Scaling and assessment of data quality

Philip Evans

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Correspondence e-mail: pre@mrc-lmb.cam.ac.uk

The various physical factors affecting measured diffraction intensities are discussed, as are the scaling models which may be used to put the data on a consistent scale. After scaling, the intensities can be analysed to set the real resolution of the data set, to detect bad regions (*e.g.* bad images), to analyse radiation damage and to assess the overall quality of the data set. The significance of any anomalous signal may be assessed by probability and correlation analysis. The algorithms used by the *CCP*4 scaling program *SCALA* are described. A requirement for the scaling and merging of intensities is knowledge of the Laue group and point-group symmetries: the possible symmetry of the diffraction pattern may be determined from scores such as correlation coefficients between observations which might be symmetry-related. These scoring functions are implemented in a new program *POINTLESS*.

1. Introduction

The diffraction intensities measured by integrating spots recorded on an area detector are not all on the same scale because they are affected by a number of physical factors from the experiment, most of which are difficult to measure directly. The process of 'data reduction' uses the redundancy of multiple measurements of symmetry-related reflections to put all observations on a common scale by fitting a scaling model which reflects the experiment. This process produces a data set which is internally consistent, within the errors of the model, though not necessarily correct on an absolute scale.

Analysis of the agreement between equivalent reflections after scaling gives estimates of the quality of the data and also highlights parts of the data which agree poorly with the rest. This allows decisions to be made about whether parts of the data should be rejected.

This paper discusses the physical reasons for the differences in scale, the scaling model and the analysis of data. This discussion is based on the *CCP*4 program *SCALA*, but the general ideas also apply to other implementations of scaling. Some more details of the program *SCALA* are given in Appendix *A*. This paper also discusses some considerations in the determination of the Laue group and hence the space group and a new program (*POINTLESS*) which scores and ranks different possible Laue groups.

2. Physical reasons for differences of scale

The various factors affecting the measured intensity can be divided into those dependent on the primary beam and the

way in which the crystal is rotated, those dependent on the diffracted beam direction and those dependent on the detector. These factors may then be combined into a model to correct the measured intensities.

2.1. Factors related to the incident X-ray beam

We generally assume that reciprocal space has been sampled by rotation of the crystal at a constant speed in an incident beam of constant or smoothly varying intensity and that adjacent images exactly abut each other in rotation angle. Variations in the rotation rate, rapid fluctuations in incident-beam intensity or errors in synchronization of the shutter cause systematic errors which are difficult or impossible either to detect or to model and ideally these factors should be explicitly monitored.

Correctable factors are slow variation in incident-beam intensity (for example on synchrotron beams), change in illuminated volume if the beam is smaller than the crystal and absorption in the primary beam. These can be grouped together into a single correction factor dependent on the crystal rotation.

2.2. Factors related to the crystal and diffracted beam

Absorption in the secondary beam direction is serious at long wavelengths and worth correcting in all cases, particularly as a relative correction for single- or multiple-wavelength anomalous scattering measurements. The most difficult systematic error is radiation damage, since radiation causes the structure to change with time, which means that different reflections change at different rates. Extrapolation to zero time (Diederichs et al., 2003; Diederichs, 2006) requires many observations of each reflection well spaced out in time and this is not generally possible in radiation-sensitive cases. The relative B factor (see §3 and §A2.2) is essentially a correction for average radiation damage.

2.3. Factors related to the detector

The detector should be properly calibrated for spatial distortion and sensitivity of response as well as for any defective regions and should be stable: detector corrections cannot easily be extracted from diffraction data. The user will usually have to tell the integration program about shadows from the beam-stop and other obstructions and it is important to do so.

3. Modelling the correction factors

The scaling model should be chosen as far as possible to describe the diffraction experiment performed. Various scaling models have been used to model the correction as a function of rotation (or time) and the direction of the diffracted beam: a good discussion of modelling the various factors, using a general exponential model, is given by Otwinowski et al. (2003). The simplest model applies a different scale factor for each image, but the scale does not usually vary sharply from one image to another, so a smooth function is

more appropriate: the function used in SCALA was inspired by the method of Kabsch (1988) (see Appendix A and Kabsch, 2000). Using separate scales for each image ('batch' scales) introduces discontinuities in the scale, even if neighbouring scales are restrained together (Otwinowski et al., 2003), which is usually undesirable. 'Batch' scaling also causes complications for partially recorded reflections in that different parts of the same reflection have different scales, so that in the determination of scales either the partial derivatives must be partitioned according to the calculated fraction or each part must be treated separately and scaled up to the full equivalent (Rossmann & van Beek, 1999); both methods use the calculated fraction, which is typically not very accurate. A smooth scale model avoids this problem by scaling each reflection after summing all its parts.

The other traditional component of the scaling model is a relative B factor, $\exp(-2B\sin^2\theta/\lambda^2)$, where B is a function of time (or rotation or image number). This provides a resolution-dependent radiation-damage correction, but it is an average correction and cannot account for localized radiation damage. Like the scale factor, this is best treated as a smooth function of time (or rotation as its proxy; see §A2.2).

Absorption in the secondary beam direction is best para-meterized as coefficients of real spherical harmonics, either in the rotating crystal frame or in the diffractometer frame (Katayama, 1986; Blessing, 1995). These two coordinate frames give very similar results if data are collected about a single rotation axis, but if a crystal is rotated about two or more axes a single absorption surface expressed in the crystal frame may in principle be used for all rotation sweeps. This assumes perfect centring of the crystal, so use of different surfaces for each sweep is likely to be better.

SCALA includes an optional and rather crude correction for errors introduced by the long tails on reflections from diffuse scattering and the inconsistency of sampling these tails by relatively coarse slicing on the rotation (see §A2.4). This may be helpful when the image width is comparable to or larger than the reflection width, when the sampling of the reflection profile is very different between reflections measured on one, two or more images. The error caused by this differential sampling is apparent in the systematic underestimation of fully recorded reflections (from one image) compared with summed partials (from two or more images), giving rise to a negative 'partial bias', defined as $\sum_h(\langle I_{\mathrm{full}}\rangle - I_{\mathrm{partial}})/\sum_h(\langle I_{\mathrm{full}}\rangle)$, where $\langle I_{\mathrm{full}}\rangle$ is the average of all fully recorded observations of the reflection (or more generally of the observations with the smallest number of parts) and I_{partial} is a summed partial observation and the summation is over all reflections which have both fully recorded and partial observations and over all summed partials. This correction should be applied with caution, since such bias can also arise from underestimation of the mosaic spread defining the width of the Bragg peak: in this case, the 'tails' correction is inappropriate. It is also unlikely to be helpful when the image width is smaller than the reflection width ('fine slicing').

research papers

4. Determining the correction factors

The correction factors are optimized to make the data as internally consistent as possible by minimizing the difference between symmetry-related observations. Note that the only information we have is the measured difference between symmetry-related observations (unless an external reference data set is used), so any systematic error which follows the crystallographic symmetry will not be corrected: in particular, crystal absorption errors may remain since the shape of a crystal often obeys its diffraction symmetry. It follows that to obtain the most accurate data symmetry-related observations should be measured in as different a way as possible (by rotation about more than one axis). Conversely, to obtain the most accurate differences for phasing (anomalous or isomorphous), equivalent observations should be be measured in as similar way as possible, with the same systematic errors.

