributions f_a and F_a for very large reflection intensities. For typical values n_a, there is no problem because values for $\mu(J_a) = \ln n_a + \gamma$ are in a range where the original distributions f_a and F_a are excellent approximations.

(*b*) *Centric case*: The cumulative distribution for a pre-picked centric reflection is $F_c(i) = \mathrm{erf}[(i/2)^{1/2}]$ according to equation (2). From the definition $F_c(i) = 1 - \exp[-u(i)]$ and the expansion of the error function for large arguments $\mathrm{erf}(x) \approx 1 - [\exp(-x^2)/(x\pi^{1/2})][1 - 1/(2x^2) + \ldots]$, one can readily deduce $u_c(i) = i/2 + \frac{1}{2}\ln i + \frac{1}{2}\ln(\pi/2) + 1/i + \ldots \approx i/2$. If the expansion is limited to the leading order, it implies that $b_{n,c} = u_c^{-1}(\ln n_c) \approx 2\ln n_c$ and $a_{n,c} = u_c'(u_c^{-1}(\ln n_c)) \approx 1/2$ according to (12). In combination with equation (17), it then follows that

$$\mu(J_c) = 2(\ln n_c + \gamma), \quad \sigma(J_c) = 2\pi/6^{1/2}, \tag{19}$$

which differs from the result for acentric reflections by an extra factor 2. To obtain a more accurate estimate, we set $u_c^{-1}(\ln n_c) = (1 + \delta)2\ln n_c$ and then exploit that δ is much smaller than 1 to evaluate it approximately. The resulting better, but still approximate, estimates are

$$\mu(J_c) = 2(\ln n_c + \gamma) - \ln(\pi \ln n_c),$$
$$\sigma(J_c) = \frac{2\pi}{6^{1/2}}\left(1 - \frac{1}{2\ln n_c}\right). \tag{20}$$

For a typical thin shell, such as the 1.5 ± 0.01 Å shell discussed below, and for typical crystals in the PDB, the number of centric reflections n_c in the shell is between 20 and 200. With such values of n_c, the correction terms to the first-order approximations for $\mu(J_c)$ are between 2 and 3 and the correction terms for $\sigma(J_c)$ are less than 1. Note that the correction terms in (20) become negligible compared to the leading terms in (19) for very large n_c.

3.4. Confidence intervals for the highest reflection intensity

The highest reflection intensity is GFT distributed both in the acentric and in the centric case [equation (15)]. Therefore, confidence intervals for the highest reflection intensity can be expressed in a universal form, which applies separately to both cases. For this, it is necessary to express the actual highest reflection intensity in terms of its deviation from $\mu(J)$, expressed in multiples m of the standard deviation $\sigma(J)$ (Table 1). Note the asymmetry of the Gumbel distribution: for a random arrangement of the scattering atoms, it is practically impossible for the actual intensity of the strongest reflection to be more than 3σ *below* the expectation value. In contrast, random arrangements that lead to a strongest intensity more than 3σ *above* the expectation value occur in more than 1% of cases (Table 1).

The results can be expressed in slightly different form as intervals around the expectation value, which cover the actual intensity of the strongest reflection for a random arrangement of scattering atoms with a predetermined level of confidence (*e.g.* 99%). To make the choice of the interval around $\mu(J)$ unique, we have additionally required that the probability of high- and low-intensity outliers should be equal. With this additional requirement and the abbreviation t for the probability $\mu - \alpha\sigma \le X \le \mu + \beta\sigma$, it is straightforward to show that α and β should be

$$\alpha = \frac{6^{1/2}}{\pi}\left\{\gamma + \ln\left[-\ln\left(\frac{1-t}{2}\right)\right]\right\},$$
$$\beta = \frac{6^{1/2}}{\pi}\left\{-\gamma - \ln\left[-\ln\left(\frac{1+t}{2}\right)\right]\right\}. \quad (21)$$

Numerical results for representative confidence levels t are collected in Table 2. For a random arrangement of scattering atoms, the intensity of the strongest reflection falls with a probability of 90% within an asymmetric interval of width $1.31\sigma + 1.87\sigma = 3.18\sigma$ around the expectation value. A wider interval of width $1.75\sigma + 3.68\sigma = 5.43\sigma$ should cover the intensity of the strongest reflection with 99% confidence (Table 2).

