Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Intensity statistics in twinned crystals with examples from the PDB

Andrey A. Lebedev,* Alexei A. Vagin and Garib N. Murshudov

Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5YW, England

Correspondence e-mail: lebedev@ysbl.york.ac.uk

Received 7 November 2005
Accepted 8 November 2005

Entries deposited in the Protein Data Bank as of February 2004 for which both model and X-ray data were available were analysed to identify cases of twinning using such simple statistics as the R factor between potential twin-related reflections. Careful consideration of all identified twins showed that in many cases twinning was ignored during structure solution and refinement. Manual analysis of the models showed that twinning often occurs in association with rotational pseudosymmetry parallel to the twinning operator. The coexistence of these two phenomena complicates the detection and diagnostics of twinning using currently available twinning tests. It was concluded that a twinning-detection step should be incorporated in every stage of structure analysis from data acquisition to refinement and validation.

1. Introduction

The Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2002) is a rich source of biological, biochemical and structural information. It also offers templates for the determination of new structures by molecular replacement. The huge number of models with experimental X-ray data provides numerous training cases of varying difficulties useful to both the practical crystallographer and software developers. These cases should be analysed before approaching real-life difficult cases and, in an ideal world, all new software should be tested against them before general release.

However, one should be careful when extracting information from the PDB because of several problems, some of which have been described by Kleywegt (1999, 2000). Currently, a new entry goes through a careful validation procedure during deposition. Nevertheless, at least one potential problem, twinning (Giacovazzo *et al.*, 1992), has not yet been addressed. Twinning tests should be included in the validation routine and, when twinning is present, the structure-factor analyses need to be adjusted accordingly. If twinning is not taken into consideration during refinement, the resulting model will inevitably be degraded. Therefore, during deposition, it is important to notify the depositor if this is the case.

The twinning phenomenon in crystals has been recognized for a long time (Friedel, 1926). For small-molecule structures, data collection and processing, structure solution and refinement against data from twinned crystals are routine (Sheldrick & Schneider, 1997). However, the situation with macro-

molecules is not yet so straightforward; the software used for data-acquisition and structure-solution procedures have not addressed this problem fully. For example, it is particularly difficult to solve a twinned structure using experimental phasing (Dauter, 2003; Rudolph *et al.*, 2003).

The phenomenon of twinning should be considered as a special case of crystal intergrowth. Crystal clusters are often observed, but usually it is possible either to optimize crystallization to grow a single crystal or to break off a single-crystal fragment. In some cases this simple approach does not work and one has to deal with diffraction data from an intergrown crystal where the diffraction patterns of two or more fragments overlap. If the fragments are orientated in a random manner relative to each other the two lattices can be identified from the first images. [A very interesting case of treatment of such data has been reported by Dauter *et al.* (2005) in which the intergrown domains have different space groups.] However, diffraction patterns from such intergrown crystals are often deceptive; if the diffraction spots from the two (or more) crystal domains completely overlap, the diffraction pattern will appear normal on initial inspection. In this case the measured observation at a given reciprocal-lattice point is in fact the sum of the twinned sets of intensities, weighted by the relative volumes ('twinning fractions') of the different components. This is called (pseudo)merohedral twinning and the term 'twinning' in macromolecular crystallography usually refers to this. The most common case seen in macromolecular crystallography is hemihedral twinning, in which there are only two crystal components related by a twofold operator. However, the situation can be more complicated, as demonstrated by Barends *et al.* (2005).

For two or more lattices to overlap completely, the unit-cell parameters and crystal symmetry must possess some special relationships. The unit-cell parameters must allow the possibility of higher symmetry than the crystal actually shows. This is most common in tetragonal, trigonal or cubic crystal classes, where the twinning operator will be one of the symmetry operators of the supergroup. However, it is also possible for triclinic, monoclinic and orthorhombic crystals when the unit-cell parameters possess some special properties (Giacovazzo *et al.*, 1992). A technique for identifying data sets where the unit-cell parameters and space group can allow (pseudo)-merohedral twinning and finding the possible twinning operators is described by Flack (1987).

In these cases, to detect whether twinning has occurred requires statistical analysis of the whole data set. When a problem is detected two options are open: (i) to discard the data set and try to obtain a new untwinned crystal or (ii) try to solve and refine the structure using the twinned data. While the first option seems to offer better data, it may turn out to be time-consuming (or even impossible). Moreover, structural genomics imposes strong constraints on the time spent on an individual protein and option (ii), *i.e.* using what is at hand, is becoming more common, with new software being developed to meet this demand.

This contribution analyzes the PDB to find out how often such a problem occurs and to generate ideas for the future automatic treatment of structures using data from twinned crystals. We also describe the major difficulties we faced in the identification of twinning in special cases using the widely available twinning tests.

2. Materials and methods

The PDB February 2004 release containing about 22 000 structures was screened and only those entries where both coordinates and structures factors had been deposited (11 367 entries) were used in the analysis. The unit-cell parameters and space group of these entries were analysed using the technique described in Appendix *A*. If (pseudo)merohedral twinning was possible then this data set was selected for further analysis. 5% deviation from ideal twinning constraints was allowed. This threshold is consistent with Mallard's rule as cited by Grimmer (2003). If observed intensities were present in a CIF file they were used directly and for other applications they were converted to structure factors using *TRUNCATE* (French & Wilson, 1978). If only observed structure amplitudes were available, estimates of the corresponding intensities were generated, although some information must be lost.

Thus, in all selected cases there is at least one potential twinning operator. $R_{\rm twin}$, defined in (1), was calculated with respect to each operator for both observed intensities and those calculated from the atomic model. The matrix for a potential twinning operator, selected from the coset of equivalent operators, and the associated $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ were calculated using a program written by one of us (AAL).

The distribution of $R_{\rm twin}^{\rm obs}$ against $R_{\rm twin}^{\rm calc}$, referred to as an RvR plot and discussed below, can give a clear indication of twinning. Detailed analysis was carried out for all likely twinned structures. The analysis involved estimation of the likely number of molecules in the asymmetric unit using *SFCHECK*, self-rotation function (Rossmann & Blow, 1962) as implemented in *MOLREP* (Vagin & Teplyakov, 1997), twinning tests based on overall reflection statistics, namely cumulative distribution of normalized intensity and moments of acentric reflections (Rees, 1980) as implemented in *TRUNCATE*, and *H*-tests (Yeates, 1997) as implemented in *SFCHECK* (Vaguine *et al.*, 1999). If the interpretation of these tests was ambiguous, then molecular replacement using *MOLREP* and refinement using *REFMAC* (Murshudov *et al.*, 1997) were carried out using the model from the PDB without substrates and with all atomic displacement parameters (ADP) reset to equal values. The models, Patterson and electron-density maps were visualized using *Coot* (Emsley & Cowtan, 2004). Further statistical analyses of the results were performed using the statistical package *R* (R Development Core Team, 2004). Some figures in this paper are based on those generated from *CCP*4 software (Collaborative Computational Project, Number 4, 1994).

3. RvR plot

Detection of twinning should ideally be performed at the stage of data acquisition before the crystal structure is known. This

task is not always trivial; for example, perfect twinning cannot be detected from merging statistics. In some instances, even finer (than $R_{\rm merge}$) statistical properties of the data are too ambiguous for assignment of crystal symmetry and detection of twinning prior to the structure determination.

Therefore, we undertook an investigation of all possible twinning cases, known or undetected, deposited in the PDB. The goal of the work was to understand the symmetry environments most frequently accompanying twinning and to pinpoint problems with its detection. Since for these data sets