The function minimized is

$$\Psi = \sum_{\mathbf{h}} \sum_{l} w_{hl}(I_{hl} - g_{hl}\langle I_h \rangle)^2 + \text{parameter restraint terms},$$

where I_{hl} is the lth observation of reflection $\mathbf{h}$, g_{hl} is its associated inverse scale, $w_{hl} = 1/\sigma^2(I_{hl})$ and $\langle I_h \rangle$ is the weighted average intensity for all observations l of reflection $\mathbf{h}$ (Hamilton et al., 1965; Fox & Holmes, 1966). The inverse scale g_{hl} is a function of all the parameters in the model.

From minimization of Ψ within one reflection,

$$\langle I_h \rangle = \sum_{l} w_{hl}g_{hl}I_{hl} \Big/ \sum_{l} w_{hl}g_{hl}^2.$$

By minimizing Ψ over all reflections, we obtain values for all the parameters. This is performed by a singular value decomposition (Fox & Holmes, 1966), eliminating two zero eigenvalues corresponding to the scales and the B factors, since the residual Ψ is unchanged by multiplying all the scale parameters by a constant or by adding a constant to all the B factors. The first scale factor is normalized to a value of 1 and the B-factor parameters are all forced negative by normalizing the largest one to 0. Parameters may be restrained by additional terms in the residual: for example, it is useful to restrain the coefficients of the spherical harmonic terms in the
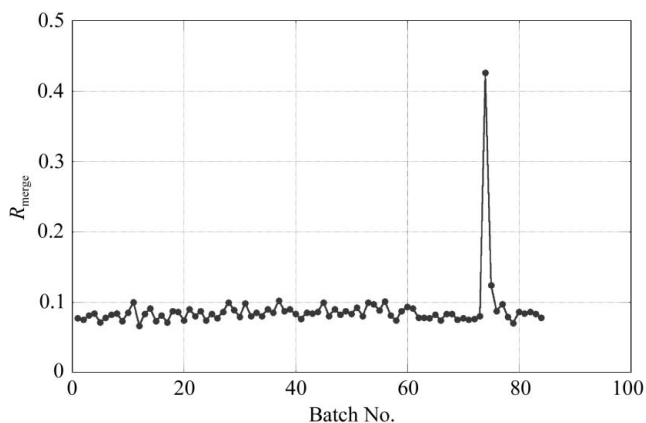

Figure 1
A plot of R_{merge} against batch number shows one wrong batch (a blank image).

absorption correction to a target value of 0 to avoid wild corrections with limited data (see §A2.3.)

5. Assessment of data quality

After applying the refined scale model, the quality of the data may be assessed in a number of ways based on the internal consistency of the data and comparison of the corrected intensities with the corrected standard deviations (see §A3). There are a number of important questions about the data which need to be answered: what is the real resolution, are there bad regions of data which should be omitted, is there any anomalous signal and what is the overall quality of the data? The internal consistency may be measured as R factors or as correlation coefficients. The conventional R_{merge} (also known as R_{sym}) is not a particularly good measure of data quality as it only measures the discrepancy between observations and takes no account of the improvement in the merged intensity by averaging many observations: indeed, R_{merge} tends to increase with increasing multiplicity. Improved multiplicity-weighted R factors have been suggested by Diederichs & Karplus (1997), Weiss & Hilgenfeld (1997) and Weiss (2001). If n_h is the number of observations of reflection $\mathbf{h}$, then

$$R_{\text{merge}} = R_{\text{sym}} = \sum_{\mathbf{h}} \sum_{l} |I_{hl} - \langle I_{\mathbf{h}} \rangle| \Big/ \sum_{\mathbf{h}} \sum_{l} \langle I_{\mathbf{h}} \rangle,$$

the traditional R_{merge},

$$R_{\text{meas}} = R_{\text{r.i.m.}} = \sum_{\mathbf{h}} \left(\frac{n_h}{n_h - 1}\right) \sum_{l} |I_{hl} - \langle I_{\mathbf{h}} \rangle| \Big/ \sum_{\mathbf{h}} \sum_{l} \langle I_{\mathbf{h}} \rangle,$$

the multiplicity-independent R factor, and

$$R_{\text{p.i.m.}} = \sum_{\mathbf{h}} \left(\frac{1}{n_h - 1}\right) \sum_{l} |I_{hl} - \langle I_{\mathbf{h}} \rangle| \Big/ \sum_{\mathbf{h}} \sum_{l} \langle I_{\mathbf{h}} \rangle,$$

the precision-indicating R factor.

$R_{\text{meas}} = R_{\text{r.i.m.}}$ is an improved version of the traditional R_{merge} and measures how well the different observations agree. $R_{\text{p.i.m.}}$ is a measure of the quality of the data after averaging the multiple measurements.

5.1. What is the real resolution?

The variance-weighted average intensities fall off with increasing resolution and may be compared with the corrected standard deviation estimate (§A3). A typical resolution cutoff is when $\langle I/\sigma(I) \rangle$ (averaging within resolution bins on $1/d^2 = 4\sin^2\theta/\lambda^2$) falls below 2.0. Beyond this point, the data are probably too weak to be useful in structure determination. The correlation coefficient between intensities averaged within two random half data sets also gives an indication of the maximum resolution (see Fig. 3b and §6). Many crystals show anisotropic diffraction and the resolution limits ought to be anisotropic, but at present no programs treat anisotropic data gracefully.

5.2. Are there bad parts of the data?

A plot of R_{merge} against 'batch' number will show if there any individual images or parts of the data which are significantly worse than the rest of the data: this might suggest that there is a bad image or that something has gone wrong with the integration. In the case illustrated in Fig. 1 there is a blank image owing to the beam disappearing.

Radiation damage causes serious degradation of data quality and shows up in several plots against batch number, but most clearly from the relative B factor: Fig. 2 shows that as the crystal dies the scale increases, the B factor becomes more negative, the R_{merge} increases and $\langle I \rangle / \langle \mathrm{Scatter} \rangle$ [where Scatter is the r.m.s. value of $(I_{hl} - \langle I \rangle)$] decreases.