4. Tests with simulated data

4.1. Limitations of the analytical treatment

The analytical treatment of the intensity of the strongest reflection is based on a number of assumptions and approximations, even for random atom configurations. (*a*) The analysis is limited to cases with many atoms in the unit cell, so that the underlying intensity distributions (1) and (2) are good approximations. (*b*) Statistical interdependencies between reflection intensities are neglected, although weak correlations are expected even for random atom configurations, as has been extensively shown for normalized structure factors in the context of direct methods (Cochran & Woolfson, 1955; Woolfson, 1987). (*c*) It is assumed that the strongest reflection is selected from many reflections, so that the GFT analysis is a good approximation. In a first step, we tested the merit of these approximations without the extra complications from non-random atom configurations in real crystal structures. We focused on four resolution shells, at 1.5 ± 0.01, 2.0 ± 0.02, 3.4 ± 0.05 and 5.0 ± 0.1 Å. The reciprocal-space volumes of these four resolution shells were in the ratio $1.00 : 0.63 : 0.18 : 0.08$.

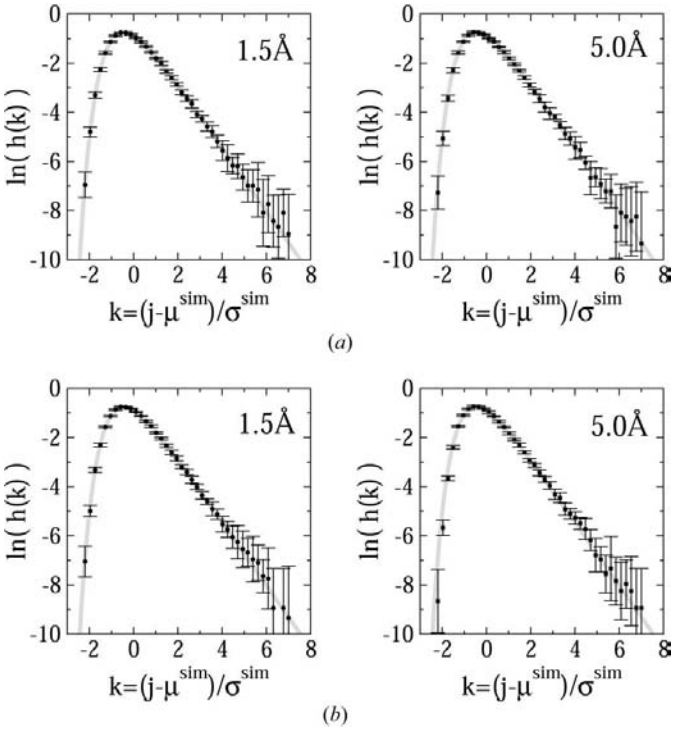

Figure 1
Distribution of the intensity of the strongest reflection for 100 000 different configurations of scattering atoms in (*a*) space group $P1$ and (*b*) space group $P\bar{1}$ in resolution shells 1.5 ± 0.01 Å and 5.0 ± 0.1 Å. Simulations were run for 50% solvent content and 4000 randomly placed C atoms (this is approximately the number of non-H atoms in 440 amino acids). $\mu^{\rm sim}(J)$ and $\sigma^{\rm sim}(J)$ were derived directly from the numerically obtained distributions and are independent of the theoretical predictions for these values. Error bars indicate uncertainties at the 1σ level.

4.2. The distribution of the highest reflection intensity

Simulations of the distribution of the strongest reflection intensity were run for test cases with Matthews coefficient 3 and 4000 atoms in $P1$ and $P\bar{1}$ unit cells. The two space groups were chosen because they have only acentric and centric

