Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning

**Jennifer E. Padilla and Todd O. Yeates***

University of California, Los Angeles, Department of Chemistry and Biochemistry, 611 Charles E. Young Drive East, Los Angeles, CA 90095-1569, USA

Correspondence e-mail: yeates@mbi.ucla.edu

Received 6 February 2003
Accepted 8 April 2003

A new approach to analyzing macromolecular single-crystal X-ray diffraction intensity statistics is presented. Instead of considering reflections in resolution shells, differences between local pairs of reflection intensities are taken and normalized separately. When the two reflections to be compared (having intensities $I_1$ and $I_2$, respectively) are chosen appropriately, the behavior of the parameter $L = (I_1 - I_2)/(I_1 + I_2)$ is insensitive to phenomena that tend to confound traditional intensity statistics, such as anisotropic diffraction and pseudo-centering. The distributions and expected values for $L$ take simple forms when the intensity data are from ordinary crystals or from perfectly twinned specimens. The robustness of the approach is demonstrated with examples using real proteins whose diffraction data appear aberrant by other methods of intensity analysis. The new statistic is better suited than other available methods for diagnosing perfect hemihedral twinning.

## 1. Introduction

In a diffraction experiment, each structure factor is the sum of separate scattering-factor vectors from the many atoms in the crystal. By drawing an analogy to random walks, diffraction intensities can be expected to follow well known distributions first described by Wilson (1949) (reviewed in Giacovazzo, 1992; Drenth, 1999). The intensity distributions take different forms under different circumstances, such as in centric zones or when hemihedral twinning occurs. Intensity statistics are therefore often used to analyze symmetry or to detect anomalies in observed data sets.

In typical applications of intensity statistics, intensities are first normalized in resolution shells. This deals with the problem of reflections at different resolutions having different expected values. Following normalization in shells, cumulative probability distributions of the intensities are examined or various moments of the intensity distribution are compared with theoretical values expected from Wilson statistics. These current approaches implicitly assume that various reflections in the same resolution shell have equal expected values. This assumption fails under a variety of situations, including when a crystal diffracts anisotropically or when pseudo-centering gives rise to weak and strong classes of reflections. In such cases, typical intensity statistics calculations give aberrant results that cease to be useful for identifying centric zones or for diagnosing twinning.

## 2. Statistics of local intensity differences

Here, we demonstrate the advantages gained by analyzing intensity statistics in an alternate fashion, by examining local differences between pairs of reflections. The difference between two intensities is divided by their sum to give a unitless quantity, $L$, defined by

$$L \equiv \frac{I(\mathbf{h}_1) - I(\mathbf{h}_2)}{I(\mathbf{h}_1) + I(\mathbf{h}_2)}, \quad (1)$$

where $h_1$ and $h_2$ are unrelated reflections and the range for $L$ is $-1$ to 1. A similar parameter, $H$, was introduced earlier (Yeates, 1988) to describe the difference between two reflections related by a twin law. A different notation, $L$, is used here because the two reflections under consideration are not twin-related, but are proximally located in reciprocal space. We assume that two such reflections will have similar expected values and probability distributions. This will be the case even when the diffraction is anisotropic, owing to the proximity of the chosen reflections. Furthermore, if the two reflections are chosen so that the differences between their corresponding indices are all even, then they will have the same parity and will have similar expected intensity values in the presence of some of the most commonly encountered kinds of pseudo-centering, including cases where one unit-cell axis becomes doubled or one unit-cell face becomes centered. More unusual kinds of centering may require reflections to be chosen differently if they are to remain in the same class of expected intensity values. Finally, because dividing by the sum of the two intensities already effectively normalizes $L$, no further normalization (*e.g.* in resolution shells) is required; the assumption is avoided that reflections in the same resolution shell have similar expected values.

We assume that it is fair to treat the two reflections being compared as having arisen independently from the same Wilson distribution (see §2.5). The cumulative probability distribution, $N(L)$, of the parameter $L$ is derived in Appendix $A$ and also follows from expressions derived previously (Yeates, 1988). For acentric data (Table 1)

$$N(|L|) = |L|. \quad (2)$$

The cumulative probability distribution of a variable is the probability that it will take on a value less than some given value, and the derivative of the cumulative probability distribution is the ordinary probability density function. The expected values for $|L|$ and $L^2$ are (see Appendix $A$)

$$\langle |L| \rangle = 1/2 \quad (3)$$

and

$$\langle L^2 \rangle = 1/3. \quad (4)$$

For centric data (Table 1), $N(|L|) = (2/\pi)\arcsin(|L|)$, $\langle |L| \rangle = 2/\pi$ and $\langle L^2 \rangle = 1/2$, although for biological molecules the centric data are fewer in number and therefore less powerful diagnostically. These expressions can be derived in a similar manner to those in Appendix $A$ and also follow from earlier ones (Yeates, 1988) by letting the twin fraction $\alpha$ equal zero.