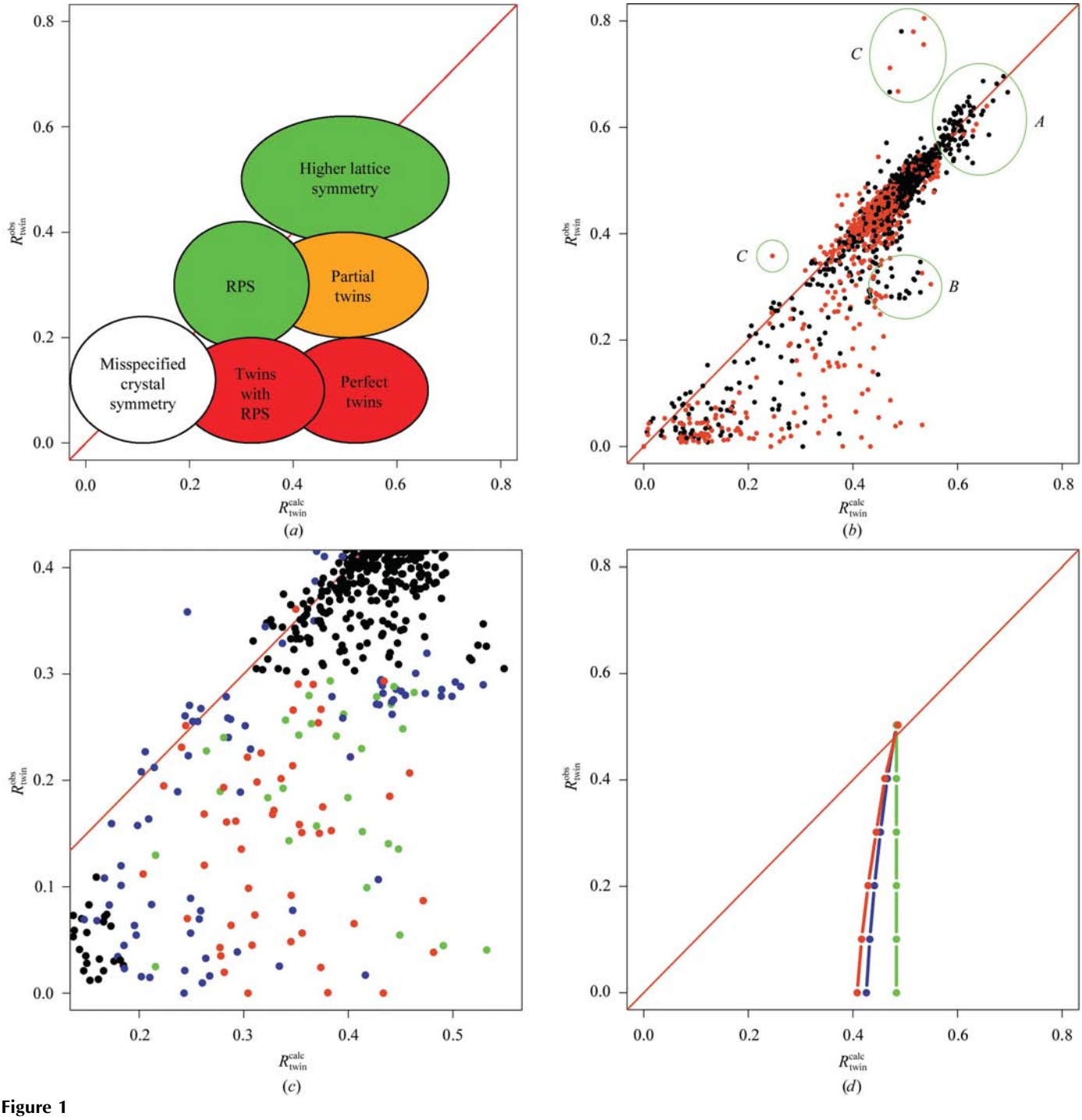

Figure 1
(*a*) Schematic view of RvR scatter plot. (*b*) Observed RvR scatter plot: red, (potential) merohedral twins; black, (potential) pseudomerohedral twins. Green ovals show the area populated by cases with TNCS (labelled *A*) and the areas corresponding to mislabelled data (labelled *B* and *C*). (*c*) Observed RvR scatter plot (enlargement of Fig. 1*b*): black, known to be untwinned and not analysed; blue, found to be untwinned after further analysis; green, twins without RPS; red, twins with some degree of RPS, but the difference between the twinning and RPS operators may be quite large. (*d*) Middle blue curve, results after refinement of PDB entry 1nqh, performed without taking twinning into consideration, against simulated data sets with twinning fractions in the range 0–0.5 with standard restraints on ADPs. Left red curve, the same calculations with relaxed restraints on the ADPs. Right green curve, results before refinement. Here $R_{\rm twin}^{\rm calc} \simeq 0.5$. It is expected that proper twin refinement would preserve this value.

both the atomic model and the experimental data are available, the analysis is considerably simplified.

3.1. *R* factor with respect to twinning operator

Let us assume that in a given crystal the combination of lattice and crystal symmetries allows twinning. This means that there is at least one potential twinning operator $S_{\rm twin}$. It can be determined using, for example, the technique described in Appendix *A*.

Let $R_{\rm twin}$ be the intensity-based *R* factor between reflections related by potential twinning operator $S_{\rm twin}$,

$$R_{\rm twin} = \frac{\sum_{\mathbf{h}} |I_{\mathbf{h}} - I_{\mathbf{h}'}|}{\sum_{\mathbf{h}} (I_{\mathbf{h}} + I_{\mathbf{h}'})}, \qquad (1)$$

where summation is over all unique reflections $\mathbf{h}$, such that intensities for both $\mathbf{h}$ and $\mathbf{h}' = S_{\rm twin}\mathbf{h}$ have been measured and $\mathbf{h} \neq \mathbf{h}'$. The definition of $R_{\rm twin}$ (1) is similar to that of $R_{\rm sym}$, except that $S_{\rm twin}$ is not an operator of the crystal point group, but belongs to the point group of the crystal lattice.

$R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ are $R_{\rm twin}$ calculated using observed intensities and (untwinned) intensities derived from the atomic model, respectively. The relationship between the two magnitudes is as follows (see also Appendix *B*)

$$\begin{aligned} R_{\rm twin}^{\rm obs} &\simeq R_{\rm twin}^{\rm calc} \quad \text{(no twinning)},\\ R_{\rm twin}^{\rm obs} &< R_{\rm twin}^{\rm calc} \quad \text{(partial twinning)},\\ R_{\rm twin}^{\rm obs} &\simeq 0 \quad \text{(perfect twinning)}. \end{aligned} \qquad (2)$$

The approximation sign in the equations above is a consequence of model and experimental errors. Our experience shows that in the majority of the cases these errors do not affect the qualitative conclusions.

If the crystal symmetry has been misspecified[1] then the analysis of unit-cell parameters and space group identifies missing elements of the point group of the crystal as twinning operators. In this case it is expected that both

$$\begin{cases} R_{\rm twin}^{\rm obs} \simeq 0 \\ R_{\rm twin}^{\rm calc} \simeq 0 \end{cases} \quad \text{(misspecified crystal symmetry)}. \qquad (3)$$

Note that a small value of $R_{\rm twin}^{\rm obs}$ can be misinterpreted, as this takes place in two different cases; see (1) and (2). In particular, false positives in detection of twinning can be found in some PDB entries with misspecified symmetry (see §3.4.2).

3.2. Twinning interfering with NCS

Let a crystal or an individual crystal of a twin possess noncrystallographic symmetry (NCS) and let one of the NCS operators be such that its rotational component is approximately equal to the (potential) twinning operator. In this case, the NCS could interfere with twinning and is further referred to as rotational pseudosymmetry (RPS). There are two reasons why twins with RPS are of special interest.

Firstly, a correlation between observations related by potential twinning operators could be caused either by RPS or by both RPS and twinning. The two cases cannot be discriminated by $R_{\rm twin}^{\rm obs}$ alone. These cases are particularly difficult for twinning detection prior to the structure determination.

Secondly, we expect a relatively high frequency of twins with RPS because of high likelihood of the following two mechanisms of their formation. The first mechanism assumes a change of crystal symmetry (we are interested in symmetry reduction) which is sometimes observed during crystallization, seeding, soaking, fast cooling and even data collection. It is physically reasonable to expect that this transition starts simultaneously in several areas of the crystal. As a result, several identical domains are formed in two or more different orientations related by the broken symmetry element and thus the crystal becomes twinned and the broken symmetry element becomes a twinning operator. At the same time it becomes an RPS operator, relating molecules in the asymmetric unit which were equivalent by crystal symmetry before the transition. In the second mechanism an individual crystal is formed by tightly packed molecular layers with symmetry that is higher than that of the interfaces between them. In these structures the whole layers are (approximately) invariant with respect to NCS operators. Consequently, any NCS operator in the layer at a twinning interface relates two twinning domains and thus the NCS is RPS. The high frequency of twinning interfaces in such symmetry environments leads to statistical crystals (Bragg & Howells, 1954; Cochran & Howells, 1954).

If there is no NCS, no pronounced anisotropy and no serious experimental errors, the expected value of $R_{\rm twin}^{\rm calc}$ can be estimated to be 0.5 as shown in Appendix *B*. However, when RPS is present, the correlation between related reflections causes a decrease in $R_{\rm twin}^{\rm calc}$. Thus, in addition to (2) the following holds,

$$\begin{aligned} R_{\rm twin}^{\rm calc} &\simeq 1/2 \quad \text{(no RPS)},\\ R_{\rm twin}^{\rm calc} &< 1/2 \quad \text{(RPS)}. \end{aligned} \qquad (4)$$

Note that even if RPS is present, (2) holds. Thus, despite the similar effects of RPS and twinning on $R_{\rm twin}^{\rm obs}$, the availability of a crystal model in principle allows us to distinguish between twinning, RPS and twinning interfering with RPS using such simple statistics as $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$.