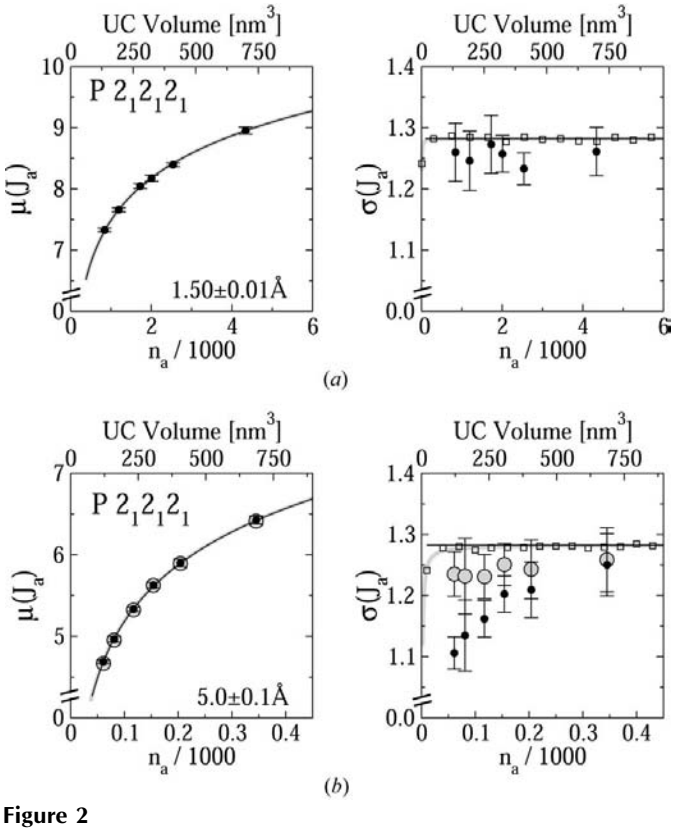

Figure 2
The expectation value μ and standard deviation σ of the intensity J_a of the strongest acentric reflection in the resolution shells 1.5 ± 0.01 Å (*a*) and 5.0 ± 0.1 Å (*b*). n_a denotes the number of crystallographically unique acentric reflections. Black lines are calculated according to the analytical equation (18), and gray lines are obtained by numerical integration of equations (7) to (9). Open boxes show the results of a simulation with uncorrelated random variables distributed according to equation (1). Circles show the results of simulations with 10 000 random atom configurations in space group $P2_12_12_1$. Small black circles are obtained when unsmoothed shell averages of the intensity are used for normalization. Large gray circles present the results for the normalization derived by the *ECALC* program, which smoothes the fluctuations of average intensities in thin resolution shells. UC stands for unit cell.

reflections, respectively. To make comparisons with theory independent of the expectation value and standard deviation of the distribution, the simulated data were plotted in terms of the reduced intensity $K = (J - \mu^{\rm sim})/\sigma^{\rm sim}$. Note that $\mu^{\rm sim}$ and $\sigma^{\rm sim}$ were derived from simulated histograms and not from the GFT predictions. By definition, the expectation value and standard deviation of K are 0 and 1, and therefore the Gumbel distribution has to be rescaled for a direct comparison. Combining equations (4) and (5), it follows immediately that the appropriate cumulative distribution is

$$H(k) = \exp\Big[-\exp\Big(-\frac{\pi}{6^{1/2}}k - \gamma\Big)\Big]. \qquad (22)$$

Differentiating this expression yields the corresponding non-cumulative distribution:

$$h(k) = \frac{\pi}{6^{1/2}}\exp\Big(-\frac{\pi}{6^{1/2}}k - \gamma\Big)\exp\Big[-\exp\Big(-\frac{\pi}{6^{1/2}}k - \gamma\Big)\Big]. \qquad (23)$$

The agreement is excellent even in the tails of the distribution for both $P1$ and $P\bar{1}$ unit cells at all tested resolutions (Fig. 1 and data not shown).

4.3. Expectation value and standard deviation of the highest reflection intensity

The atom density was kept constant at 1 Da $(3\ \text{Å}^3)^{-1}$, but 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, 10 000 random atom configurations were generated.

(*a*) *Acentric case*: Initial tests were run in space group $P2_12_12_1$. In this space group, reflections with all indices different from 0 are acentric. For the 1.5 ± 0.01 Å shell, all analytical and numerical predictions for $\mu(J_a)$ and $\sigma(J_a)$ are consistent with simulation results (Fig. 2*a*). For the 5.0 ± 0.1 Å resolution shell, which has approximately tenfold smaller volume, predictions and simulations agree for $\mu(J_a)$ (Fig. 2*b*, left panel), but diverge for $\sigma(J_a)$, particularly for small reflection numbers n_a (Fig. 2*b*, right panel, note the break in the ordinate). The discrepancy depends on the normalization procedure: it is less severe if smoothed shell intensity averages are used.