5.3. Outlier rejection

Occasionally, individual observed intensities are just wrong, for one of a number of reasons. These include (i) spots which do not belong to the main crystal lattice but overlap a predicted position, from ice crystals, salt crystals or another crystal, (ii) zingers, *i.e.* events on the detector which do not arise from X-rays, and (iii) spots which lie outside the active area of the detector, *e.g.* behind the beamstop.
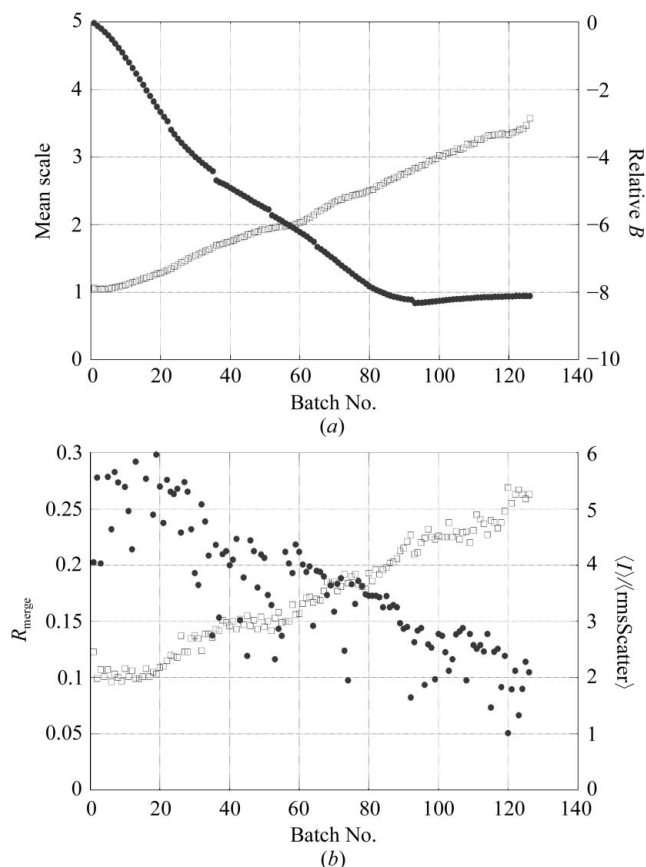

Figure 2
Radiation damage: with increasing time of exposure (or batch number), the scale increases as the intensity decreases (*a*, open squares), the relative B factor gets becomes negative (*a*, filled circles), R_{merge} increases (*b*, open squares) and $\langle I \rangle / \langle \mathrm{Scatter} \rangle$ decreases (*b*, filled circles). [$\langle \mathrm{Scatter} \rangle$ is r.m.s.$(I_{hl} - \langle I \rangle)$.]

Detecting outliers is reasonably easy if the reflection has been measured many times, but is not possible for a reflection measured only once or twice: this is a major reason for measuring data with a high multiplicity. The outlier rejection algorithm used in *SCALA* is described in §A5. Note that outlier detection generally assumes that the majority of observations of a reflection are correct: one common case where this may cause problems is with spots behind a slightly miscentred beamstop, when it is possible that the majority of observations are wrong and the program will reject the correct ones. It is important to tell the integration program (*e.g.* *MOSFLM*) the position of the beamstop explicitly.

Spurious observations arising from ice or salt spots are often very large and may be rejected if they have an intensity much larger than would be expected (Read, 1999). This test is performed on the normalized amplitudes E, normalized as a function of resolution such that $\langle E^2 \rangle = 1$. An E of > 4 is very unlikely, but because of the errors in normalization, particularly at low resolution where the mean intensity is changing rapidly with resolution and there are relatively few reflections, or with anisotropic data, it is better to reject only observations with $E > 8$–10.

6. Scaling of multiple-wavelength data sets and detection of anomalous signal

When multiple data sets have been collected from the same crystal (or indeed different crystals) at different wavelengths for a MAD experiment, the relative systematic errors may be reduced by scaling them together, assuming for the purposes of scaling that the differences between the data sets arising from different anomalous scattering are small. Similarly, the differences between Bijvoet pairs I^+ and I^- within a data set are usually small and may be ignored in the scaling step. Scaling data sets together forces all observations to be as similar as possible within the scaling model and improves the signal, since the scaling model varies slowly in reciprocal space while the desired signal (anomalous or dispersive differences) varies more rapidly, so the differences remaining after the relative systematic errors have been removed are closer to the true signal. This was discussed by Evans (1997), but in retrospect the scaling seems to work well without the reference data set recommended there.

It is useful to know if there is any significant anomalous or dispersive signal before attempting to locate anomalous scatterers. The observed anomalous differences may be compared with their estimated standard deviations using a normal probability plot of the normalized differences $\delta_{\mathrm{anom}} = I^+ - I^- [\sigma^2(I^+) + \sigma^2(I^-)]^{1/2}$ (Howell & Smith, 1992). The slope of the central region of this plot will be >1 if the anomalous differences are larger than expected from their standard deviations (Fig. 3*a*). Another way of detecting a signal is from the correlation coefficient between differences in different data sets (Fig. 3*b*): this will fall off with resolution and may be used to set a suitable maximum resolution limit for initial trials to locate anomalous scatterers (Schneider & Sheldrick, 2002). If only one data set is available (SAD), it may be split randomly

into two half data sets, provided the multiplicity is high enough, and the correlations calculated between the two halves (Fig. 3c).

Another way of analysing the significance of the anomalous signal from the half data sets is from a scatter plot: for each reflection we divide the I^+ and the I^- observations randomly into two sets, average them within the sets and subtract them to obtain $\Delta I_1 = \langle I^+\rangle_1 - \langle I^-\rangle_1$, $\Delta I_2 = \langle I^+\rangle_2 - \langle I^-\rangle_2$ for each reflection, then plot ΔI_1 against ΔI_2. For perfect data where $\Delta I_1 = \Delta I_2$, this plot would have all points lying along the diagonal. The correlation coefficient is the slope of the least-squares straight line fitted through these points, but it is very sensitive to a few outliers and makes no use of the fact that the slope should be 1.0 for ideal data. Real data (Fig. 4a) shows a distribution which is roughly elliptical. The width of the distribution along the diagonal is a measure of the signal and its width perpendicular to the diagonal is a measure of the error, so the ratio of these, the r.m.s. correlation ratio = (r.m.s. deviation along diagonal)/(r.m.s. deviation perpendicular to diagonal), can be used as a measure of the significance of the signal and may be plotted as a function of resolution (Fig. 4b). In the absence of any anomalous signal, the distribution is spherical (Fig. 4c) and the r.m.s. correlation ratio is close to 1 (Fig. 4d). This measure seems somewhat more robust than the

correlation coefficient, with less variation between resolution bins, but leads to similar conclusions about a suitable resolution at which to truncate the data to preserve a strong signal: for the peak wavelength in the example in Figs. 3, 4(a) and 4(b), a good signal extends to about 3.6 Å resolution with correlation coefficients between and within data sets of above about 0.3 and an r.m.s. correlation ratio of above 1.5.