The relations (2)–(4) are illustrated in Fig. 1(*a*). The figure shows areas corresponding to different combinations of RPS and twinning.

3.3. Selection of twinning cases

The simplest possible way to select twinning cases from the PDB would be to extract the relevant information from the PDB headers and/or related papers. However, this approach is not sufficient because the researchers depositing data and/or writing papers either may have not noticed or not discussed twinning (false negatives) or may have misinterpreted higher crystal symmetry as twinning (false positive). Therefore, it was

[1] We say that crystal symmetry is misspecified when the space group reported in the PDB file is a subgroup of the true space group of the crystal, *e.g.* *P*4 instead of *P*422. Accordingly, in such cases the PDB file contains more molecules than should be in the asymmetric unit of the crystal, but some of these molecules are actually related by the missing symmetry operator(s).

decided to analyse PDB entries directly. This direct approach may also lead to a better understanding of the problems with the detection of twinning.

We analysed unit-cell parameters and the reported crystal symmetry of 11 367 entries present in the PDB at February 2004 containing both an atomic model and X-ray data. Entries where twinning is impossible or where the data were corrupted and unreadable by our software were rejected from further consideration. For the remaining 4086 entries, potential twinning operators were determined and $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ were computed. If there were two or more (non-equivalent) potential twinning operators (as, for example, in $P3$), then that which gave the lowest value of $R_{\rm twin}^{\rm obs}$ was chosen. Thus, each selected entry was characterized by only two quantities, $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$, and the corresponding point was drawn on the plot of $R_{\rm twin}^{\rm obs}$ *versus* $R_{\rm twin}^{\rm calc}$ (RvR plot; see Fig. 1*b*).

For each structure represented in Fig. 1(*b*), we analysed whether the twinning, if present, is merohedral or pseudomerohedral using the technique described in Appendix *A*. The points in Fig. 1(*b*) are coloured according to the results of this analysis.

All cases belonging to 'twinning areas' in the RvR plot (Fig. 1*a*) were analysed in detail to validate the presence or absence of twinning and to characterize the NCS if present. The specific areas and some peculiarities of the RvR plot are discussed below.

Table 1
Frequency of twinning in different symmetry environments.

Crystal symmetry or type of merohedry	No. of twins					
	Total	RPS	TNCS	RPS + TNCS	DNA	Detwinned
$P1$	1	—	—	—	1	—
$P2_1$	13	11	3	3	—	—
$C2$	1	—	—	—	—	1
$P2_12_12$	1	—	—	—	—	—
$P2_12_12_1$	2	2	—	—	—	—
$C222_1$	1	1	1	1	—	—
Pseudomerohedral total	19	14	4	4	1	1
$P4_1$	4	1	1	—	—	—
$P4_3$	6	4	1	1	—	—
$P4_2$	1	1	1	1	—	—
$I4$	3	2	—	—	—	—
$P3_1$	7	5	—	—	—	1
$P3_2$	6	4	2	1	—	—
$P321$	2	2	—	—	—	—
$P3_121$	2	—	—	—	—	—
$P3_221$	1	1	—	—	—	—
$P3_212$	2	1	—	—	1	—
$H3$	15	3	1	1	2	—
$P6_3$	6	3	—	—	—	1
$P6_5$	3	3	2	2	—	—
$I2_13$	1	—	—	—	1	—
Merohedral total	59	30	8	6	4	2
Total	78	44	12	10	5	3

3.4. Observed RvR plot

3.4.1. Main cluster. A large cluster around (0.5, 0.5) corresponds to untwinned crystals with no pronounced pseudosymmetry. However, twinning is not forbidden by the unit-cell parameters and space group and could occur for related crystals. Some of these points correspond to data sets which were detwinned before deposition. These cases were not included in further analysis. In particular, all untwinned crystals in space groups $P3_x$, $P3_x21$, $P3_x12$, $P4_x$, $I4_x$, $P6_x$, $P2_x3$, $I2_x3$ and $F23$ belong to this area. Since the lattices of these space groups have higher rotational symmetry than that of crystals, no extra constraints on the unit-cell parameters are needed for twinning to occur (Giacovazzo *et al.*, 1992; Schlessman & Litvin, 1995).

3.4.2. Misspecified crystal symmetry. The cluster at the origin corresponds to structures in which the crystal symmetry is misspecified and is actually higher than that used in the refinement and reported in the PDB entry. Thus, both $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ are expected to be close to 0.0. Several randomly chosen cases from this cluster have been successfully refined in the higher symmetry space group. It is interesting to note that for two of them twinning was reported, presumably on the basis of the low $R_{\rm twin}^{\rm obs}$; these are examples of false positives.

The first reliable case of twinning has $R_{\rm twin}^{\rm calc} = 0.2$. At the same time, in some cases where the space group was misspecified $R_{\rm twin}^{\rm calc}$ goes up to 0.3; these are mainly low-resolution structures where it is easy to overfit the model and to generate significant differences between independently modelled symmetry-related molecules.

3.4.3. RPS. The lower tail of the main cluster corresponds to untwinned crystals with RPS. Most points in this area are located on the diagonal, with $R_{\rm twin}^{\rm calc} \simeq R_{\rm twin}^{\rm obs}$ in the range 0.35–0.4. This tail extends along the diagonal down to about 0.2. Here we find one of the most extreme examples, the untwinned 1i1j (Lougheed *et al.*, 2001), in which the root-mean-square deviation of C^{α} atoms from the positions corresponding to higher crystal symmetry is about 0.15 Å.

3.4.4. Translational noncrystallographic symmetry (TNCS). The main cluster has an upper diagonal tail around (0.6, 0.6) corresponding to structures with TNCS, in which the set of TNCS vectors and consequently the modulation of intensities in the reciprocal space caused by TNCS are not invariant with respect to $S_{\rm twin}$. Thus, the intensities related by $S_{\rm twin}$ are modulated differently and the assumptions required for the relation (10) in Appendix *B* to be valid are violated. The numerator in (1) for both $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ increases, increasing their values.

Among such structures we observed no cases of twinning (note the empty area below this cluster in the RvR plot).

3.4.5. Mislabelled and corrupt data. There are some extra features on the RvR plot arising from mislabelling of columns in the CIF file. Two small clusters shown by circles in Fig. 1(*b*), which are just above and below the main cluster, are worth mentioning. In the first one, at about (0.5, 0.4) on the RvR plot, the structure amplitudes are present in the CIF file but are labelled as intensities. In the second one, at about (0.5, 0.6), the intensities are labelled as structure amplitudes.

These peculiarities may in theory be identified by simple statistical techniques. However, if such factors as twinning, pseudosymmetry or anisotropy affect the data or several deposition inaccuracies (for example, deposition of the