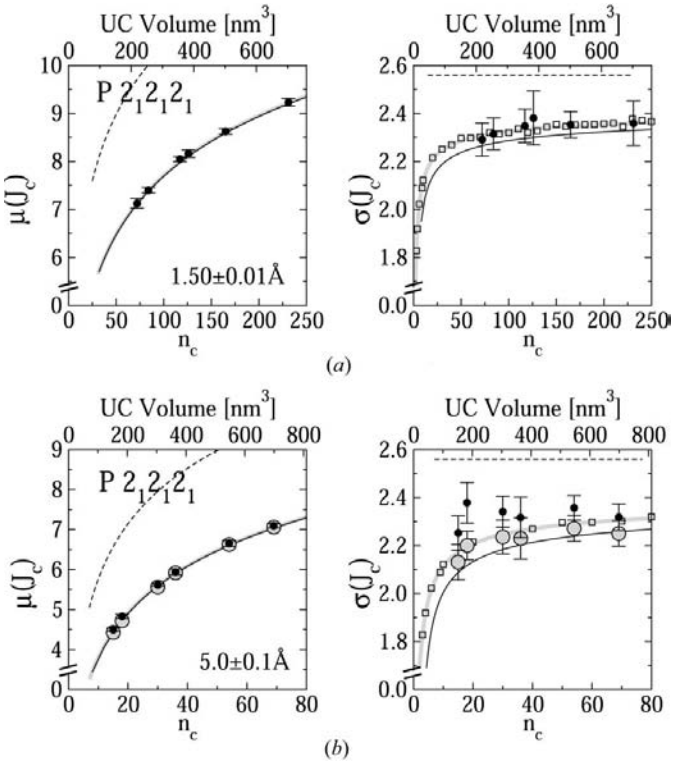

Figure 3
The expectation value μ and standard deviation σ of the intensity J_c of the strongest centric reflection in the resolution shells 1.5 ± 0.01 Å (*a*) and 5.0 ± 0.1 Å (*b*). n_c denotes the number of crystallographically unique centric reflections. Dashed black lines are according to equation (19) and continuous black lines according to equation (20). Open boxes show the results of a simulation with uncorrelated random variables distributed according to equation (2). Gray lines and circles have meanings analogous to in Fig. 2.

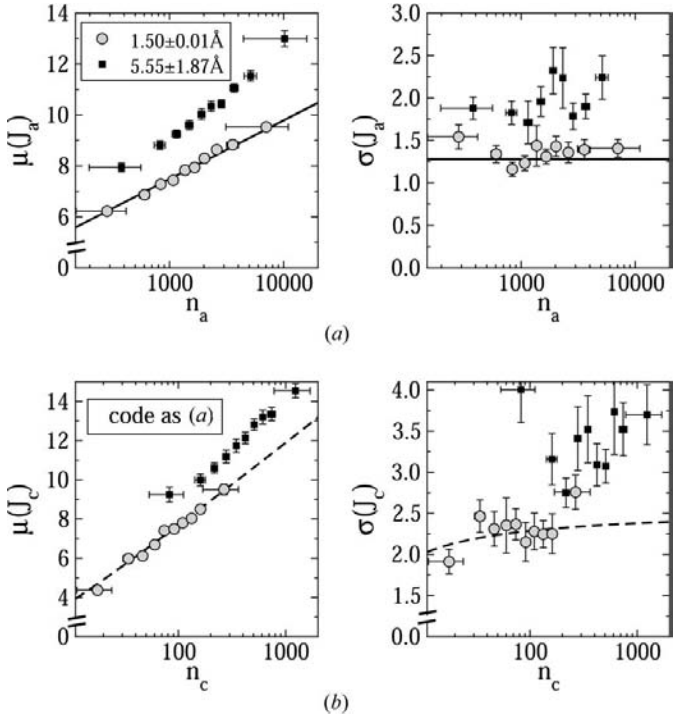

Figure 4
The expectation value μ and standard deviation σ of the intensity (*a*) J_a of the strongest acentric reflection and (*b*) J_c of the strongest centric reflection in the resolution shells 1.5 ± 0.01 Å (gray circles) and 5.55 ± 1.87 Å (black squares). Bins for n values were chosen as a compromise between the conflicting requirements for large bins to collect sufficient statistics for $\mu(J)$ and $\sigma(J)$ and for small bins to keep the spread of n low. In the 55 ± 1.87 Å shell, two structures with very regular β architectures (1k5c, 1jl0) that lead to extreme outliers were excluded from the analysis.

As expected, checks in space group $P2_12_12_1$ for other resolution shells confirm the trend that the predictions for $\mu(J_a)$ are good throughout, and predictions of $\sigma(J_a)$ tend to be better for higher-resolution shells with more reflections (Figs. S1*A*, *B*).[1] Simulations support the validity of the analytical formulas also for other space groups and across the resolution range (Figs. S1*C*–*J*). In centered space groups such as *C*121, systematically absent reflections do not contribute to n_a in equation (18) (Figs. S1*C*–*F*).