**Table 1**
Relations for $L$.

$L$ is defined as $[I(h_1) - I(h_2)]/[I(h_1) + I(h_2)]$ for unrelated reflections $h_1$ and $h_2$. $N(L)$ is the theoretical cumulative distribution of $L$, valid for $-1 \le L \le 1$. $N(|L|)$ is the theoretical cumulative distribution of $|L|$, for $0 \le |L| \le 1$. $\langle |L| \rangle$ and $\langle L^2 \rangle$ are the expected values of $|L|$ and $L^2$. More complex expressions are given in Appendix $C$ for the case of partial twinning.

| Type of reflection intensity | $N(L)$ | $N(\|L\|)$ | $\langle \|L\| \rangle$ | $\langle L^2 \rangle$ |
|---|---|---|---|---|
| Acentric, untwinned | $(L + 1)/2$ | $\|L\|$ | 1/2 | 1/3 |
| Centric, untwinned | $\arccos(-L)/\pi$ | $(2/\pi)\arcsin(\|L\|)$ | $2/\pi$ | 1/2 |
| Acentric, perfectly twinned | $(L + 1)^2(2 - L)/4$ | $\|L\|(3 - L^2)/2$ | 3/8 | 1/5 |

### 2.1. An example with anisotropic data

A diffraction data set showing significant anisotropy was selected from the PDB (Berman *et al.*, 2000). The structure is that of cyclophilin A ($P4_3$; $a = b = 54.2$, $c = 63.4$ Å; PDB code 1awu; Vajdos *et al.*, 1997), for which diffraction was stronger along the direction of the $c$ axis compared with other directions. The acentric intensity statistics for observed data between 3.5 and 2.5 Å were first examined in the usual way and were found to be significantly perturbed (Fig. 1*a*). Fig. 1(*a*) illustrates that the observed cumulative intensity distribution falls nearly midway between the curves expected for centric and acentric data. Also, a value of 2.54 was obtained for $\langle I^2 \rangle / \langle I \rangle^2$ over acentric data in the resolution range given above; the theoretical values for acentric and centric data are exactly 2 and 3, respectively. The statistics of the local intensity differences were then examined using the same data. The observed distribution of the parameter $L$ follows the theoretical distribution for acentric data much more closely (Fig. 1*b*). The observed values of $\langle |L| \rangle$ and $\langle L^2 \rangle$ are 0.52 and 0.36, which are much closer to the theoretical values for acentric data (1/2 and 1/3) than those for centric data ($2/\pi$ and 1/2) (Table 1). The results show that the problematic effects of anisotropy are diminished greatly by the present method. While it may be true that traditional methods will work if anisotropy can be modelled simply and corrected (Blessing, 1987), the present approach should also apply to various kinds of anisotropy that are hard to model because they go beyond simple elliptical corrections.

### 2.2. An example with pseudo-centering

When non-crystallographic symmetry causes molecules to pack in the unit cell with an approximate crystallographic translation, the diffraction pattern may display pseudo-centering, with strong and weak reflections in an alternating arrangement. A diffraction data set displaying significant pseudo-centering was selected from the PDB (Berman *et al.*, 2000). The structure is that of clam hemoglobin ($P2_12_12_1$; $a = 101.4$, $b = 93.8$, $c = 127.1$ Å; PDB code 1sct; Royer *et al.*, 1995). In this crystal, there are two $\alpha_2\beta_2$ tetramers in the asymmetric unit arranged to be nearly centered in the $ab$ plane, resulting in anomalously strong diffraction at reflections for which $h + k$ is even. Data were analyzed from 8 to 3.5 Å resolution. Traditional intensity statistics showed aberrant results (Fig. 1*c*), with a value for $\langle I^2 \rangle / \langle I \rangle^2$ of 2.73, in

disagreement with the theoretical value of 2. In contrast, when the statistics of the local intensity differences were examined, comparing nearby reflections of the same parity, the distribution of the parameter $L$ followed the theoretical curve very closely (Fig. 1*d*), with values for $\langle|L|\rangle$ and $\langle L^2\rangle$ of 0.49 and 0.32, again in excellent agreement with the theoretical values of 1/2 and 1/3 (Table 1). This example of pseudo-centering provides a particularly compelling illustration of the advantage of the local intensity statistic.