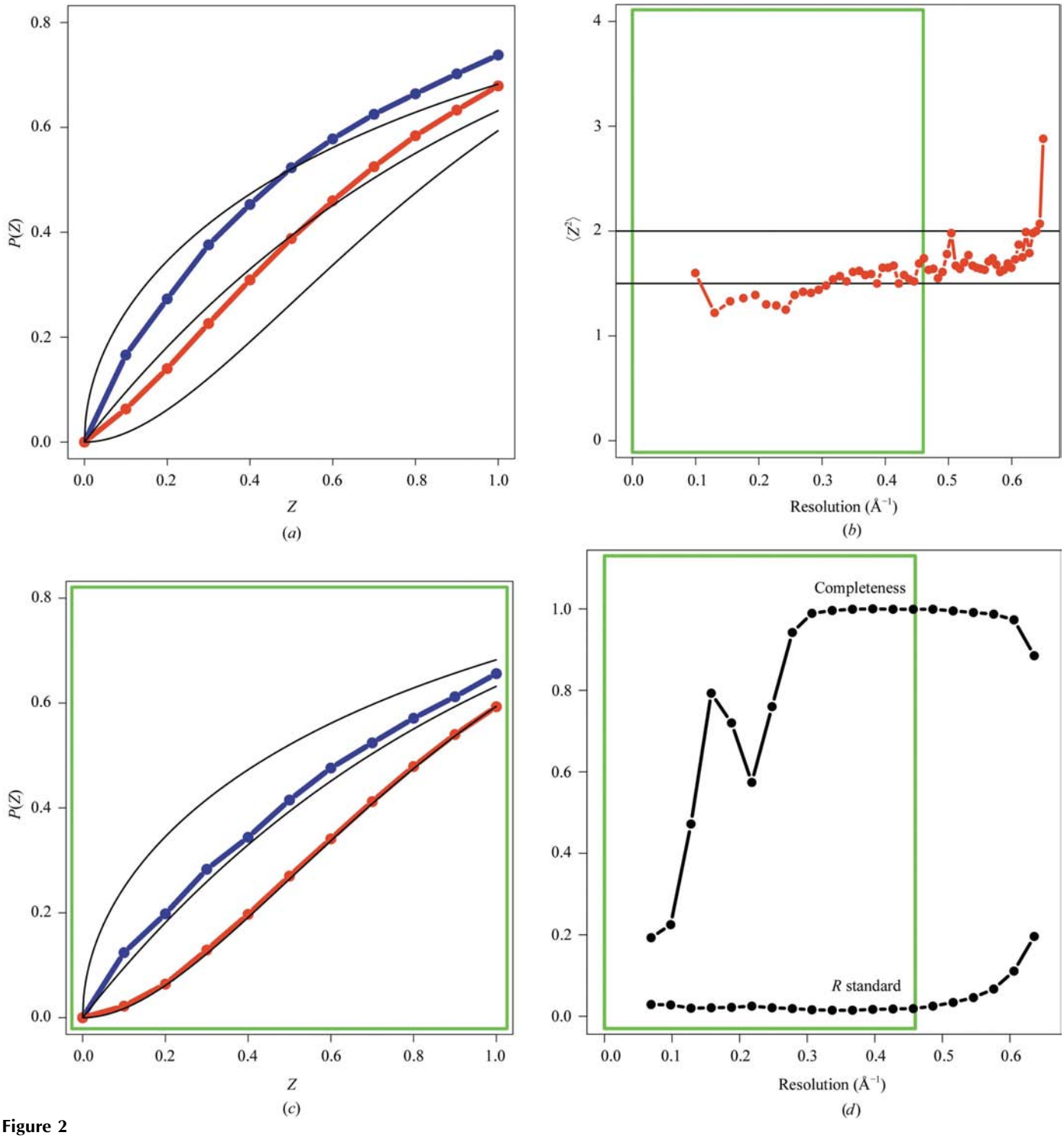

Figure 2
Effect of the resolution cutoff on the experimental cumulative distributions of Z. The plots were produced using X-ray data from PDB entry 1l2h (Rudolph *et al.*, 2003). (*a*) Experimental cumulative distributions of Z for all the data, resolution range 18.6–1.54 Å, (*b*) the experimental second moment of Z for acentric reflections *versus* resolution, (*c*) experimental cumulative distributions of Z in the resolution range 18.6–2.20 Å, (*d*) completeness and R standard *versus* resolution. The resolution range used in (*c*) is shown by green boxes in (*b*) and (*d*). The thick blue and red curves correspond to centric and acentric reflections, respectively. The thin black curves in (*a*) and (*c*) are the reference (theoretical) cumulative distributions of Z corresponding to (top curves) centric reflections in the untwinned case, (middle curves) centric reflections in the perfectly hemihedrally twinned case (*i.e.* with two twinning fractions) and acentric reflections in the untwinned case and (bottom curves) acentric reflections in the perfectly hemihedrally twinned case. The thin black curves in (*c*) are theoretical second moments of Z for acentric reflections *versus* resolution for (top line) the untwinned and (bottom line) perfectly hemihedrally twinned cases.

detwinned instead of the measured data) are present simultaneously then such analysis becomes complicated, if possible.

3.4.6. Areas of the RvR plot indicating twinning is likely. The points below the diagonal should, at least in theory, correspond to twins. In particular, points that deviate from the diagonal with $R_{\rm twin}^{\rm calc}$ significantly less than 0.5 should correspond to twins with RPS (see §3.2) and with an adequate refinement protocol, the deviation from the diagonal should correlate with twinning fraction.

The cases with $R_{\rm twin}^{\rm calc}$ in the range 0.2–0.6 and $R_{\rm twin}^{\rm obs}$ in the range 0.0–0.3, as well as some randomly chosen cases from other areas, were further investigated (coloured circles in Fig. 1*c*). The protocol of analysis included validation of the model, various twinning tests performed with both observed

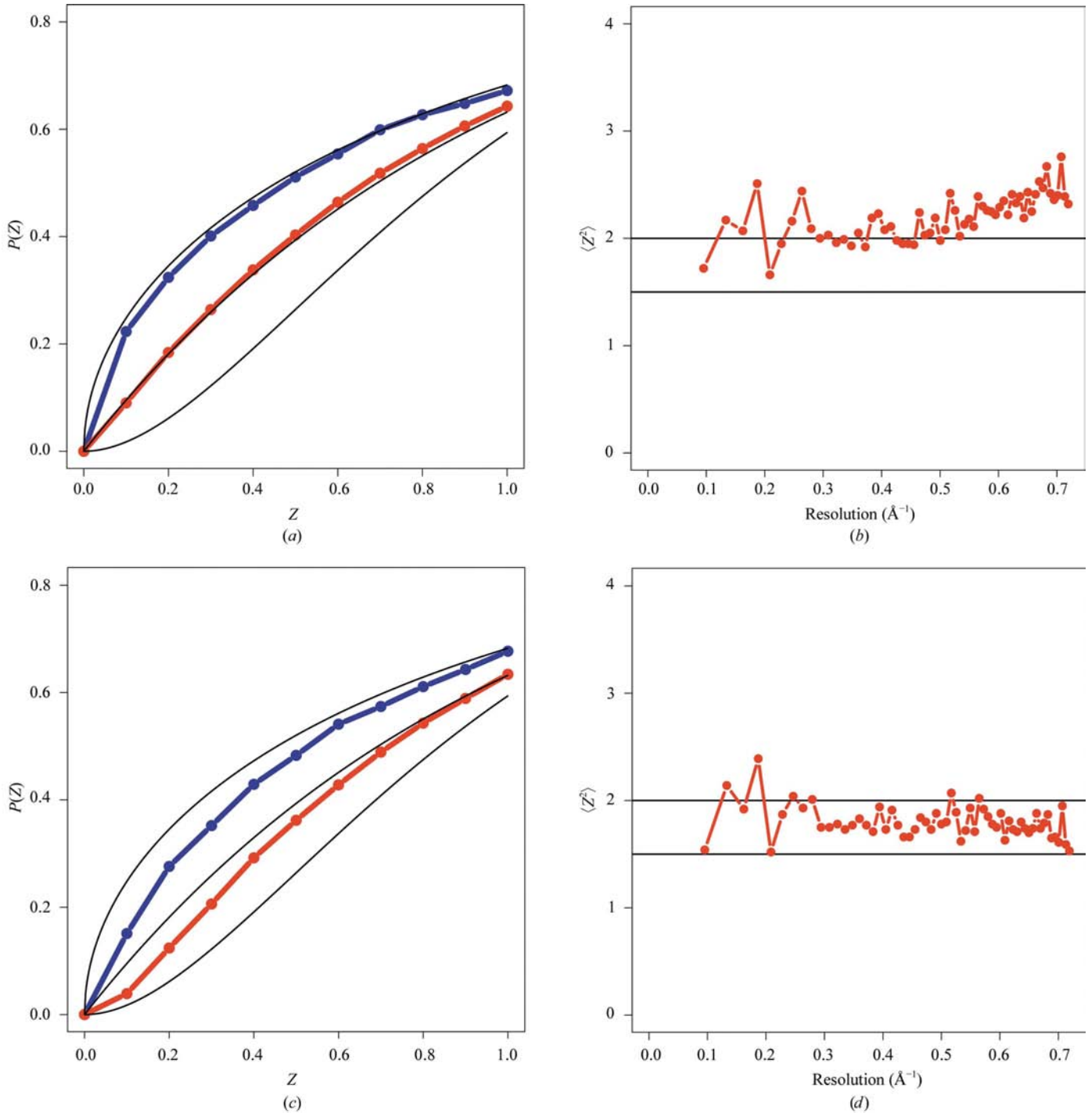

Figure 3
The effect of RPS on the perfect twinning tests. The plots were drawn using X-ray data from PDB entry 1i1j (Lougheed *et al.*, 2001). (*a*) and (*c*) Sample cumulative distribution of Z and (*b*) and (*d*) second moment of Z for acentric reflections *versus* resolution for (*a*) and (*b*) original data and (*c*) and (*d*) data with simulated hemihedral twinning. The colour legend is the same as for the similar plots in Fig. 2.

and calculated intensities with different resolution cutoffs and characterization of the NCS. In particular, NCS operators if present were compared with potential twinning operators to identify RPS. If there was spatial pseudosymmetry, attempts were made to refine structures in the corresponding higher symmetry space group to ensure that the reported crystal symmetry was correct.