(*b*) *Centric case*: As above, initial tests were run in space group $P2_12_12_1$. In this space group, reflections in the $h = 0$, $k = 0$ and $l = 0$ planes of reciprocal space are centric. As anticipated, we find that the formulas of equation (19) for $\mu(J_c)$ and $\sigma(J_c)$ overestimate both quantities severely, but the next order approximations of equation (20) agree well with simulations (compare the dashed and continuous black lines in Fig. 3). For centric reflections, the extreme-value approximations that lead to equation (20) introduce a small extra error into the prediction of $\sigma(J_c)$, which can be avoided by the numerical integration according to equations (7) to (9) (compare the agreement of the continuous black and gray lines with the open boxes). As in the case of the acentric data, *ECALC* smoothing of shell average reflection intensities improves the agreement between the analytical results and the simulations (Fig. 3).

We have confirmed that the formulas for $\mu(J_c)$ and $\sigma(J_c)$ hold at other resolutions (Figs. S2*A*, *B*) and for other non-centrosymmetric space groups (Figs. S2*C*–*F*). As expected, the formulas of equation (20) apply to centrosymmetric space groups such as $P\bar{1}$, which have only centric reflections, as well. In these space groups, the analytical approximations are better than in non-centrosymmetric space groups, because there are more centric reflections in each thin resolution shell (compare Figs. S2*C*–*F* and Figs. S2*G*–*J*).

5. Tests with real data from the Protein Data Bank (PDB)

5.1. Selection of test cases

Structures with a resolution better than 1.5 Å without nucleic acids were selected from the Protein Data Bank (PDB, release date 18 April 2006). Duplicates with over 90% sequence identity were removed to avoid bias due to the presence of multiple nearly identical structures. Calculated structure factors were used throughout because experimental structure factors (*a*) were not available for all structures, (*b*) could be tainted by the presence of spurious ice or salt peaks, (*c*) would require correction for overall anisotropic *B* factors, which can be done, but introduces additional complications. As before, we focused primarily on two resolution shells, which are representative for high- and low-resolution data, respectively.

[1] Supplementary figures, indicated with letter S before the number, are available from the IUCr electronic archive (Reference: WE5015). Details for accessing these data are given at the back of the journal.

5.2. Expectation value and standard deviation of the highest reflection intensity

In contrast to the situation for simulated data, which can be generated in any desired quantity to obtain reliable statistics for $\mu(J)$ and $\sigma(J)$, there is typically only one structure for a given n. 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 $\mu(J)$ and $\sigma(J)$ and for small bins to keep the spread of n low. The result of this analysis is presented in Fig. 4. For both acentric and centric reflections, we find excellent agreement between the predictions and the results for real data for the high-resolution shell. In contrast, there is a significant discrepancy for the low-resolution shell, where the highest intensities $\mu(J_a)$ and $\mu(J_c)$ and also $\sigma(J_a)$ and $\sigma(J_c)$ are larger than predicted.

5.3. Confidence interval for the highest reflection intensity

The above analysis does not directly test whether the predicted asymmetry of the J_a and J_c distributions is present in real data. Therefore, these data were expressed in a different way to look for this feature in the 1.5 ± 0.01 Å resolution shell (Figs. 5*a*, *b*). If the approximations for random data were applicable to real data, then, according to equations (18), (20) and (21) and Table 2, the region $\mu - 1.31\sigma \leq J \leq \mu + 1.87\sigma$ (the orange stripe in Figs. 5*a*, *b*) should cover 90% of all highest reflection intensities, with 5% outliers each above and below this interval. The wider interval $\mu - 1.75\sigma \leq J \leq \mu + 3.68\sigma$ (the yellow stripe in Figs. 5*a*, *b*) should even cover 99% of all real cases, again with an equal number of outliers above and below the interval. Qualitatively, the scatter plots are in excellent agreement with the predictions from the analytical formulas. As predicted from the asymmetry of the GFT distribution, the scatter plots of J_a and J_c have very sharp lower borders, but fade out much more gradually towards high J_a and J_c values (Figs. 5*a*, *b*).

For a more detailed qualitative comparison between the (approximate) predictions for random data and the results for real data, it is necessary to quantify the percentage of outliers. This has been done for the predicted 90% confidence interval, initially for the 1.5 ± 0.01 Å resolution shell, and subsequently for many other thin shells at different resolutions. The number of large and small J_a and J_c outliers was then plotted as a function of the average resolution of the shell in the range for the resolution range from 1.0 to 6 Å. Based on the analytical treatment (for random atom configurations), there should be 5% large J and 5% small J outliers. In agreement with the results in Fig. 4, we find that the predictions are good for high-resolution shells (Figs. 5*c*, *d*) but break down for shells at low resolution (Figs. 5*e*, *f*).