### 2.3. Application to data perfectly twinned by hemihedry

In certain space groups, a specimen may grow so that it is composed of distinct crystal domains whose orientations are

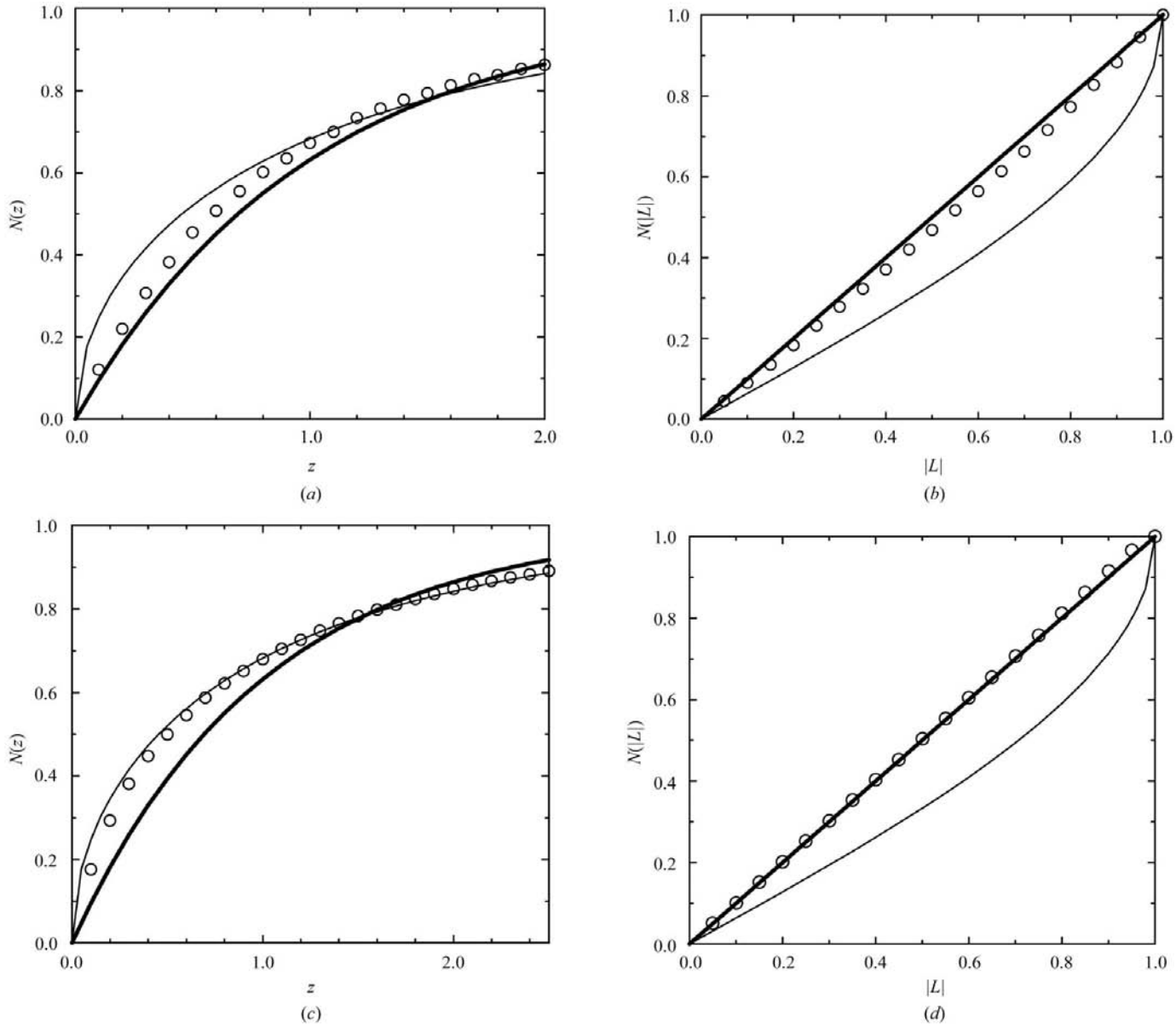


**Figure 1**
Robustness of local intensity difference statistics in the presence of anisotropic scattering and pseudo-centering. Theoretical distributions for acentric data are shown by bold curves, while those for centric data are shown by the thinner curves. Distributions for observed acentric data are shown by open circles. An example (PDB code 1awu) where anisotropic scattering was evident is shown in (*a*) and (*b*). The traditional method of analyzing the cumulative distribution of intensity values normalized in resolution shells, where $z = I/\langle I\rangle$, is shown in (*a*), where the observed distribution falls between the two theoretical curves. In (*b*), the local statistic $L = (I_1 - I_2)/(I_1 + I_2)$, where $I_1$ and $I_2$ are unrelated intensities, follows the acentric curve closely. An example (PDB code 1sct) where pseudo-centering was prominent is shown in (*c*) and (*d*). For the acentric data, the traditional cumulative intensity distribution nearly follows the theoretical curve for centric data owing to the effects of pseudo-centering (*c*). In (*d*), the local statistic follows the theoretical curve for acentric data very closely. In evaluating the local statistics in (*b*) and (*d*), two reflections were compared if they were of the same parity but were otherwise nearly adjacent (*i.e.* the difference between corresponding reflection indices was required to be 0 or 2). Every such pair of reflections was included.

crystallographically distinct, but whose reciprocal lattices are superimposed (Buerger, 1960; Donnay & Donnay, 1974; Yeates, 1997); this is referred to as hemihedral twinning. In cases where the two twin domains represent equal or similar volumes (*i.e.* the twin fraction $\alpha$ nearly equals 1/2), the observed diffraction pattern acquires an erroneous artificially high apparent symmetry. This pathological situation, referred to as perfect twinning, is generally discovered by analyzing intensity statistics, as they are distinctly affected. However, as discussed earlier, a number of effects (including anisotropy and pseudo-centering) can perturb traditional intensity statistics dramatically. These perturbations tend to counteract and mask the presence of twinning. On the other hand, the statistics of local intensity differences are relatively immune to the effects of anisotropy and pseudo-centering, while still being sensitive to the effects of twinning, as shown below. The statistics of the parameter $L$ can therefore be applied effectively to the problem of detecting perfect twinning. Twinning statistics may also be influenced by intensity measurement errors (Dumas *et al.*, 1999), but the effects of errors are not examined here.