The rectangular area of the RvR plot under consideration overlaps with the areas discussed above and therefore includes a number of other data sets which turned out to be untwinned but which had special 'features' such as misspecified symmetry, mislabelled structure amplitudes or which displayed RPS and therefore lay on the diagonal with both $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ below 0.5.

78 cases of twinning have been identified, verified and characterized. They are marked in Fig. 1(*c*) with red and green circles corresponding to the presence and absence of RPS interference, respectively. These cases are further discussed in the next subsection.

3.5. Twinning cases

All the structures and their data belonging to the twinning areas of the RvR plot (Figs. 1*a* and 1*c*) were analysed in detail to identify actual twinning cases. Table 1 contains symmetry and NCS information for the 78 twinning cases identified with a high degree of confidence. NCS for DNA structures was not analysed.

There are two features of this table that are worth mentioning. Firstly, pseudomerohedral twinning is not unusual. Secondly, the cases where twinning interferes with RPS are more frequent than simple twinning, especially for the pseudomerohedral twins.

One of the important practical conclusions from these analyses is as follows. If the perfect twinning tests show the presence of twinning, it does not necessarily mean that the twinning is merohedral. Thus, even if data from a perfect twin have point-group symmetry *P*422, it is not necessary, neither theoretically nor in practice, that the point-group symmetry of the individual crystal is *P*4 and that the twinning is merohedral, generated by twofold axis orthogonal to crystallographic fourfold. For example, the crystal symmetry of 1upp (Karkehabadi *et al.*, 2003) is $C222_1$ and there are two twinning domains, with one possible choice of twinning operator being a fourfold axis along one of the crystallographic twofold axes.

3.6. False negatives

It is interesting to note that only one third of the cases identified as twins by our analysis were reported as such in the PDB submission, although in some of these cases analysis of intensities derived from atomic models shows that the twinning was actually taken into account during refinement. Nevertheless, in a significant number of cases this was not done.

The effect of ignoring twinning on $R_{\rm twin}^{\rm calc}$ is illustrated by the following simulated experiment. The 3.1 Å data from an untwinned crystal were artificially twinned to produce six data sets with twinning fractions of 0.0, 0.1, ..., 0.5. The model from the PDB was refined against all data sets following the same protocol, without model rebuilding and ignoring twinning. $R_{\rm twin}^{\rm obs}$ and $R_{\rm twin}^{\rm calc}$ were computed for these six data sets and for the intensities calculated from the appropriate 'refined' models. The result is shown as the central blue curve in Fig. 1(*d*). If twinning had been properly taken into account during refinement then $R_{\rm twin}^{\rm calc}$ would remain constant throughout all these refinements (vertical green line on the right in Fig. 1*d*). Note that 'incorrect refinements' have been carried out starting from the correct model. Even in these cases the points clearly drift towards the left on the RvR plot. Since real-life crystal structure solution requires many cycles of refinement alternated with model building, it is anticipated that this drift to the left is much more serious than in this simulation. To analyse this trend further 'refinements' were carried out with relaxed restraints on ADPs. The results are plotted in red in Fig. 1(*d*) and show further reduction of $R_{\rm twin}^{\rm calc}$.

This simulated experiment helps to explain why only some of the twinning cases without RPS (green points in Fig. 1*c*) show $R_{\rm twin}^{\rm calc} \simeq 0.5$. In all of them proper refinement accounting for twinning has been performed. In some twinning cases without RPS, $R_{\rm twin}^{\rm calc}$ is significantly less than 0.5 and, judging by the simulated results shown in Fig. 1(*d*), we expect that the refinement protocol was not adequate.

The above simulated experiment is one of the cases where the so-called 'model bias' arises because of an insufficient number of parameters and the addition of only one extra parameter, the twinning fraction, would substantially reduce it. Generally speaking, model bias is not so much a consequence of a large number of parameters, but of incorrect parameterization; bias is best corrected by re-parameterization of the model rather than by removing a part of it.

4. Performance of twinning tests

During the verification of twinning in the cases selected using the RvR plot a number of problems were encountered. Some of these problems are of a general nature and are worthy of special attention. This section discusses the influences of experimental error and pseudosymmetry on perfect twinning tests and the influence of RPS on one particular partial twinning test, the *H*-test.

4.1. Effect of experimental errors on the perfect twinning tests

In the perfect twinning tests, the observed intensities normalized within resolution shells are assumed to be sampled from the one-dimensional distributions of the random variable *Z*. Two different distributions are considered, one for centric and one acentric reflections. Derivation of these distributions (Rees, 1980) is based on the Wilson distribution of structure factors (Wilson, 1949) for untwinned crystals. Two of the major tests are based on comparison of the theoretical and observed curves of (i) the cumulative distribution of *Z versus Z* and (ii)

the second moment of Z *versus* resolution, shown in Figs. 2, 3 and 4.

It is necessary to use sensible resolution cutoffs to be able to draw any reliable conclusions from these tests. The reason for this is that high-resolution reflections as a rule have larger experimental errors, but the theory does not take these into account.

Our experience shows that the low-resolution cutoff is not necessary; however, it is important to remove high-resolution data, where R standard = $\langle\sigma(F)\rangle/\langle F\rangle$ starts growing and/or where a large variation of the second moment of Z for acentric reflections is observed. The required plots are available from various software, *e.g. TRUNCATE* and *SFCHECK*. An example of how this rule of thumb works is shown in Fig. 2. In this example, the experimental cumulative distribution of Z clearly indicates perfect twinning with a high-resolution cutoff at 2.2 Å. In contrast, the same test but with all data is misleading and the experimental curves are close to the theoretical curves for untwinned crystals.

It is important to emphasize that high-resolution reflections do contain useful information about the structure despite a resolution cutoff being needed for some applications.

4.2. Effect of RPS on the perfect twinning tests

Our experience shows that RPS affects perfect twinning tests only in the presence of twinning, when it partially compensates for the effect of twinning.

The following numerical experiment illustrates this effect. The 1i1j X-ray data set represents an untwinned crystal with RPS. The data set with perfect twinning was simulated from the original untwinned data. All parameters except the twinning fraction are the same in the two data sets. The experimental second moment of Z *versus* resolution and the cumulative distribution of Z are shown in Fig. 3. The experimental curves for the original data set match theoretical predictions (Figs. 3*a* and 3*b*); however, this is not so for the simulated data set, where only a marginal deviation from the theoretical curves for untwinned data towards those for perfect twins is observed (Figs. 3*c* and 3*d*). This example demonstrates that the simple theory based on Wilson's distribution assuming uniform distribution of atoms in the asymmetric unit fails for twins with RPS.

Such behaviour can intuitively be understood by imaginary traversing of the RvR plot (Fig. 1). If we travel from the main cluster at (0.5, 0.5) towards the origin along the diagonal, we start from a point without anomalies of any kind and finish at the point where crystal symmetry is higher than reported symmetry but also with no anomalies. At both ends we have untwinned data statistics and the same statistical distributions could be expected all along the diagonal pathway (where untwinned data sets with RPS are located). Another limiting path is from the point (0.5, 0.0), below the main cluster, towards the origin along the abscissa. On this path the transition occurs from perfect twin statistics to untwinned statistics. The above example with simulated twinning is located on this path at a point where this transition is almost accomplished.

This behaviour of perfect twinning tests has been observed in a number of real cases where RPS interferes with twinning (Table 1; see also an example in Dauter *et al.*, 2005). The closer the NCS operator generating RPS comes to an operator of higher space group, the less contrast there is between the results of perfect twinning tests for untwinned and twinned data. This lack of contrast creates difficulties for diagnostics. Fortunately, the atomic structure in such circumstances can frequently be solved and refined to a first approximation in a higher symmetry space group and when the refinement sticks at an unreasonably high R factor, the structure can be resolved and further refined in the correct space group. In this scenario problems with uncertain diagnosis are avoided, but it is necessary to collect data in the lower symmetry space group and to keep them unmerged. Reliable diagnosis of twinning therefore becomes an important component of both data collection and refinement.

The effect of RPS on intensities decreases and therefore effect of twinning becomes more pronounced in higher resolution shells, where the intensities are affected by the small difference between NCS-related molecules. However, as noted above the data in higher resolution shells are less reliable for twinning tests because of experimental errors (see §4.1).