5.4. Good agreement at high resolution

For thin shells at high resolution in the range from 1.0 to 2.5 Å (Figs. 5*c*, *d*), the predictions agree well with data calculated for real crystal structures. Independent of the precise method of intensity normalization (by *ECALC* with

smoothing or simply division without smoothing), there are slightly too many large J outliers and slightly too few small J outliers. Apart from this minor discrepancy, the good agreement between predictions and results for real crystal structures shows that the statistics for the highest reflection intensity in this resolution range are not strongly influenced by the non-random features of real crystal structures (Figs. 5*c*, *d*). Analogous findings are obtained for the predicted 99% confidence interval (Figs. S3*A*, *B*)

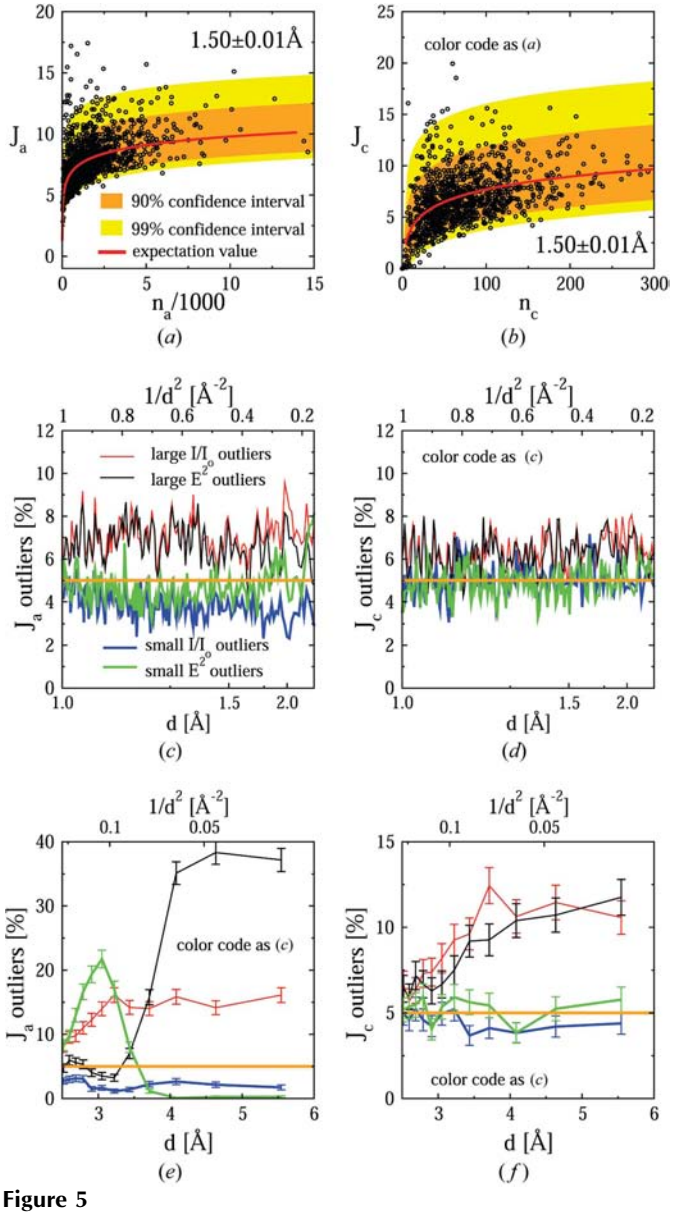

Figure 5
Strongest acentric (*a*, *c*, *e*) or centric (*b*, *d*, *f*) reflection in a thin resolution shell for real structures from the Protein Data Bank (PDB, release date 18 April 2006). (*a*), (*b*) Intensity of the strongest reflection in the shell 1.5 ± 0.01 Å. The red line indicates the prediction for the expectation value according to (*a*) equation (18) or (*b*) equation (20), and the orange and yellow regions show the predicted 90 and 99% confidence intervals according to equation (21) and Table 2. (*c*, *d*, *e*, *f*) Quantification of (*a*) and (*b*), not only for the 1.5 ± 0.01 Å shell but for many thin shells of equal volume centered at various resolutions. Theory predicts 5% outliers each above and below the predicted 90% confidence interval. The red and black lines show the actual percentage of large J outliers and blue and green lines indicate the actual percentage of small J outliers for real structures from the PDB. Red and blue lines are based on the normalization with respect to actual shell intensity averages (label I/I_0), and black and green lines are based on intensities calculated from normalized structure factors, which in turn are based on smoothed intensity averages (label E^2). The high (*c*, *d*) and low (*e*, *f*) resolution ranges are presented separately. In (*d*) and (*f*), structures with fewer than 10 centric reflections in a thin resolution shell were excluded from the analysis.