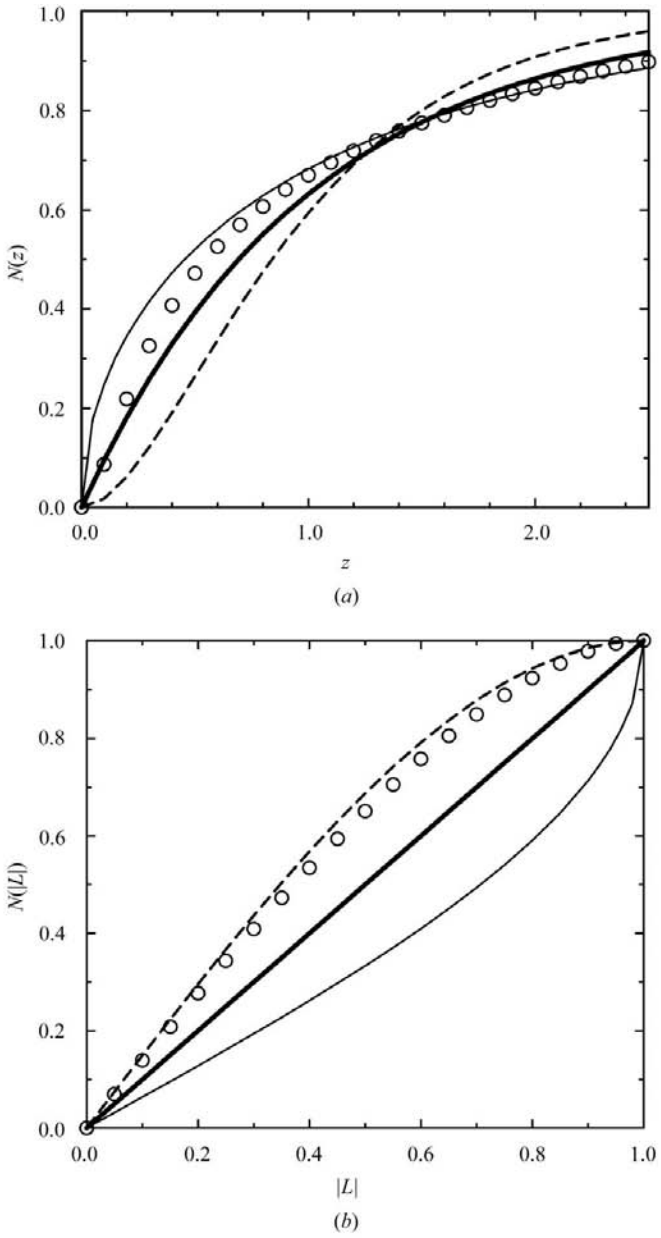


**Figure 2**
Utility of local intensity difference statistics in detecting perfect twinning in the presence of confounding factors. Theoretical distributions for acentric data are shown by bold curves, those for centric data are shown by the thinner curves and those for perfectly twinned (acentric) data are dashed. Distributions for observed acentric data are shown by open circles. The data are from a crystal of the VP5CT protein from rhesus rotavirus (Philip Dormitzer, personal communication). The specimen was highly merohedrally twinned and also displayed pseudo-centering. In (*a*), traditional intensity statistics, where $z = I/\langle I\rangle$, give no indication of twinning, as the effects of pseudo-centering dominate. In (*b*), the local statistic nearly follows the theoretical curve for twinned data, despite the presence of pseudo-centering.

It can be shown (see Appendix *B*) that when observed intensities arise from perfect twinning, the cumulative probability distribution of $L$ takes a form different from the case with ordinary intensities (Table 1). Specifically, for acentric reflections,

$$N(|L|) = |L|(3 - L^2)/2. \quad (5)$$

The expected values for $\langle|L|\rangle$ and $L^2$ are (see Appendix *B*)

$$\langle|L|\rangle = 3/8 \quad (6)$$

and

$$\langle L^2\rangle = 1/5. \quad (7)$$

It is worth re-emphasizing that the two reflections being compared here are locally proximal and not twin-related to each other. This is in contrast to earlier analyses (Dunitz *et al.*, 1972; Fisher & Sweet, 1980; Yeates, 1988), which applied to the situation of partial twinning rather than perfect twinning and which involved the comparison of twin-related reflections.