7. Determination of Laue group, point group and space group

The true space group of a crystal cannot be known with certainty until the structure has been solved and refined, since it is easy to be misled by pseudosymmetry and perhaps by twinning, but the space group does impose itself on the diffracted intensities and from these it is possible to propose the likely space group or at least a range of possibilities. It is useful to find the likely symmetry as early as possible during the initial examination of a crystal, since it affects the data-collection strategy (how much rotation range is needed for a
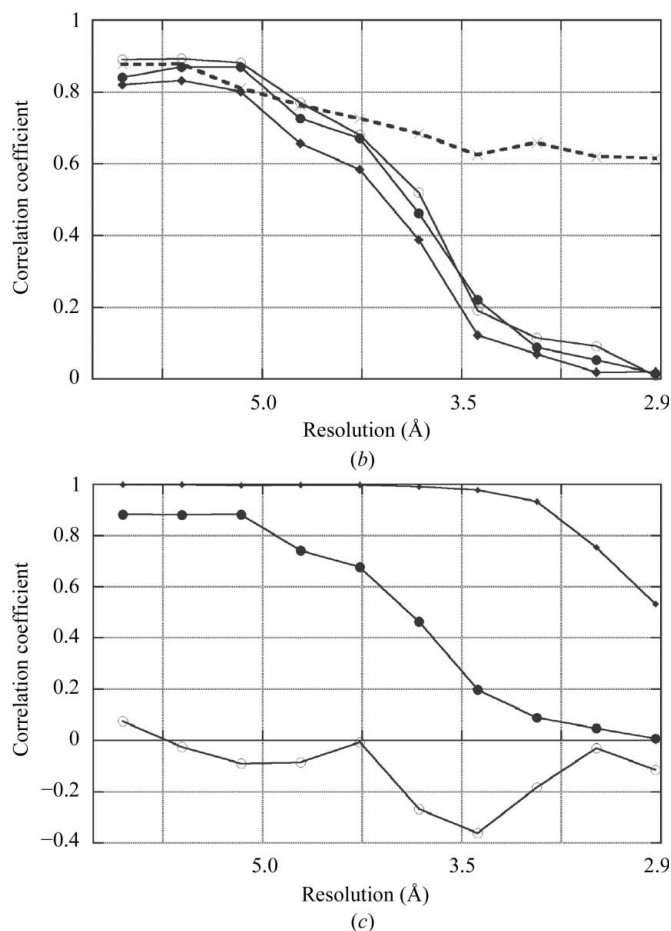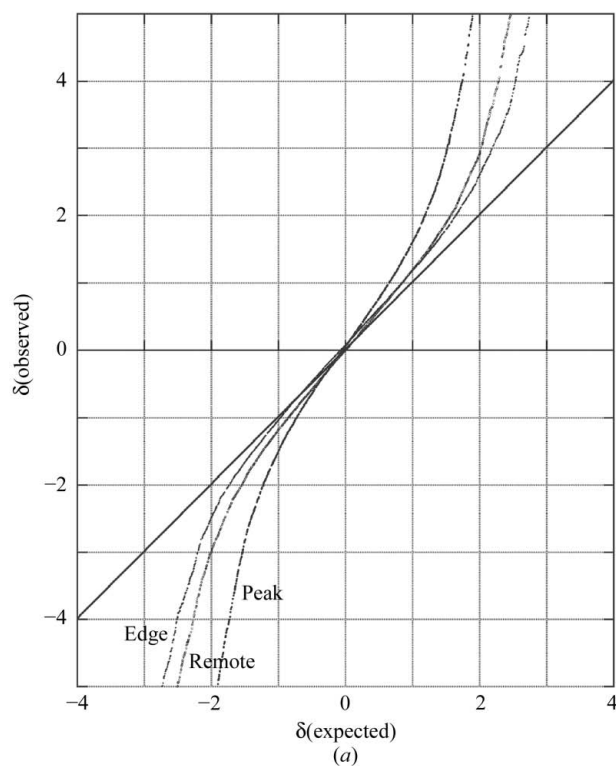

Figure 3
Significance of anomalous signal in a three-wavelength MAD data set (peak, edge, remote). (a) Normal probability plot of anomalous differences $\delta_{anom} = (I^+ - I^-)/[\sigma^2(I^+) + \sigma^2(I^-)]^{1/2}$ for each wavelength. The central slope indicates the strength of the anomalous signal relative to the estimated errors, Peak > Remote > Edge. (b) Correlation coefficients between pairs of different wavelengths: filled circles, peak to edge; diamonds, edge to remote; open circles, peak to remote. The dashed line is the correlation coefficient between dispersive differences, peak–remote to edge–remote. (c) Correlation coefficients between random half data sets with the peak data set: filled circles, anomalous differences (acentric); open circles, Bijvoet differences for centric data (should be 0); diamonds, $\langle I\rangle$, showing decrease in the quality of the intensities themselves at high resolution.

complete data set). Scaling and merging depends on the Laue group (or more strictly, the point group; see below), since this controls which spots are related by symmetry. This section describes the methods which are used in a new program to determine the Laue group, *POINTLESS*, which will be distributed in the *CCP*4 suite.

7.1. Stages in space-group determination

The determination of space group can be considered as a series of stages of increasing difficulty: determining successively the lattice symmetry, the Laue group, the point group and the space group. At all stages, distinguishing between the possibilities may be uncertain owing to either a small number of observations or pseudosymmetry (see §7.2).

7.1.1. Lattice symmetry: crystal class. Autoindexing determines the unit-cell parameters of the observed lattice initially without constraints, but the crystal class imposes restrictions on the allowed cell (*e.g.* $a = b$, $\alpha = \beta = 90$, $\gamma = 120°$ for a hexagonal lattice) and lattice centring restricts which indices are present (*e.g.* $h + k + l$ even for an *I*-centred lattice). When indexing a diffraction pattern, the user (or the program) chooses a lattice which fits geometrically to the observed spot positions within an acceptable limit on some penalty function (see, for example, Leslie, 2006), but the apparent cell restrictions may occur accidentally (*e.g.* $\beta \simeq 90°$ in a monoclinic cell) and at the indexing stage the intensities are not available to indicate that the wrong choice has been made.

7.1.2. Laue group symmetry. The Laue group is the symmetry of the diffraction pattern, plus any lattice centring. It corresponds to the space group without any translations, with an added centre of symmetry from Friedel's law. The Laue group may be inferred from the observed symmetry of the diffraction pattern (see §7.2).