5.5. Poor agreement at low resolution

In contrast, there are major discrepancies for shells at low and very low resolution in the range from 2.5 to 6 Å (Figs. 5*e*, *f*). The discrepancies are particularly serious for J_a values calculated by the *ECALC* program (termed E^2 outliers in the figure). The resolution dependence of the percentage of outliers is remarkable. There are far too many small J_a outliers and far too few large J_a outliers in the resolution range around 3 Å. At still lower resolution, the discrepancies are even larger, but now the effects are reversed and too many large J_a outliers and too few small J_a outliers are observed (Figs. 5*e*, *f*). Similar discrepancies are found for the predicted 99% confidence interval (Fig. S3*C*, *D*).

After much testing, this complicated behavior was traced to an unexpected source, namely the detailed behavior of the *ECALC* program (Collaborative Computational Project, Number 4, 1994). We had assumed that the E^2 values from *ECALC* would be 1 on average in all resolution shells. This was indeed true if the *ECALC* was applied to thin shells of diffraction data as in Figs. 2 and 3, but turned out to be untrue if *ECALC* was applied to thick resolution shells (Fig. S4). In this case, we found that a correction factor greater than 1 was required around 3.0 Å and a correction factor smaller than 1 was required at even lower resolution for average E^2 values to be 1 in each shell. Without the correction, there are too many small J_a and too few large J_a outliers around 3.0 Å. At even lower resolution, there are too many large J_a and too few small J_a, exactly as would be expected (Figs. 5*e*, *f*).

After correcting for this feature of *ECALC*, so that average normalized intensities are truly 1 at all resolutions, there are still too few small J_a and too many large J_a outliers (Fig. 6*a*), but now the results for the *ECALC* procedure agree with the results for the simple normalization protocol (compare Fig. 6*a* and Fig. 5*e*). We believe that the remaining discrepancy can be attributed to the features of real crystal structures, which tend to enhance some reflections at the expense of others. The extreme example of a near-perfect local symmetry, which almost extinguishes every second reflection and strongly enhances the rest, readily explains why an uneven distribution of reflection intensities increases J and therefore leads to too many *large J* outliers if it is not taken into account. Therefore, structures with strong pseudo-origin peaks were excluded from the analysis from the beginning (see *Materials and methods* for details).

Numerical tests were run to pinpoint precisely which non-random features of real crystal structures are responsible for the deviations from the predicted largest reflection intensity. In short, we modified real structures from the PDB so that

they retained some non-random features, but lacked others. As it is the closest approximation to the situation with real data, we have kept the protein mask but randomly redistributed the scattering atoms and then applied van der Waals repulsion between atoms, but no bonding terms to account for the short atom–atom distances in crystal structures (Fig. 6*b*). As expected, there is still substantial discrepancy at low resolution. Essentially the same is found if atoms are randomly redistributed between the mask, but not subjected to any interactions (Fig. 6*c*). A substantial discrepancy at very low resolution remains even after atom positions have been fully randomized (Fig. 6*d*). The discrepancy becomes insignificant only after setting all B factors to 0 (Fig. 6*e*). Note that the excellent agreement between predictions and calculations in Fig. 6(*e*) shows that intensity correlations, which are neglected in the analytical approach, apparently do not play a major role.