4.3. TNCS and twinning

Table 1 shows that when twinning and TNCS coexist the third ingredient, RPS, is usually also present (see, for example, PDB entry1upp; Karkehabadi *et al.*, 2003). This is not surprising because the reduction of symmetry resulting in twinning must involve reduction of crystal point group, *i.e.* formation of RPS.

In these structures higher point-group symmetry and shorter crystallographic translations can be accommodated by small, sometimes less than 1 Å, displacements of atoms. Thus, the modulation of intensities in the reciprocal space caused by TNCS can be considered in terms of sublattices with different mean intensities (pseudocentering). Note that in these structures the sublattices are invariant with respect to twinning operator (compare with §3.4.4).

The modulation of intensities owing to TNCS has an effect on the perfect twinning tests which is opposite to that of twinning. This effect is present in both twinned and untwinned crystals, *e.g.* the second moment of Z for acentric reflections becomes greater than two in the absence of twinning. The effect becomes stronger when the deviation from higher crystal symmetry decreases. Demodulation (normalization accounting for TNCS) or examination of the separate sublattices may reduce the effect of TNCS, but the effect of RPS remains. For different sublattices, the effect of the RPS is different and depends differently on the deviation from higher crystal symmetry, but it always partially compensates for the effect of twinning.

The analysis of the data in terms of sublattices can be avoided by using the perfect twinning test suggested by Padilla & Yeates (2003) and implemented in *DATAMAN* (Kleywegt & Jones, 1996). With a proper resolution cutoff this test indicates perfect twinning, but the contrast is less than that theoretically predicted (Padilla & Yeates, 2003). The presence of RPS in most twins with TNCS explains this observation, as the effect of RPS partially compensates the effect of twinning even if the modulation of intensities owing to TNCS is accounted for.

4.4. Partial twinning tests

There are two partial twinning tests most frequently used in macromolecular crystallography, the Britton test (Britton, 1972) and the *H*-test (Yeates, 1997). In these tests the data are assumed to be processed in the correct group or its subgroup and not 'over-merged' in a higher group. These tests are applied to a given potential twinning operator suggested by the crystal and lattice symmetries. Therefore, all non-equivalent twinning operators (*e.g.* there are two of them in *P*3) have to be tested individually. Unfortunately, neither of these tests can distinguish between higher symmetry and perfect twinning. Nevertheless, in the case of partial twinning they both indicate twinning and estimate the twinning fraction. We discuss the *H*-test in more detail with an emphasis on the effect of pseudosymmetry on its behaviour.

In the *H*-test the joint two-dimensional distribution of the intensities related by potential twinning operator is of interest. Thus, extra information is used compared with the previously discussed twinning tests, which are based on one-dimensional distribution of intensities derived from the Wilson distribution. The idea of this test is that the cumulative distribution $P(H)$ of a random variable H is a straight line over the whole range of possible H. In theory, which unfortunately does not take account of the effects of RPS, the linearity holds for both twinned and untwinned data and the slope of the plot $P(H)$

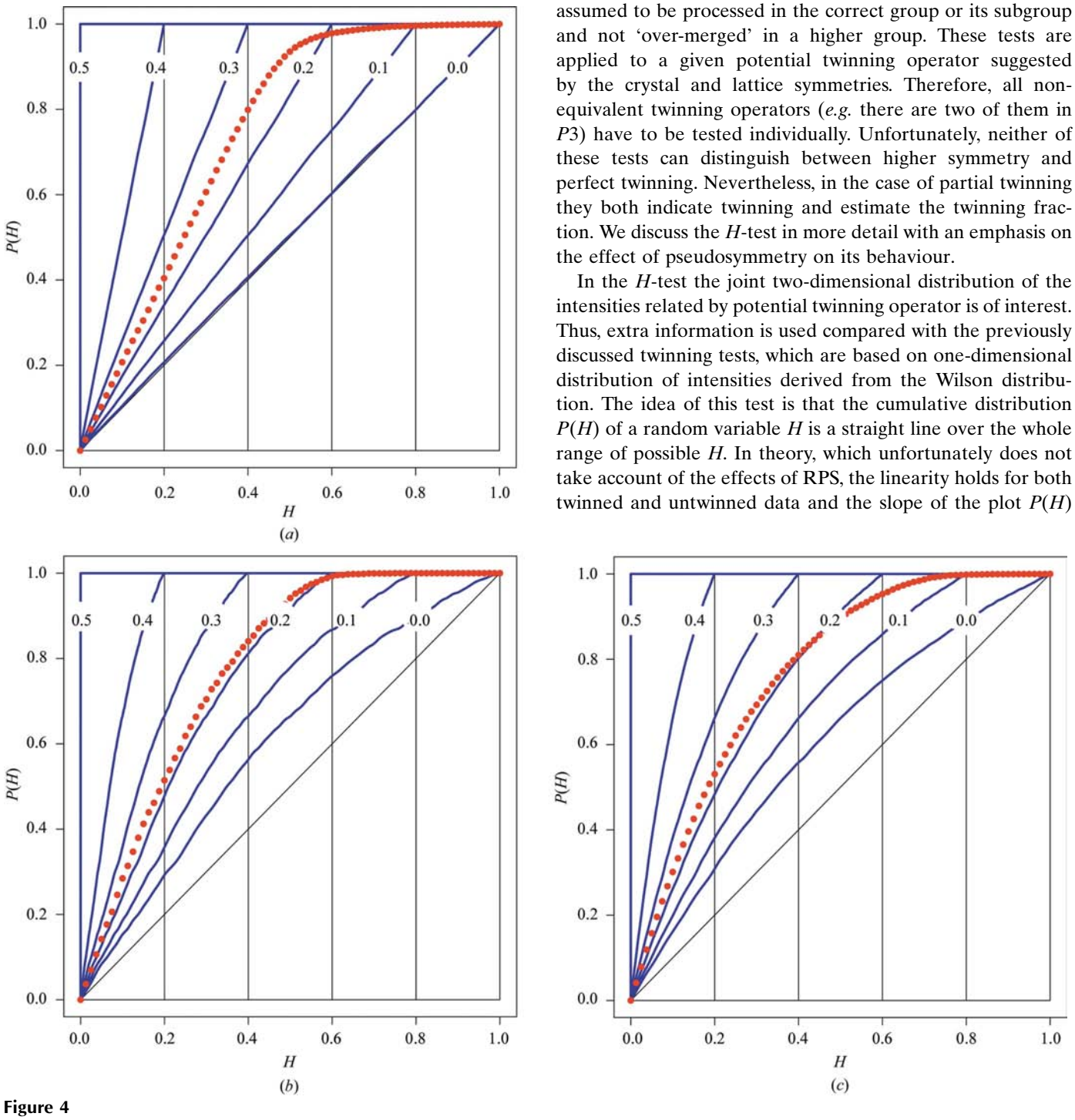

Figure 4
Cumulative distributions of H for three twins: (*a*) 1rxf (Morgan *et al.*, 1994), (*b*) 1ku5 (Li *et al.*, 2002) and (*c*) 1gwy (Mancheno *et al.*, 2003). In 1ku5 and 1gwy, RPS interferes with twinning. Experimental distributions are represented by red dotted lines. The intensities derived from atomic models were used to simulate cumulative distribution of H for different twinning fractions (thick blue lines), the values of the twinning fractions being given over the corresponding lines.

versus H depends on the twinning fraction (blue lines in Fig. 4*a*).

In the original version of the H-test the linearity is essential, as the twinning fraction is estimated from mean value of H. However, theoretically impossible difference between intensities related by twinning operator may appear as a result of radiation damage to the crystal if there was a long time interval between the two measurements. Too large differences could also occur if the X-ray beam was focused at different parts of the crystal during these two measurements. The presence of such outliers distorts the experimental distribution of H at larger H and causes non-linearity as in Fig. 4(*a*). This type of non-linearity is typical of twins without interfering RPS, although the range of H where the plot deviates from the straight line varies. Such cases can be treated by a modified H-test in which the twinning fraction is estimated using the slope of the plot at the origin (Yeates & Fam, 1999).

If RPS interferes with twinning then all pairs of reflections involved in the H-test are affected and the cumulative distribution of H becomes non-linear over the whole range of the argument (Fig. 4*b*) and both versions of H-test fail to give a reasonable estimate of the twinning fraction. In cases similar to that in Fig. 4(*b*), however, the twinning fraction can be estimated from the value of H at the point where the experimental curve approaches the line $P(H) = 1$. In this formulation the H-test is equivalent to the Britton test. A disadvantage of such a formulation is that the estimate of twinning fraction is based on the right tail of the distribution, which can be seriously corrupted by the effects of experimental errors mentioned above (see Fig. 4*c*). Thus, further improvement of the test can only be achieved by accurate modelling of the effect of RPS, TNCS and anisotropy and by accounting for outliers, while keeping the advantage of the original version of the H-test in which the whole data set is utilized.