7.1.3. Point-group symmetry. To take anomalous dispersion into account, intensity observations should be averaged according to the point group that can be derived from the space group by removing the lattice type and translations. For chiral space groups (*i.e.* for all macromolecular crystals), there is only one possible point group corresponding to each Laue group. For many non-chiral space groups, the point group may be inferred by determination of which principal zones of the reciprocal lattice are centric, which can be performed from intensity statistics: a centre of symmetry makes all reflections centric and a twofold axis (rotation or screw) makes the perpendicular zone centric, while a mirror or glide plane does not. However, in practice tests on zone statistics are unreliable, particularly in the presence of heavy atoms or pseudosymmetry (G. M. Sheldrick, personal communication).

7.1.4. Space-group symmetry. The space group is the point group plus translations (screw axes for chiral space groups). Screw axes are only visible in the diffraction pattern as systematic absences along the axes and these are not always very reliable as there may be few reflections and there may also be accidental absences. Determination of the translational part of the space group from axial absences must be considered as a hypothesis to be confirmed by structure solution. In non-chiral cases, possible glide planes introduce absences throughout a zone which may be detected more reliably.

7.2. Scoring functions for determination of Laue group symmetry

To distinguish between possible Laue groups, we need to compare observations which might be related by potential symmetry and score their agreement. There are two problems which need to be addressed in choosing a suitable method of scoring. Firstly, we would like to be able to obtain a preliminary idea of the symmetry from a very partial data set, from the first few images, even before a complete data collection and we want a method which is robust to limited data and will not give a spurious high score from a few accidental agreements. Secondly, we would like a score function which is insensitive to the scale between observations, since we need to know the symmetry to scale the data.
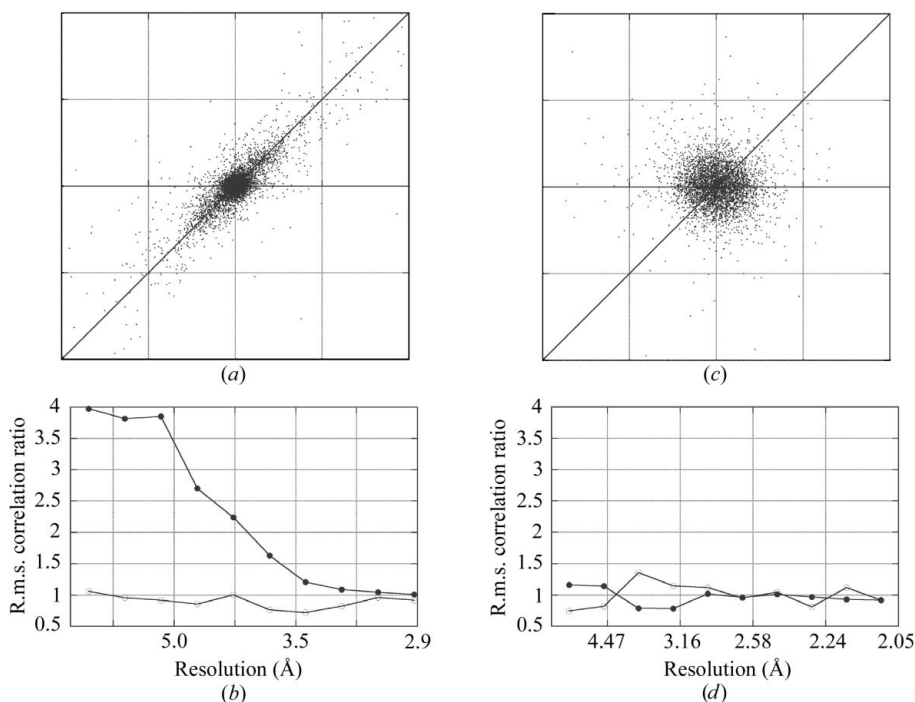
Two sorts of scoring functions have been tried.

Figure 4
Significance of anomalous signals from random half data sets. (*a*) Scatter plot of ΔI_{anom} pairs for the peak data set shown in Fig. 3, showing a strong correlation. (*b*) R.m.s. correlation ratio (see text) as a function of resolution: filled circles, acentric data; open circles, centric data. (*c*) Scatter plot for a native data set, showing no anomalous signal. (*d*) R.m.s. correlation ratio for native data set.

research papers

(i) Difference functions, which are sensitive to the unknown scale

$$\mathrm{RMSdifference} = -\left(\sum_{\mathbf{h}}\{(I_1 - I_2)^2/[\sigma^2(I_1) + \sigma^2(I_1)]\}\right)$$

summed over all pairs of observed intensities I_1 and I_2 which might be related by symmetry. This would be the log(probability) if the errors $\sigma^2(I)$ were random and I_1 and I_2 were on the same scale.

(ii) Product functions such as the correlation coefficient, which are relatively insensitive to the unknown scale. A correlation coefficient does however assume that all observations come from the same underlying distribution, so it needs to be calculated from normalized intensities E^2 to avoid the artificial correlation arising from the change in $\langle I\rangle$ with resolution. Since use of E^2 enhances the weak high-resolution intensities, thus inflating their errors, it is necessary to truncate the resolution of the data to remove very weak data. At present, *POINTLESS* uses a cutoff $\langle I\rangle/\langle\sigma(I)\rangle > 1.5$. Surprisingly, correlation coefficients with contributions weighted by $1/\sigma^2(I)$ seemed in several trials to be less discriminating than the standard unweighted coefficients.

Use of the correlation coefficient reduces the problem of the unknown scales, but the problem of small samples remains. The approach used in *POINTLESS* is to calculate the score given by all possible intensity pairs related by a potential symmetry element (the test score) and to compare this score with scores from the same size groups of unrelated pairs. The many pairs at the same resolution which cannot be related by symmetry are divided into groups of the same size as the test sample (with a maximum size of say 200, since larger groups should not be very different), the score is calculated for each group and then the mean and standard deviation of these scores used to convert the test score into a Z score,

$$Z(\text{test score}) = \frac{\{\text{Score(test)} - \text{Mean[Score(unrelated)]}\}}{\sigma[\text{Score(unrelated)}]}.$$

7.3. Determining the Laue group in *POINTLESS*

POINTLESS reads unmerged integrated intensities from, for example, *MOSFLM* and determines the lattice with highest possible symmetry compatible with the unit-cell parameters, within a rather generous limit (currently 3°; Le Page, 1982). The symmetry in the file is ignored. Most of the symmetry handling in the program uses the cctbx library (Grosse-Kunstleve *et al.*, 2002). Each symmetry element (rotation axis) in this lattice symmetry is scored separately using all pairs of observations related by that rotation. All the possible combinations of these elements are then scored, giving all the possible subgroups. For each subgroup the score for elements belonging to the lattice group but not to the subgroup are subtracted from the score for elements which do belong to the subgroup

$$\mathrm{Net}Z = Z(\text{for}) - Z(\text{against}).$$

Table 1
Example 1: an orthorhombic case with $a \simeq b$, scores for each symmetry element.