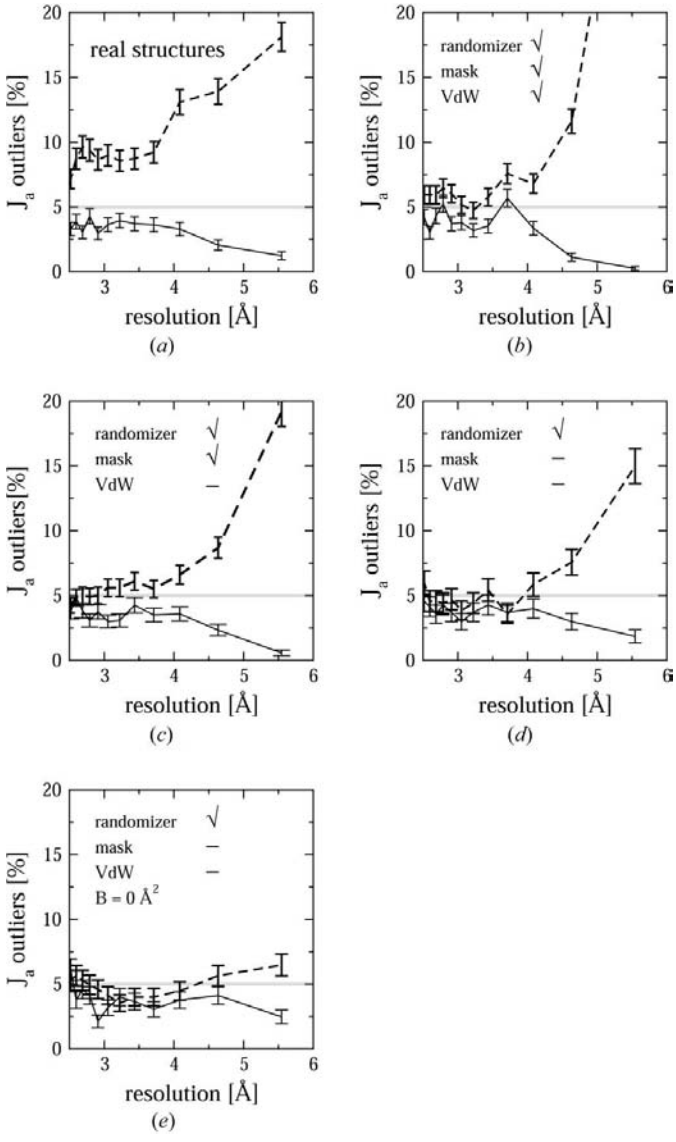

Figure 6
Influence of various non-random features on the percentage of outliers outside the predicted 90% confidence interval. The gray line indicates the theoretical prediction (5%), the dashed curve the actual percentage of large J_a outliers and the continuous curve the actual percentage of small J_a outliers. In all cases, smoothed intensity averages from *ECALC* are used for normalization and corrected so that the average E^2 equals 1. Results are presented for crystal structures (*a*) without modification, (*b*) after randomization of atom positions within the protein mask and subsequent imposition of van der Waals restraints, (*c*) after randomization of atom positions within the protein mask without additional restraints, (*d*) after full randomization of atom positions in the asymmetric unit and (*e*) after full randomization of atom positions in the asymmetric unit and resetting all B factors to 0.

6. Generalizations and applications

In this work, we have focused on intensities in thin resolution shells, so that expectation values could be normalized either to their expectation value or to the shell intensity average. Without the latter, the limitation to thin resolution shells can be dropped and all reflections, including those in special positions, can be treated equally, with only the distinction between acentric and centric reflections. However, the 'strongest' reflection in this sense is no longer the reflection with the highest intensity, but the most 'unusual' reflection in the sense that its intensity exceeds the expectation value by the largest factor. Therefore, our formulas, applied to either thin shells or wide shells, should be applicable as an alternative to the usual E-value-based criteria to reject unlikely reflections at the stages of X-ray data integration or scaling. Our tests with real data show that our formulas should work well from the highest possible resolution to 2.5 Å, except for structures with a strong pseudo-origin peak. Owing to the influence of non-random features in real data on the highest-reflection intensity at low resolution, our analysis should not be applied to data below 2.5 Å resolution, or the confidence interval should be chosen wider than would be required for data from crystals with atoms in fully random positions.

This work was supported by the Polish Ministry of Scientific Research and Information Technology grant to MB (MNiI, decision KO89/PO4/2004). We are grateful to Professor Robert Huber for reading an early version of this manuscript, and to Dr Honorata Czapinska for proof-reading the final version. MB thanks the European Molecular Biology Organization (EMBO) and the Howard Hughes Medical Institute (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.
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.
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 Computing Commission Newsletter*, pp. 4–9.
Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M. & Rossi, F. (2005). *GNU Scientific Library: Reference Manual*. Network Theory Limited, Bristol.
Gnedenko, B. (1943). *Ann. Math.* **44**, 423–453.
Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press.
Kotz, S. & Nadarajah, S. (2001). *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press.
Matsumoto, M. & Nishimura, T. (1998). *ACM Trans. Model. Comput. Simul.* **8**, 3–30.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Sivapalan, M. & Bloeschl, G. (1998). *J. Hydrology*, **204**, 150–167.
Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005–1007.
Wilson, A. J. C. (1942). *Nature (London)*, **150**, 151–152.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wilson, A. J. C. (1950). *Acta Cryst.* **3**, 258.
Woolfson, M. M. (1987). *Acta Cryst.* A**43**, 593–612.