An example illustrates the interplay between twinning and other factors that can affect intensity statistics. Crystals of the VP5CT protein from rhesus rotavirus (Dormitzer *et al.*, 2001) were suspected of being twinned, but traditional intensity statistics were confounded by the presence of pseudo-centering (Philip Dormitzer, personal communication). As illustrated in Fig. 2(*a*), the observed intensity distribution was shifted away from rather than towards the theoretical curve expected for twinning. As another indicator, for acentric data between 20 and 3.5 Å resolution, $\langle I^2\rangle/\langle I\rangle^2$ was 2.37, which is closer to the theoretical value of 2 expected for ordinary data than it is to the theoretical value of 1.5 expected for perfectly twinned data. In contrast, the distribution of the local parameter $L$ gave a clear indication of twinning, as seen in Fig. 2(*b*). Also, the average values of $|L|$ and $L^2$ were 0.40 and 0.22, which are likewise consistent with significant twinning according to Table 1. The presence of a high degree of twinning was subsequently confirmed by other analyses, such as a finding that $\langle I^2\rangle/\langle I\rangle^2$ fell to 1.64 when the calculation included only reflections for which the indices were all even.

### 2.4. The case of partial twinning

Table 1 gives the theoretical cumulative distribution of $L$, along with expected values for $|L|$ and $L^2$, for ordinary data and for perfectly twinned data. Partial twinning is the intermediate situation in which the distinct twin domains are of significantly unequal volumes; the twin fraction $\alpha$ is less than 1/2. In the presence of partial twinning, the statistics for $L$ depend on $\alpha$, given by complicated expressions discussed in Appendix *C*. These expressions for the expected values of $|L|$ and $L^2$ were verified by comparison to numerical simulations of ideal Wilson-distributed data transformed by a twinning operation (Fig. 3). Owing to the shape of the curves, the parameter $L$ is not suited to giving accurate estimates of the twin fraction $\alpha$ for specimens that are highly twinned. Even in cases of partial twinning, where $\alpha$ is relatively small, experience has shown that other methods that directly compare twin-related reflections (Britton, 1972; Fisher & Sweet, 1980; Murray-Rust, 1973; Rees, 1982; Yeates, 1988) give more reliable estimates of $\alpha$ than methods such as the present one. The present method is especially relevant to the problem of detecting perfect twinning; it may also give an indication of partial twinning, but other methods are recommended for subsequently estimating the twin fraction.

### 2.5. On the independence of neighboring reflections

Our analysis assumes that the intensities of neighboring reflections are statistically independent. However, the atoms in a molecular structure are not distributed randomly, leading to potential correlations between the structure factors for distinct reflections. The limits of this assumption must therefore be examined. To do so, we analyzed a group of diffraction data sets extracted from the PDB (Berman *et al.*, 2000). The data sets were chosen as the first five from an alphabetical list satisfying certain criteria, such as a resolution of at least 2.2 Å, a protein with at least 200 amino acids with only one molecule

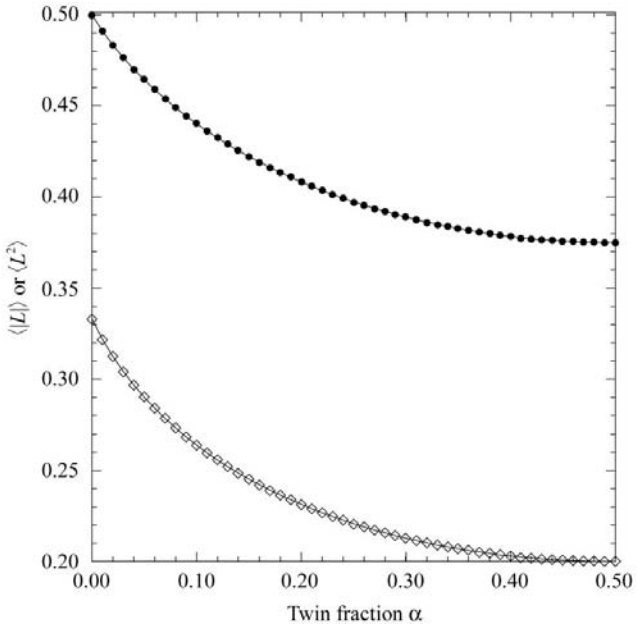


**Figure 3**
The dependence of the expected values of the local intensity statistic, $L$, on the twin fraction for partially twinned specimens. The expected value for $|L|$ is shown by filled dots. The expected value for $L^2$ is shown by unfilled diamonds. The data points (dots and diamonds) represent values obtained by numerical simulation. The curves are drawn according to the theoretical equations derived in Appendix *C*. Precise agreement is obtained between the theoretical equations and the simulated values.