Unit-cell parameters $a = 44.67$, $b = 46.10$, $c = 117.89$ Å, $\alpha = \beta = \gamma = 90°$, tested in tetragonal lattice $P4/mmm$. Even in the limited data set from images 1–5 (5°, left-hand part of table) the twofold axes along c [001] and a [100] are clearly present, but the twofold along b [010] has only four pairs of observations so is undetermined. The potential fourfold axis along c [001] is not obviously present and the diagonal twofolds are absent. With the full 90° data set (right-hand part of table) the Z scores are larger mainly because σ[CC(unrelated)] is smaller. CC is the correlation coefficient between potentially related pairs of E^2 and Z-CC is the Z score for correlation coefficients as defined in the text.

Symmetry element	Images 1–5				All data		
	Z-CC	CC	No.		Z-CC	CC	No.
Twofold [001]	**1.51**	**0.48**	**22**	+++	**11.0**	**0.73**	**24337**
Twofold [100]	**2.85**	**0.73**	**33**	+++	**11.4**	**0.75**	**33259**
Twofold [110]	−1.02	−0.13	45		2.23	0.14	26701
Twofold [010]	**−1.36**	**−0.76**	**4**	+++	**11.0**	**0.73**	**19199**
Twofold [1−10]	−1.10	−0.15	37		0.93	0.05	28477
Fourfold [001]	−0.68	−0.01	72	+	3.72	0.24	60928

This favours the highest symmetry consistent with a good score in preference to lower symmetries with good Z(for) scores.

7.3.1. Example 1: an orthorhombic case with $a \simeq b$. A crystal indexed and integrated with unit-cell parameters $a = 44.67$, $b = 46.10$, $c = 117.89$ Å, $\alpha = \beta = \gamma = 90°$ was tested in the possible tetragonal lattice $P4/mmm$ using either just the first 5° of data or a full 90° data set. Table 1 shows the scores for the individual possible symmetry elements: the twofold axes along c [001] and a [100] are clearly present, but the twofold along b [010] has only four pairs of observations and thus is indeterminate. The potential fourfold axis along l [001] is not obviously present and the diagonal twofolds are absent. With the full 90° data set (right-hand part of table) the Z scores are larger, mainly because σ[CC(unrelated)] is smaller and the twofold along b is now clear. Table 2 shows the scores for all the possible Laue groups, showing that even with the very limited 5° of data the correct Laue group $Pmmm$ is reasonably clear.

7.3.2. Example 2: pseudo-hexagonal $Cmmm$. A hexagonal lattice may be indexed as C-centred orthorhombic in three different ways, related by 60° rotations. Conversely, a true C-centred orthorhombic lattice with $b = 3^{1/2}a$ can be indexed as hexagonal. In this case, an autoindexing program has only a one in three chance of picking the correct orthorhombic lattice.

In the case illustrated in Tables 3 and 4, the unit cell has $b \simeq 3^{1/2}a$ and was indexed incorrectly. The scores on the individual symmetry elements (Table 3) clearly pick out the correct 222 set of rotations and the combination (Table 4) selects the correct $Cmmm$ setting.

7.4. Future directions

Future developments of *POINTLESS* will include assessment of intensity statistics and systematic absences in order to score possible space groups and to detect twinning and comparison with previously collected data sets to choose

Table 2
Example 1: an orthorhombic case with $a \simeq b$, possible Laue groups ranked by Net$Z$-CC.

Z-CC+ is the Z(for) score for symmetry elements belonging to the subgroup; Z-CC− is the Z(against) score for symmetry elements belonging to the lattice group but not the subgroup.

Laue group	Images 1–5				All data			Reindex operator
	Net$Z$-CC	Z-CC+	Z-CC−		Net$Z$-CC	Z-CC+	Z-CC−	
Pmmm	**3.79**	**3.01**	**−0.78**	+++	**8.15**	**11.12**	**2.97**	**[h, k, l]**
P12/m1	3.43	2.85	−0.58	++	6.44	11.38	4.94	[k, h, −l]
P12/m1	1.63	1.51	−0.12	++	6.03	10.99	4.96	[−h, l, k]
P4/m	0.24	0.28	0.05	++	5.86	11.02	5.17	[h, k, l]
P4/mmm	−0.04	0.04	0.0	++	5.70	5.70	0.00	[h, k, l]
P−1	−0.04	0.0	0.04		−1.02	5.28	6.30	[h, k, l]
C12/m1	−1.26	−1.1	0.16		−1.16	4.97	6.12	[h + k, −h + k, l]
P12/m1	−1.44	−1.36	0.08		−3.85	2.23	6.08	[h, k, l]
Cmmm	−1.75	−0.65	1.10		−5.44	0.93	6.36	[h + k, −h + k, l]
C12/m1	−1.88	−1.02	0.86		−5.70	0.00	5.70	[h − k, h + k, l]

Table 3
$C222$ pseudo-hexagonal lattice.

Unit-cell parameters $a = 74.72$, $b = 129.22$, $c = 184.25$ Å, $\alpha = \beta = \gamma = 90°$, $b \simeq 3^{1/2}a$, tested in hexagonal lattice group $P6/mmm$. Twofold axes are present along the hexagonal axes a, $c \times a$ [−1 2 0] and c.

Symmetry element	Z-CC	CC
Twofold [001]	**10.22**	**0.70**
Twofold [1−10]	−0.52	−0.03
Twofold [2−10]	0.11	0.02
Twofold [100]	**11.37**	**0.78**
Twofold [110]	−0.83	−0.05
Twofold [010]	0.22	0.02
Twofold [−120]	**11.60**	**0.79**
Threefold [001]	−0.03	0.01
Sixfold [001]	0.70	0.06

Table 4
$C222$ pseudo-hexagonal lattice, discriminating between the three possible orthorhombic cells.

R_{meas} is the multiplicity-weighted R_{merge}, but calculated for unscaled normalized (E^2) intensities, hence its high value.

Laue group	Net$Z$-CC	Z-CC+	Z-CC−	CC	R_{meas}	Reindex operator
Cmmm	**10.94**	**10.97**	**0.03**	**0.75**	**0.19**	**[3/2h + 1/2k, 1/2h − 1/2k, −l]**
Cmmm	2.48	4.68	2.21	0.33	0.47	[1/2h + 1/2k, 3/2h − 1/2k, −l]
Cmmm	−0.71	2.42	3.13	0.17	0.48	[h, k, l]
P6/mmm	2.86	2.86	0.00	0.20	0.51	[1/2h + 1/2k, 1/2h − 1/2k, −l]

between alternative valid but non-equivalent indexing schemes. Ultimately, it is intended that all the facilities of *SCALA* will also be included.

APPENDIX *A*
Algorithms used in *SCALA*

This appendix describes some of the details of the scaling and analysis calculations in *SCALA*. It is not comprehensive, but covers the most important and commonly used functions. The description here refers to version 3.2.13. Most of the algorithms also described in the documentation for *SCALA* distributed by CCP4.