5. Conclusions

This analysis of the PDB shows that combinations of crystal and lattice symmetries can allow twinning in more than 30% of cases, with both merohedral and pseudomerohedral cases widespread. For easy identification of twinning, the RvR plot was designed, which utilizes both observed intensities and intensities derived from the atomic models. Careful analysis of suspected twins identified from this plot has flagged 78 cases with a high degree of confidence. However, since refinement of the atomic model ignoring twinning causes model bias and thus distorts this picture, we expect there may be more actual twins. Moreover, since twinning is one of the factors that often prevents structure solution, there are almost certainly many cases of twinning that have not been fully analysed and deposited in the PDB.

Analysis of all the identified cases showed that RPS coexists with twinning more frequently than we expected, affecting the intensity distributions and thus increasing the difficulty of detecting twinning and hence the analysis of the structure. We found that all twinning tests can fail to give convincing results. The situation becomes even more serious when TNCS is added to the picture. Ideally, one should also consider other crystal-growth anomalies, such as statistical crystals, non-merohedral twinning and split crystals.

As a result of in-depth analysis of the identified twinning cases, we arrived at the conclusion that it is important to check for this and other crystal-growth anomalies at every stage of structure analysis: starting from data acquisition and ending with refinement and validation. To do this correctly, it is important to build a model accounting for various 'abnormalities' and utilizing all the information available up to the current stage. For example, at the data-collection stage an awareness of twinning may help to choose the correct strategy; during refinement proper modelling of this phenomenon can reduce the noise in the electron density and hence help to reveal finer details of the molecular structure.

APPENDIX *A*

A1. Algorithms used in the determination of twinning operators and their type of merohedry

Several authors (Flack, 1987; Le Page, 2002; Grimmer, 2003) have already described the automatic identification of potential twinning operators using unit-cell parameters and space group. A necessary step in all these algorithms is reducing the cell to a minimum primitive cell, either a Buerger or Niggli cell (see, for example, Mighell & Rodgers, 1980, and references therein). Here, we describe an algorithm designed by one of us (AAL) and implemented in a set of routines.

Given unit-cell parameters and crystal symmetry, potential twinning operators are determined as follows. The cell is reduced to the primitive cell with the shortest unit-cell edges and the point-group operators derived from crystal symmetry are transformed accordingly to give G, a group of 3×3 matrices with elements from $\{-1, 0, 1\}$. The set of 504 matrices with elements from $\{-1, 0, 1\}$ of finite order with respect to matrix multiplication and with determinant equal to one is then generated. This set includes all operators of all rotational point groups expressed in fractional coordinates, provided that the unit cell with shortest edges is chosen. These operators are sorted according to the perturbation that they cause to the metric derived from the primitive unit-cell parameters. The best 24 or less generators satisfying the perturbation threshold of 5% are then used sequentially according to the above sorting order to expand G to H, the rotational point group of the lattice,

$$H = G \cup Ga \cup GaGa \cup \ldots, \qquad (5)$$

where a is the tested generator. Expansion (5) is carried until $a(Ga)^n$ contains no new elements or a new element is inconsistent with H being finite group. Any new consistent $q \in a(Ga)^n$ generates coset Gq of new elements that are added to existing subset of H. This procedure simultaneously produces H and its coset decomposition with respect to G. Representatives of the cosets, one from each excluding G are potential (non-equivalent) twinning operators.

To draw Fig. 1(*b*), we also analysed the type of merohedry of potential twinning operators using the following method. Let G be a rotational point group and M be a metric represented as a set of 6×6 matrices and as a 6-vector, respectively. Let M be invariant with respect to G, *i.e.*

$$gM = M, \quad g \in G. \tag{6}$$

Consequently, the projector

$$\pi = |G|^{-1} \sum_{g \in G} g$$

is such that $\pi M = M$. Let R be a 6×6 matrix representing a potential twinning operator. If

$$R\pi = \pi, \tag{7}$$

then

$$RM = R\pi M = \pi M = M$$

and no constraints are needed for M to be invariant with respect to R in addition to those imposed by (6). Therefore, if (7) holds, then the twinning generated by R is merohedral. This type of merohedry check requires no tables and can be performed in integers if the 6×6 matrix representation of G is generated from its 3×3 matrix representation in fractional coordinates.

APPENDIX *B*

Consider, for example, a threefold twinning operator $S_{\rm twin}$ relating three twinning domains. Let $\mathbf{h}' = S_{\rm twin}\mathbf{h}$ and $\mathbf{h}'' = S_{\rm twin}\mathbf{h}'$ and $\mathbf{h}$, $\mathbf{h}'$ and $\mathbf{h}''$ be different (be in a general position with respect to $S_{\rm twin}$) and corresponding intensities $I^{\rm obs}_{\mathbf{h}}, I^{\rm obs}_{\mathbf{h}'}, I^{\rm obs}_{\mathbf{h}''}$ be measured. For twinning fractions $\alpha_1, \alpha_1, \alpha_1$ and neglected errors,

$$\begin{aligned} 1 &= \alpha_1 + \alpha_2 + \alpha_3, \\ I^{\rm obs}_{\mathbf{h}} &= \alpha_1 I^{\rm calc}_{\mathbf{h}} + \alpha_2 I^{\rm calc}_{\mathbf{h}'} + \alpha_3 I^{\rm calc}_{\mathbf{h}''}, \\ I^{\rm obs}_{\mathbf{h}'} &= \alpha_3 I^{\rm calc}_{\mathbf{h}} + \alpha_1 I^{\rm calc}_{\mathbf{h}'} + \alpha_2 I^{\rm calc}_{\mathbf{h}''}, \\ I^{\rm obs}_{\mathbf{h}''} &= \alpha_2 I^{\rm calc}_{\mathbf{h}} + \alpha_3 I^{\rm calc}_{\mathbf{h}'} + \alpha_1 I^{\rm calc}_{\mathbf{h}''}. \end{aligned} \tag{8}$$

Relations (2) can be verified as follows.

In the absence of twinning, $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = 0$, and therefore for all $\mathbf{h}$ with $I^{\rm obs}_{\mathbf{h}}$ measured, $I^{\rm obs}_{\mathbf{h}} = I^{\rm calc}_{\mathbf{h}}$ and hence $R^{\rm obs}_{\rm twin} = R^{\rm calc}_{\rm twin}$.

For perfect twinning $\alpha_1 = \alpha_2 = \alpha_3$ and $I^{\rm obs}_{\mathbf{h}} = I^{\rm obs}_{\mathbf{h}'} = \mathrm{I}^{\rm obs}_{\mathbf{h}''}$ and $R^{\rm obs}_{\rm twin} = 0$.

For partial twinning

$$\begin{aligned} &|I^{\rm obs}_{\mathbf{h}} - I^{\rm obs}_{\mathbf{h}'}| + |I^{\rm obs}_{\mathbf{h}'} - I^{\rm obs}_{\mathbf{h}''}| + |I^{\rm obs}_{\mathbf{h}''} - I^{\rm obs}_{\mathbf{h}}| \\ &\quad = |\alpha_1(I^{\rm calc}_{\mathbf{h}} - I^{\rm calc}_{\mathbf{h}'}) + \alpha_2(I^{\rm calc}_{\mathbf{h}'} - I^{\rm calc}_{\mathbf{h}''}) + \alpha_3(I^{\rm calc}_{\mathbf{h}''} - I^{\rm calc}_{\mathbf{h}})| + \ldots \\ &\quad \le \alpha_1|I^{\rm calc}_{\mathbf{h}} - I^{\rm calc}_{\mathbf{h}'}| + \alpha_2|I^{\rm calc}_{\mathbf{h}'} - I^{\rm calc}_{\mathbf{h}''}| + \alpha_3|I^{\rm calc}_{\mathbf{h}''} - I^{\rm calc}_{\mathbf{h}}| + \ldots \\ &\quad = (\alpha_1 + \alpha_2 + \alpha_3)(|I^{\rm calc}_{\mathbf{h}} - I^{\rm calc}_{\mathbf{h}'}| + |I^{\rm calc}_{\mathbf{h}'} - I^{\rm calc}_{\mathbf{h}''}| + |I^{\rm calc}_{\mathbf{h}''} - I^{\rm calc}_{\mathbf{h}}|) \\ &\quad = |I^{\rm calc}_{\mathbf{h}} - I^{\rm calc}_{\mathbf{h}'}| + |I^{\rm calc}_{\mathbf{h}'} - I^{\rm calc}_{\mathbf{h}''}| + |I^{\rm calc}_{\mathbf{h}''} - I^{\rm calc}_{\mathbf{h}}|. \end{aligned} \tag{9}$$

If intensities are not equal to zero, then the two sides of the above relation are equal only if $I^{\rm calc}_{\mathbf{h}} = I^{\rm calc}_{\mathbf{h}'} = I^{\rm calc}_{\mathbf{h}''}$. Hence, assuming that the crystal symmetry is correctly specified and thus there are at least some non-zero nonequal triplets of calculated intensities, we have $R^{\rm obs}_{\rm twin} < R^{\rm calc}_{\rm twin}$.