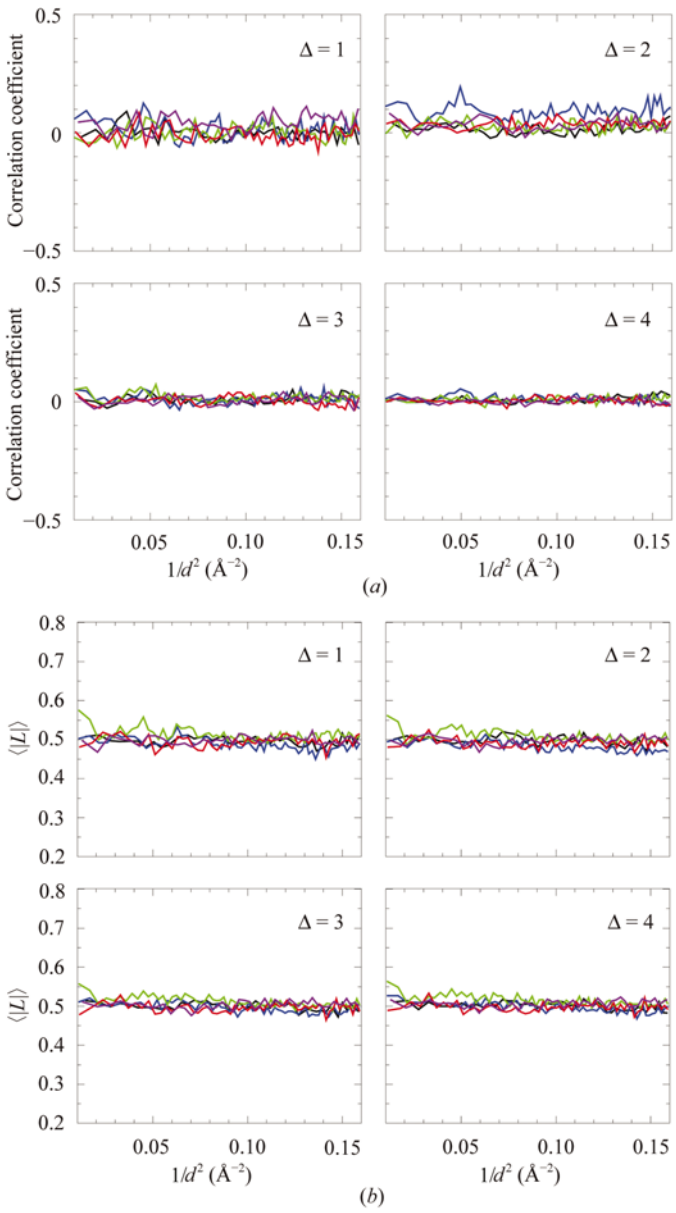


**Figure 4**
An examination of potential correlations between the intensities of neighboring reflections in macromolecular diffraction data. (*a*) The correlation coefficient is shown between pairs of neighboring reflection intensities as a function of resolution. The difference, Δ, is the sum of the absolute values of the differences between the three corresponding reflection indices for two reflections. Small but measurable correlations are evident when Δ is small, but vanish for values of Δ greater than 2. Five arbitrarily selected data sets were examined (1a48, black; 1a53, blue; 1a6q, green; 1a8f, red; 1a8q, purple). (*b*) The behavior of the parameter $L$ as a function of the distance between reflections, calculated over the same data sets as in (*a*). Deviations from the theoretical value ($\langle|L|\rangle = 0.5$) are small, especially for $\Delta > 2$.

per asymmetric unit to avoid potential pseudocentering and a unit cell typical of protein crystals (*i.e.* axes longer than 50 Å); the chosen data sets were 1a48, 1a53, 1a6q, 1a8f and 1a8q. We calculated the correlation coefficient between neighboring reflection intensities as a function of resolution and as a function of the difference between the indices of the two reflections being compared. As shown in Fig. 4(*a*), the correlations are generally weak, with values only occasionally exceeding 0.1 in magnitude. The small correlations evident in the case where the reflections are adjacent ($\Delta = 1$) essentially vanish as the reflections being compared have combined differences in their indices exceeding two ($\Delta > 2$).

The relative independence of neighboring reflection intensities over the 10–2.5 Å resolution range tested is borne out by the behavior of the parameter $L$ (Fig. 4*b*). Observed deviations from theoretical behavior ($\langle|L|\rangle = 0.5$) are small even for $\Delta = 1$. For reflection pairs with $\Delta > 2$, the variations from ideality are small enough to be attributed to random noise from finite sampling. The results suggest that an optimal strategy for evaluating the parameter $L$ may be to compare reflections which are relatively close together but whose indices differ (in total) by more than two.

## 3. Summary

The statistics of local intensity differences are insensitive to the effects of anisotropic scattering and, if reflection pairs are chosen appropriately, to the effects of pseudo-centering, effects which confound traditional intensity statistics. The parameter $L$ is more robust for discriminating between acentric and centric data and for detecting perfect twinning. Because this new statistic appears to perform as well or better than traditional intensity statistics, we suggest that the distribution of $L$ should supplement or perhaps supplant some of the intensity analyses in current use.