A1. Files, data sets, runs and batches: data organization

Unmerged intensity data is read from a file in the *CCP*4 MTZ format, which represents a hierarchy of data organization. This file typically comes from the integration program *MOSFLM*, but intensities from other programs may be imported *via* the *CCP*4 programs *COMBAT* or *DTREK2-SCALA*. With *COMBAT*, geometric information may be lost in this process, so not all scaling options may be available. The file may contain several data sets (*e.g.* collected at different wavelengths for MAD), each of which is divided by the program into 'runs'. Each run consists of spots from a set of

contiguous images ('batches') and has its own set of scaling parameters, *i.e.* the scales vary smoothly within the run but are different between runs. *SCALA* automatically divides data into runs at any discontinuity in batch number or rotation angle. A 'reflection' consists of all 'observations' of symmetry-related intensities and each observation may consist of a number of 'parts'. Parts are summed to form a complete observation, provided that either the flags from *MOSFLM* are consistent (*e.g.* all parts 1–3 of three present) or the total calculated fraction (read from the file) is within limits (usually 0.95–1.05).

A2. Scaling model

The inverse scale factor for an observation I_{hl} (*i.e.* the *l*th observation of reflection **h**) is composed of four parts

$$g_{hl} = (\text{Scale } C_{hl}) \times (B \text{ factor } T_{hl}) \times (\text{Absorption } S_{hl})$$
$$\times (\text{Tails correction } L_{hl}).$$

A2.1. Scale factor. The scale term C_{hl} for a reflection measured at rotation angle φ is smoothly interpolated with Gaussian weights from a series of scales at intervals, typically $5°$ ($\Delta\varphi$), covering the range of the data in a run. For the normalized rotation angle $r = (\varphi - \varphi_0)/\Delta\varphi$ (where φ_0 is the initial rotation angle),

$$\text{scale}C(r) = \sum_i C_i \exp[-(r - r_i)^2/V_r],$$

research papers

where C_i are the scale factors at positions r_i and the summation i is over all scales close to position r, *i.e.* with $(r - r_i)^2/V_r <$ ProbLim (default = 3.0). V_r is the 'variance' of the weight (default value 1.0). This is similar to the method of Kabsch (1988).

A2.2. B factor. The B-factor term is similarly derived from a smoothed function of 'time' t (usually time is taken as equivalent to φ), with B factors defined at intervals (default interval $\Delta t = 20°$ on φ).

$$\text{Normalized time } t = (t_{hl} - t_0)/\Delta t,$$

where t_0 is the initial time,

$$B\text{-factor scale } T(t) = \exp[+2B(t) \sin^2 \lambda/\lambda^2],$$

$$B(t) = \sum_i B_i \exp[-(t - t_i)^2/V_B],$$

where B_i are the B factors at positions t_i, the summation is for $(t - t_i)^2/V_B <$ ProbLim and V_B is the smoothing weight (default = 0.5). Note the positive sign in the exponent arises because this is the inverse scale.

A2.3. Absorption correction. The absorption term is derived from summing a series of spherical harmonic terms (Katayama, 1986; Blessing, 1995) as a function of the diffracted beam vector $\mathbf{s}_2$, expressed either in the diffractometer frame or the crystal frame.

$$S_{hl}(\mathbf{s}_2) = 1 + \sum_{l=1}^{l_{\max}} \sum_{m=-l}^{+l} C_{lm} Y_{lm}(\mathbf{s}_2),$$

where C_{lm} are the coefficients to be determined and $Y_{lm}(\mathbf{s}_2)$ are the spherical harmonic functions. Harmonics up to order $l_{\max} = 4$ or 6 are sufficient to give a good fit. The initial '1' in this equation is essentially the zeroth-order term ($l = 0$). Ideally, absorption should be identical if the beam is reversed

[*i.e.* $S(\mathbf{s}_2) = S(-\mathbf{s}_2)$] which implies that terms with l odd should have zero coefficients, but inclusion of the odd-order terms allows for crystal mis-centring and other approximations and provides a useful correction to errors in anomalous differences, since I^+ and I^- observations then have different corrections applied, even for inverse-beam experiments.

The coordinate frame used for $\mathbf{s}_2$ is not critical, but usually the diffractometer frame is used,

$$\mathbf{s} = \mathbf{UBh},$$

where $\mathbf{B}$ is the crystal orthogonalization matrix, $\mathbf{U}$ is the orientation matrix, Φ is the diffractometer rotation matrix and $\mathbf{s}$ is the diffraction vector

$$\mathbf{s}_2 = \mathbf{s} - \mathbf{s}_0,$$

where $\mathbf{s}_0$ is the incident-beam vector. The diffractometer frame

$$\mathbf{s}_{2d} = [-\Phi]\mathbf{s}_2$$

and the permuted crystal frame

$$\mathbf{s}_{2c} = [-\mathbf{Q}][-\mathbf{U}][-\Phi]\mathbf{s}_2,$$

where $\mathbf{Q}$ is a permutation matrix

To keep the absorption surface smooth in regions where there is no data, the coefficients C_{lm} are restrained to a value of 0 with a quadratic penalty function added to the total residual,

$$R = \sum_{lm} w_s C_{lm}^2,$$

where the weight $w_s = 1/\sigma_s^2$ with a default value of $\sigma_s = 0.001$. Otwinowski *et al.* (2003) have suggested using tighter restraints on high order terms,
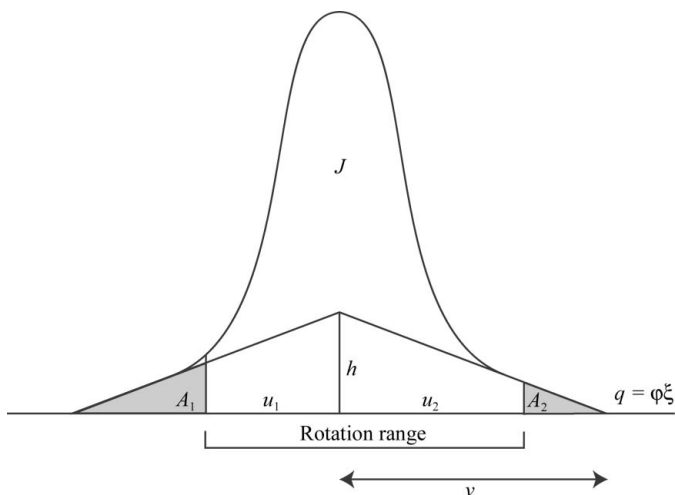
$$R = \sum_{lm} w_s l^2 C_{lm}^2.$$

Figure 5
Tails correction. The Bragg peak (intensity J) is superimposed on a thermal diffuse scattering peak modelled as a triangle of half-width v and height h in the reciprocal-space coordinate $q = \varphi\xi$, where φ is the rotation angle and ξ is the radius of the spot from the rotation axis. If the spot is integrated over the rotation range shown, the two areas A_1 and A_2 of the diffuse scattering are not measured.