Relations (2) can be similarly derived for any kind of twinning, including the usual case of two twinning fractions.

Let us estimate the expected value of $R^{\rm calc}_{\rm twin}$ (in the absence of any twinning) defined in (1), assuming that (i) there is no RPS and (ii) the overall ADP tensor and the set of TNCS vectors (if TNCS is present) are invariant with respect to $S_{\rm twin}$. Formally, these mean (i) mutual independence of all random variables $I_{\mathbf{h}}$ and (ii) identical exponential distribution of random variables $I_{\mathbf{h}}$ and $I_{\mathbf{h}'}$, $\mathbf{h}' = S_{\rm twin}\mathbf{h}$. In particular, the random variables $I_{\mathbf{h}}$ and $I_{\mathbf{h}'}$ possess the following joint probability distribution density

$$\begin{aligned} p(I_{\mathbf{h}}, I_{\mathbf{h}'}) = p(I_{\mathbf{h}})p(I_{\mathbf{h}'}) = \beta^2_{\mathbf{h}} \exp(-\beta_{\mathbf{h}} I_{\mathbf{h}} - \beta_{\mathbf{h}} I_{\mathbf{h}'}), \\ I_{\mathbf{h}} > 0, \;\; I_{\mathbf{h}'} > 0, \end{aligned} \tag{10}$$

where the multipliers ($\beta_{\mathbf{h}}$) at $I_{\mathbf{h}}$ and $I_{\mathbf{h}'}$ are the same.

For new variables

$$r_{\mathbf{h}} = I_{\mathbf{h}} + I_{\mathbf{h}'}, \quad s_{\mathbf{h}} = I_{\mathbf{h}} - I_{\mathbf{h}'}, \tag{11}$$

the joint probability distribution density is

$$p(s_{\mathbf{h}}, r_{\mathbf{h}}) = \tfrac{1}{2}\beta^2_{\mathbf{h}} \exp(-\beta^2_{\mathbf{h}} r_{\mathbf{h}}), \quad -r_{\mathbf{h}} < s_{\mathbf{h}} < r_{\mathbf{h}}, \tag{12}$$

and, in particular, the conditional probability distribution density of $s_{\mathbf{h}}$ given $r_{\mathbf{h}}$ is

$$p(s_{\mathbf{h}}|r_{\mathbf{h}}) = \frac{p(s_{\mathbf{h}}, r_{\mathbf{h}})}{p(r_{\mathbf{h}})} = \frac{1}{2r_{\mathbf{h}}}, \quad -r_{\mathbf{h}} < s_{\mathbf{h}} < r_{\mathbf{h}}. \tag{13}$$

Thus, the expected value of $|s_{\mathbf{h}}|$ given $r_{\mathbf{h}}$ is

$$\mathcal{E}(|s_{\mathbf{h}}||r_{\mathbf{h}}) = \frac{1}{2\,r_{\mathbf{h}}} \int_{-r_{\mathbf{h}}}^{r_{\mathbf{h}}} |s_{\mathbf{h}}|\, \mathrm{d}s_{\mathbf{h}} = \frac{r_{\mathbf{h}}}{2}. \tag{14}$$

Finally,

$$\mathcal{E}(R_{\rm twin}) = \mathcal{E}\left[\frac{\sum\limits_{\mathbf{h}} |s_{\mathbf{h}}|}{\sum\limits_{\mathbf{h}} r_{\mathbf{h}}}\right] = \mathcal{E}\left[\frac{\sum\limits_{\mathbf{h}} \mathcal{E}(|s_{\mathbf{h}}||r_{\mathbf{h}})}{\sum\limits_{\mathbf{h}}\; r_{\mathbf{h}}}\right] = \mathcal{E}\left(\frac{1}{2}\right) = \frac{1}{2}. \tag{15}$$

The last equation means that $R_{\rm twin}$ averaged over all possible structures obeying the above conditions (i) and (ii) equals one half exactly. For a particular structure, this means the approximate equation in (4).

These calculations can also be applied to two unrelated structures, as was performed by Srinivasan & Parthasarathy (1976) for a similar problem but for the conventional R factor.

It is important to stress that this interpretation does not mean that in the X-ray experimental data the resolution shells with $R_{\rm merge}$ higher than 50% are useless. In the case of experimental data, experimental errors are necessarily present and their distribution is different from that used to derive the above relation. The estimation of the resolution cutoff is a completely different problem and has to be approached using different notions, such as the informational content of the data or the informational content of the data per unit of synchrotron time.

We thank Eleanor Dodson, George Sheldrick, Ian Tickle, Olga Moroz and Vladimir Levdikov for helpful discussions and practical examples. This work was supported by BBSRC (AAL and AAV, grant reference B10670) and the Wellcome Trust (GNM).

References

Barends, T., deJong, R., van Straaten, K., Thunnisen, A.-M. & Dijkstra, B. (2005). *Acta Cryst.* D**61**, 613–621.
Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.*, **112**, 535–542.
Bragg, W. L. & Howells, E. R. (1954). *Acta Cryst.* **7**, 409–411.
Britton, D. (1972). *Acta Cryst.* A**28**, 296–297.
Cochran, W. & Howells, E. R. (1954). *Acta Cryst.* **7**, 412–415.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z. (2003). *Acta Cryst.* D**59**, 2004–2016.
Dauter, Z., Botos, I., LaRonde-LeBlanc, N. & Wlodawer, A. (2005). *Acta Cryst.* D**61**, 967–975.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Flack, H. (1987). *Acta Cryst.* A**43**, 564–568.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.
Friedel, G. (1926). *Leçons de Cristallographie.* Paris: Blanchard.
Giacovazzo, H. L., Monaco, H. L., Viterbo, D., Scordari, F., Gilli, G., Zanotti, G. & Catti, M. (1992). *Fundamentals of Crystallography.* Oxford University Press.
Grimmer, H. (2003). *Acta Cryst.* A**59**, 287–296.
Karkehabadi, S., Taylor, T. C. & Andersson, I. (2003). *J. Mol. Biol.* **334**, 65–73.
Kleywegt, G. J. (1999). *Acta Cryst.* D**55**, 1878–1884.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.
Le Page, Y. (2002). *J. Appl. Cryst.* **35**, 175–181.
Li, T., Ji, X., Fun, F., Gao, R., Cao, S., Peng, Y. & Rao, Z. (2002). *Acta Cryst.* D**58**, 870–871.
Lougheed, J. C., Holton, J. M., Alber, T., Bazan, J. F. & Handel, T. M. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 5515–5520.
Mancheno, J., Martin-Benito, J., Martinez-Ripoll, M., Gavilanes, J. & Hermoso, J. (2003). *Structure*, **11**, 1319–1328.
Mighell, A. D. & Rodgers, J. R. (1980). *Acta Cryst.* A**36**, 321–326.
Morgan, N., Pereira, I., Andersson, I., Adlington, R., Baldwin, J., Cole, S., Crouch, N. & Sutherland, J. (1994). *Bioorg. Med. Chem. Lett.* **4**, 1595–1600.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Padilla, J. & Yeates, T. (2003). *Acta Cryst.* D**59**, 1124–1130.
R Development Core Team (2004). *R: A language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
Rees, D. (1980). *Acta Cryst.* A**36**, 578–581.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Rudolph, M. G., Kelker, M. S., Schneider, T. R., Yeates, T. O., Oseroff, V., Heidary, D. K., Jennings, P. A. & Wilson, I. A. (2003). *Acta Cryst.* D**59**, 290–298.
Schlessman, J. & Litvin, D. B. (1995). *Acta Cryst.* A**51**, 947–-949.
Sheldrick, G. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Srinivasan, R. & Parthasarathy, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon.
Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Yeates, T. (1997). *Methods Enzymol.* **276**, 345–358.
Yeates, T. O. & Fam, B. C. (1999). *Structure Fold. Des.* **7**, R25–R29.