## APPENDIX *A*
## The statistics of local differences between unrelated acentric reflection intensities from an untwinned specimen

For acentric data from an untwinned specimen, the probability density function for intensities is

$$P(I) = k\exp(-kI), \tag{8}$$

where $k = 1/\langle I\rangle$. Denote two unrelated intensities by $I_1$ and $I_2$. Then from (1), $L \equiv (I_1 - I_2)/(I_1 + I_2)$. This leads to $I_2 = I_1(1 - L)/(1 + L)$, which sets the integration limit for evaluating the cumulative distribution for $L$, $N(L)$,

$$N(L) = \int_{I_1=0}^{I_1=\infty} \int_{I_2=I_1(1-L)/(1+L)}^{I_2=\infty} P(I_1, I_2)\,\mathrm{d}I_2\,\mathrm{d}I_1. \tag{9}$$

Substituting from (8),

$$N(L) = \int_{I_1=0}^{I_1=\infty} \int_{I_2=I_1(1-L)/(1+L)}^{I_2=\infty} k^2\exp[-k(I_1 + I_2)]\,\mathrm{d}I_2\,\mathrm{d}I_1, \tag{10}$$

which integrates to give

$$N(L) = (L + 1)/2. \tag{11}$$

Noting that $P(L) = [\mathrm{d}N(L)/\mathrm{d}L]$, we take the derivative of (11) to obtain the probability density function

$$P(L) = \tfrac{1}{2}. \tag{12}$$

From (12) we can integrate again to obtain the cumulative distribution of $|L|$,

$$N(|L|) = |L|. \tag{13}$$

From (12), the expectation values for $|L|$ and $L^2$ can be calculated,

$$\langle|L|\rangle = \int_{-1}^{0} -LP(L)\,\mathrm{d}L + \int_{0}^{1} LP(L)\,\mathrm{d}L, \tag{14}$$

yielding

$$\langle|L|\rangle = 1/2 \tag{15}$$

and

$$\langle L^2\rangle = \int_{-1}^{1} L^2P(L)\,\mathrm{d}L, \tag{16}$$

yielding

$$\langle L^2\rangle = 1/3. \tag{17}$$

## APPENDIX *B*
## The statistics of local differences between unrelated acentric reflection intensities from a perfectly twinned specimen

From the probability density function for intensities for acentric data from an untwinned specimen, (8), it can be shown that the probability density function for intensities from a perfectly twinned specimen is

$$P(I) = k'^2I\exp(-k'I), \tag{18}$$

where $k' = 2/\langle I\rangle$, (Rees, 1982; Stanley, 1972).

Denote two unrelated intensities, both perfectly twinned, by $I_1$ and $I_2$. Again from (1), the same integration limits as in Appendix *A* are found for evaluating the cumulative distribution for $L$, $N(L)$. Substituting from (18) into (9),

$$N(L) = \int_{I_1=0}^{I_1=\infty} \int_{I_2=I_1(1-L)/(1+L)}^{I_2=\infty} k'^4I_1I_2\exp[-k'(I_1 + I_2)]\,\mathrm{d}I_2\,\mathrm{d}I_1, \tag{19}$$

which integrates to give

$$N(L) = (L + 1)^2(2 - L)/4. \tag{20}$$

As in Appendix *A*, we take the derivative of (20) to obtain the probability density function,

$$P(L) = 3(1 - L^2)/4. \tag{21}$$

From here we can integrate to obtain

$$N(|L|) = |L|(3 - L^2)/2. \quad (22)$$

From (21), the expected values for $|L|$ and $L^2$ are found to be

$$\langle |L| \rangle = 3/8 \quad (23)$$

and

$$\langle L^2 \rangle = 1/5. \quad (24)$$

## APPENDIX C
## The statistics of local differences between unrelated acentric reflection intensities from a partially twinned specimen

For acentric data from a partially twinned specimen, the observed twinned intensities are the sum of intensities, $I_A$ and $I_B$, from each of the twin domains, weighted by the volume fraction of the twin domain,

$$I_{\rm twin} = (1 - \alpha)I_A + \alpha I_B. \quad (25)$$

From the probability density function for untwinned acentric reflections (8), we can write out a new density function for partially twinned reflections,

$$P(I_{\rm twin}) = \int_{I=0}^{I_{\rm twin}} \frac{k}{\alpha} \exp\left(\frac{-kI}{\alpha}\right) \frac{k}{(1-\alpha)} \exp\left[-\frac{k(I_{\rm twin} - I)}{(1-\alpha)}\right] {\rm d}I, \quad (26)$$

which integrates to

$$P(I_{\rm twin}) = \frac{k}{(1-2\alpha)} \left\{ \exp\left[-\frac{k}{(1-\alpha)} I_{\rm twin}\right] - \exp\left(-\frac{k}{\alpha} I_{\rm twin}\right) \right\}. \quad (27)$$

