Since PCV = $1/\langle I/\sigma(I)\rangle$, this quantity also provides an indication as to whether the standard deviations $\sigma(I)$ have been estimated appropriately. Finally, the statistic $R_{\rm mrgd\text{-}F}$, used for assessing the quality of the reduced data, enables a direct comparison of this merging R value with the refinement residuals R and $R_{\rm free}$.

Ideally, merging statistics should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.2. *Completeness*

Data completeness can be assessed by calculating what fraction of the unique reflections within a range of Bragg spacings that could in theory be observed has actually been measured. Ideally, completeness should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.3. *Redundancy*

Redundancy is defined as the number of independent observations (after merging of partial reflections) per unique reflection in the final merged and symmetry-reduced data set. Ideally, average redundancy should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.4. *Signal strength*

The average strength or significance of the observed intensities can be expressed in different ways. Values that are often quoted include the percentage of reflections for which $I/\sigma(I)$ exceeds a certain value (usually 3.0) and the average value of $I/\sigma(I)$. Ideally, these numbers should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.5. *Resolution*

The nominal resolution limits of a data set are chosen by the crystallographer, usually at the data-processing stage, and ought to reflect the range of Bragg spacings for which useful intensity data have been collected. Unfortunately, owing to the subjective nature of this process, resolution limits cannot be compared meaningfully between data sets processed by different crystallographers. Careful crystallographers will take factors such as shell completeness, redundancy and $\langle I/\sigma(I)\rangle$ into account, whereas others may simply look up the minimum and maximum Bragg spacing of all observed reflections. Bart Hazes (personal communication) has suggested defining the effective resolution of a data set as that resolution at which the number of observed reflections would constitute a 100% complete data set. Alternatively, Vaguine *et al.* (1999) define the effective (or optical) resolution as the expected minimum distance between two resolved peaks in the electron-density map and calculate this quantity as $2\Delta_P/2^{1/2}$, where Δ_P is the width of the origin Patterson peak. One day, hopefully, the term 'resolution' will be replaced by an estimate of the information content of data sets. Randy Read (personal communication) has carried out preliminary work along these lines.

21.1.7.1.6. *Unit-cell parameters*

The accuracy of unit-cell parameters has been shown to be grossly overestimated for small-molecule crystal structures (Taylor & Kennard, 1986). Not intimidated by this observation, some macromolecular crystallographers routinely quote unit-cell axes of 100–200 Å with a precision of 0.01 Å. An analysis of several high-resolution protein crystal structures has revealed that surprisingly large errors in the unit-cell parameters appear to be quite common (at least if synchrotron sources are used for data collection; EU 3-D Validation Network, 1998). Such errors can be detected *a posteriori* by checking if the bond lengths in a model show any systematic, perhaps direction-dependent, variations from their target values.

21.1.7.1.7. *Symmetry*

From the symmetry of the diffraction pattern, the point-group symmetry of the crystal lattice can usually be derived. It is important to merge the data in the point group with the highest possible symmetry (usually assessed using merging statistics) in order to minimize the chance of making an incorrect space-group assignment (Marsh, 1995, 1997; Kleywegt *et al.*, 1996). Once the first data set has been processed, it is always useful to compute a self-rotation function. A non-origin peak of comparable strength to the origin peak will indicate that the true space group has higher symmetry. [Similarly, a self-Patterson function can be calculated at this stage to detect any purely translational NCS (Kleywegt & Read, 1997).] Once the final model is available, a search for possibly missed higher symmetry can be carried out, *e.g.* using the method developed by Hooft *et al.* (1994).

Sometimes crystallographic symmetry breaks down (pseudo-symmetry): an apparent higher symmetry at low resolution does not hold at higher resolution. In some cases, this is a consequence of the chemistry of the system studied (*e.g.* an asymmetric ligand bound by a symmetric protein dimer). In other cases, it may go undetected and complicate space-group determination and solution and refinement of the structure.

When it comes to space-group determination, many of the lessons learned by small-molecule crystallographers also apply to macromolecular crystallography (Marsh, 1995; Watkin, 1996).

21.1.7.2. *Model quality, coordinates*

Many criteria (and computer programs) are available to check for structural outliers based only on analysis of Cartesian coordinate sets.

21.1.7.2.1. *Geometry and stereochemistry*

The covalent geometry of a model can be assessed by comparing bond lengths and angles to a library of 'ideal' values. In the past, every refinement and modelling program had its own set of 'ideal' values. This even made it possible to detect (with 95% accuracy) with which program a model had been refined, simply by inspecting its covalent geometry (Laskowski, Moss & Thornton, 1993). Nowadays, standard sets of ideal bond lengths and bond angles derived from an analysis of small-molecule crystal structures from the CSD (Allen *et al.*, 1979) are available for proteins (Engh & Huber, 1991; Priestle, 1994) and nucleic acids (Parkinson *et al.*, 1996). For other entities, typical bond lengths and bond angles can be taken from tables of standard values (Allen *et al.*, 1987) or derived by other means (Kleywegt & Jones, 1998; Greaves *et al.*, 1999).

For bond lengths, the r.m.s. deviation from ideal values is invariably quoted. Deviations from ideality of bond angles can be expressed directly as an angular r.m.s. deviation or in terms of angle distances (*i.e.* the angle $\angle ABC$ is measured by the 1–3 distance $|AC|$; note that this distance is also implicitly dependent on the bond

references