A2.4. Tails, a correction for diffuse scattering. Diffuse scattering causes long tails on reflections and tails in the direction of rotation (φ) are often truncated by the integration program by an underestimate of the reflection width on φ, the 'mosaic spread'. A spot may be integrated on one or more images: a fully recorded observation is integrated over a narrower rotation range than a partial, so will include less of the 'tail' of the spot. This leads to a systematic difference between fulls and summed partials, a negative partial bias. *SCALA* implements a very crude correction for this systematic difference (Evans, 1997), based on some ideas from Blessing (1987).

Thermal diffuse scattering is proportional to the Bragg intensity J, so the measured intensity $I = J(1 + \alpha)$. The constant of proportionality varies with resolution (and may be aniso-tropic): it is modelled in *SCALA* as $\alpha = (\sin\theta/\lambda)^2\alpha_1$, where α_1 is a refinable parameter. The width of the thermal diffuse scattering peak is assumed to be constant in reciprocal space, a refinable parameter v. The peak may be modelled as a triangle of height h in the reciprocal-space coordinate q (Fig. 5), given by the fraction of the complete peak area $J\alpha = hv$. If the total

scan range from the start of the first image to the end of the last image is less than $2v$, the diffuse scattering peak is truncated by the areas marked A_1 and A_2 in Fig. 5. We can calculate a correction to the equivalent full scan.

Intensity for full scan corrected for diffuse scattering is given by

$$J = I - hv = I/(1 + \alpha).$$

Intensity for partial scan from point u_1 to point u_2 is given by

$$J = I - (hv - A_1 - A_2) = I - hv(1 - C_1 - C_2)$$
$$= I/[1 + \alpha(1 - C_1 - C_2],$$

where

$$C_j = A_j/hv = \begin{cases} 0 & u_j > v \\ \frac{1}{2}[(v - u_j)/v]^2 & 0 < u_j < v \\ 1 - \frac{1}{2}[(v + u_j)/v]^2 & U_j < 0 \end{cases}.$$

Correction factor (dividing scale factor) $= [1 + \alpha(1 - C_1 - C_2)]/(1 + \alpha) = f(v, \alpha_1)$.

A3. Estimation of errors

Integration programs such as *MOSFLM* (Leslie, 2006) produce good estimates of intensities, but the estimates of the individual errors are less reliable and are typically underestimated. After scaling, the error estimates can be improved by comparing the observed scatter between observations and the estimated standard deviation, making them equal on average. If the standard deviations $\sigma(I_{hl})$ are correct, then the normalized deviations $\delta_{hl} = (I_{hl} - \langle I'_h \rangle)/\sigma(I_{hl})$ (where $\langle I'_h \rangle$ is averaged over all observations of reflection **h** excluding the lth observation) should be distributed with a mean 0.0 and standard deviation 1.0. A simple correction to give improved error estimates is $\sigma'(I_{hl}) = Sdfac[\sigma^2(I_{hl}) + (Sdadd\, g_{hl}\langle I_h \rangle)^2]^{1/2}$. These correction factors Sdfac and Sdadd have at least some physical justification: Sdadd allows for the fact that many potential errors are proportional to the true intensity, for example fluctuations in the incident beam and errors in the exact rotation. *SCALA* uses a default value of Sdadd = 0.02. The factor Sdfac is a more general correction for unknown errors, but includes uncertainty in the detector gain which converts detector-readout values to photon counts which are used to estimate Poissonian errors. To determine Sdfac, *SCALA* uses a normal probability analysis (Abrahams & Keve, 1971; Howell & Smith, 1992) of δ_{hl} and sets factor Sdfac equal to the slope of the central part of the normal probability plot, thus forcing the slope to be 1.0 after correction. Using just the central part of the plot for this avoids fitting outliers in the distribution.

A4. Averaged intensities

Individual observations I_{hl} are averaged using the variance as weight,

$$\langle I_h \rangle = \frac{\sum_l w_{hl}g_{hl}I_{hl}}{\sum_l w_{hl}g_{hl}^2},$$

$$\sigma(\langle I_h \rangle) = \frac{1}{\sum_l w_{hl}g_{hl}^2},$$

$$w_{hl} = 1/\sigma^2(I_{hl}).$$

A5. Outlier rejection algorithm

Flow-chart for rejection algorithm.

(i) If there are three or more observations: for each I_{hl} (l = 1, n), calculate the weighted mean of all other observations $\langle I'_h \rangle$ and its error estimate $\sigma(\langle I'_h \rangle)$.

(ii) Calculate normalized deviations $\delta_{hl} = (I_{hl} - g_{hl}\langle I'_h \rangle)/[\sigma^2(I_{hl}) + (g_{hl}\langle I'_h \rangle)^2]^{1/2}$.

(iii) Find the largest deviation: if $\max|\delta_{hl}| >$ SdRej (default value 6) then reject one observation, but which one?

(iv) Count number of observations for which $\delta_{hl} > 0$ (N_+) and $\delta_{hl} < 0$ (N_-).

(v) If either of N_+ or $N_- = 1$, then one observation is a long way from all others, so reject this one. Otherwise, reject the observation with the largest $|\delta_{hl}|$.

(vi) If there are still three or more observations, iterate from step (i).

(vii) If there are two observations left, by default keep both.

References

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157–165.
Blessing, R. H. (1987). *Crystallogr. Rev.* **1**, 3–58.
Blessing, R. H. (1995). *Acta Cryst.* A**51**, 33–38.
Diederichs, K. (2006). *Acta Cryst.* D**62**, 96–101.
Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.
Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Acta Cryst.* D**59**, 903–909.
Evans, P. R. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances In Phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 97–102. Warrington: Daresbury Laboratory.
Fox, G. C. & Holmes, K. C. (1966). *Acta Cryst.* **20**, 886–891.
Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D (2002). *J. Appl. Cryst.* **35**, 126–136.
Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *Acta Cryst.* **18**, 129–130.
Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kabsch, W. (2000). In *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann & E. Arnold. Dordrecht: Kluwer Academic Publishers.
Katayama, C. (1986). *Acta Cryst.* A**42**, 19–23.
Le Page, Y. (1982). *J. Appl. Cryst.* **15**, 255–259.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.

Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Acta Cryst.* A**59**, 228–234.

Read, R. J. (1999). *Acta Cryst.* D**55**, 1759–1764.

Rossmann, M. G. & van Beek, C. G. (1999). *Acta Cryst.* D**55**, 1631–1640.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.

Weiss, M. S. & Hilgenfeld, R. (1997). *J. Appl. Cryst.* **30**, 203–205.