Now we let $I_1$ and $I_2$ represent two intensities that are each partially twinned, but not twin-related to each other, each following the probability density in (27). The integration limit for the cumulative distribution function is set up as in Appendices *A* and *B*. Substituting into (9), we obtain

$$N(L) = \int_{I_1=0}^{\infty} \int_{I_2=\frac{(1-L)}{(1+L)} I_1}^{\infty} \frac{k^2}{(1-2\alpha)^2} \left\{ \exp\left[-\frac{k}{(1-\alpha)} I_1\right] - \exp\left(-\frac{k}{\alpha} I_1\right) \right\} \times \left\{ \exp\left[-\frac{k}{(1-\alpha)} I_2\right] - \exp\left(-\frac{k}{\alpha} I_2\right) \right\} {\rm d}I_2\, {\rm d}I_1, \quad (28)$$

which evaluates to

$$N(L) = \frac{1}{(1-2\alpha)^2} \left\{ \frac{[\alpha^2 + (1-\alpha)^2](1+L)}{2} - \frac{\alpha(1-\alpha)^2(1+L)}{[1+(1-2\alpha)L]} - \frac{\alpha^2(1-\alpha)(1+L)}{[1-(1-2\alpha)L]} \right\}. \quad (29)$$

Taking the derivative, using *Mathematica* (Wolfram Research Inc., 1999), we obtain the probability density function

$$P(L) = \frac{1}{(1-2\alpha)^2} \left\{ \frac{\alpha^2 + (1-\alpha)^2}{2} - \frac{\alpha(1-\alpha)^2}{[1+(1-2\alpha)L]} - \frac{\alpha^2(1-\alpha)}{[1-(1-2\alpha)L]} + \frac{\alpha(1-\alpha)^2(1-2\alpha)(1+L)}{[1+(1-2\alpha)L]^2} - \frac{\alpha^2(1-\alpha)(1-2\alpha)(1+L)}{[1-(1-2\alpha)L]^2} \right\}. \quad (30)$$

This can be used to obtain expressions for the expected values $\langle |L| \rangle$ and $\langle L^2 \rangle$ as functions of the twin fraction $\alpha$,

$$\langle |L| \rangle = \frac{(1-2\alpha)^2(1-6\alpha+6\alpha^2) - 8(1-\alpha)^2\alpha^2 \ln[4\alpha(1-\alpha)]}{2(1-2\alpha)^4} \quad (31)$$

$$\langle L^2 \rangle = \frac{(1-2\alpha)(1-12\alpha-4\alpha^2+32\alpha^3-16\alpha^4) + 24(1-\alpha)^2\alpha^2 \ln\left(\frac{1-\alpha}{\alpha}\right)}{3(1-2\alpha)^5}. \quad (32)$$

Plots of these expected value functions *versus* $\alpha$ are found in Fig. 3.

The authors thank Rob Grothe and Dick Marsh for helpful comments, and Philip Dormitzer for valuable contributions including the diffraction data from VP5CT crystals. This work was supported by the NSF (DBI 98-07896) and the NIH (GM 31299).

## References

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. (1987). *Crystallogr. Rev.* **1**, 3–58.
Britton, D. (1972). *Acta Cryst.* A**28**, 296–297.
Buerger, M. J. (1960). *Crystal-structure Analysis.* New York: Wiley.
Donnay, G. & Donnay, J. D. H. (1974). *Can. Mineral.* **12**, 422–425.
Dormitzer, P. R., Greenberg, H. B. & Harrison, S. C. (2001). *J. Virol.* **75**, 7339–7350.
Drenth, J. (1999). *Principles of Protein X-ray Crystallography*, 2nd ed. New York: Springer.
Dumas, P., Ennifar, E. & Walter, P. (1999). *Acta Cryst.* D**55**, 1179–1187.
Dunitz, J. D., Gehrer, H. & Britton, D. (1972). *Acta Cryst.* B**28**, 1989–1994.
Fisher, R. G. & Sweet, R. M. (1980). *Acta Cryst.* A**36**, 755–760.
Giacovazzo, C. (1992). *Fundamentals of Crystallography.* IUCr/Oxford University Press.
Murray-Rust, P. (1973). *Acta Cryst.* B**29**, 2559–2566.
Rees, D. C. (1982). *Acta Cryst.* A**38**, 201–207.
Royer, W. E. Jr, Heard, K. S., Harrington, D. J. & Chiancone, E. (1995). *J. Mol. Biol.* **253**, 168–186.
Stanley, E. (1972). *J. Appl. Cryst.* **5**, 191–194.
Vajdos, F. F., Yoo, S., Houseweart, M., Sundquist, W. I. & Hill, C. P. (1997). *Protein Sci.* **6**, 2297–2307.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wolfram Research Inc. (1999). *Mathematica*, version 4. Wolfram Research Inc., Champaign, Illinois, USA.
Yeates, T. O. (1988). *Acta Cryst.* A**44**, 142–144.
Yeates, T. O. (1997). *Methods Enzymol.* **276**, 344–358.