

2. BASIC CRYSTALLOGRAPHY

hkl for which observations exist in both data sets across the entire resolution range or within a particular resolution shell. CC_{anom} is a reliable indicator of the strength of the anomalous signal. Values above 0.30 are considered good.

R.m.s. correlation ratio. This is another statistic based on randomly partitioned data sets, which is calculated by the program *SCALA* (Evans, 2006; Collaborative Computational Project, Number 4, 1994). It is an analysis of the scatterplot of $[I_1(hkl) - I_1(\bar{h}\bar{k}\bar{l})]$ versus $[I_2(hkl) - I_2(\bar{h}\bar{k}\bar{l})]$, where the subscripts 1 and 2 identify the two half data sets. The analysis assumes that the correlation is ideally 1.0. The r.m.s. correlation ratio is defined as the ratio of the r.m.s. widths of the scatterplot distribution along the diagonal and perpendicular to the diagonal. This statistic seems to be more robust than CC_{anom} to the presence of outliers. It cannot, however, be applied to analysing the correlations between different data sets.

Mean anomalous signal-to-noise ratio, $\langle d''/\sigma(d'') \rangle$. The anomalous signal-to-noise ratio of an individual reflection measurement $d''(hkl)/\sigma[d''(hkl)]$ is defined as the ratio of the observed anomalous intensity difference $d'' = |I(hkl) - I(\bar{h}\bar{k}\bar{l})|$ and the corresponding estimated standard uncertainty $\sigma(d'')$ in the measurement of this anomalous difference. The average of the anomalous signal-to-noise ratios for all reflections within a certain resolution range is used as an indicator of utility for phasing. A value of $(2/\pi)^{1/2} \simeq 0.8$ for mean $d''/\sigma(d'')$ of a resolution shell, for example, is taken to indicate that no anomalous signal is present (G. Sheldrick & G. Bunkoczi, personal communication).

Decay R factor, R_d . The decay R factor R_d is defined as a pairwise R factor based on the intensities of symmetry-related reflections occurring on different diffraction images (Diederichs, 2006). An increase in R_d as a function of difference in image-collection times is a good indicator of radiation damage occurring during data collection. In fractional form, this is

$$R_d = 2 \sum_{hkl} \sum_{m-n} |I_m(hkl) - I_n(hkl)| / \sum_{hkl} \sum_i [I_m(hkl) + I_n(hkl)], \quad (2.2.2.14)$$

where $I_m(hkl)$ and $I_n(hkl)$ are the intensities of the reflection hkl occurring on images m and n . The only program in which this is currently implemented is *XDSSTAT*.

Wilson-plot B factor, B_{Wilson} . A Wilson plot (Wilson, 1949) is a plot for a contiguous series of resolution shells of the logarithm of the mean intensity in a given resolution shell divided by the sum of the squared atomic form factors for all atoms in the unit cell evaluated at the mean of the resolution limits of the shell. From a least-squares fit of a straight line to the linear part of the Wilson plot, the B factor B_{Wilson} can be derived. Typically, data of lower than 4.5 Å resolution are excluded from the fit. The more meaningful determinations of B_{Wilson} come from Wilson plots that are linear all the way to the nominal resolution d_{min} and minimize the occurrence of spikes due to ice rings.

$$\ln(\langle I_{\text{obs}} \rangle / \sum_i f_i^2) = -\ln K_{\text{Wilson}} - 2B_{\text{Wilson}}/d^2, \quad (2.2.2.15)$$

where $\langle I_{\text{obs}}(hkl) \rangle$ is the mean over the intensities of all observed reflections hkl in a given resolution shell. The sum \sum_i runs over all atoms in the structure. The parameter d is the midpoint of the resolution shell over which I_{obs} has been averaged. K_{Wilson} is an absolute scale factor.

2.2.3. Comparing different diffraction data sets

In MX, there are many instances when two or more diffraction data sets have to be compared to each other. An important decision which has to be made is, for instance, whether data sets collected from different crystals are isomorphous enough so that they can be merged together. Another task is the comparison of a native data set and a heavy-atom derivative. The question here is how much of the observed difference is due to non-isomorphism and how much is due to isomorphous replacement.

Scaling R factor, R_{scale} . The scaling R factor R_{scale} between two data sets is defined as the difference in the structure-factor amplitudes of one data set relative to the structure-factor amplitudes of the other data set after the two data sets have been scaled to each other.

$$R_{\text{scale}} = \sum_{hkl} |F_1 - F_2| / \sum_{hkl} |F_1|. \quad (2.2.3.1)$$

However, the above formula is asymmetric with respect to data sets 1 and 2. An alternative symmetric formula is

$$R_{\text{scale}} = 2 \sum_{hkl} |F_1 - F_2| / \sum_{hkl} (F_1 + F_2). \quad (2.2.3.2)$$

R_{scale} is also often given based on intensities rather than structure-factor amplitudes.

Gradient from a normal probability analysis. For each reflection hkl in two data sets that have been scaled to each other, the quantity $\Delta(\text{real})$ is computed and compared with its expectation value $\Delta(\text{expected})$ assuming a normal distribution of errors (Howell & Smith, 1992). If $\Delta(\text{real})$ is plotted against $\Delta(\text{expected})$, random differences between the two data sets give rise to a slope of 1.0. Significant differences result in slopes significantly larger than 1.0. Such an analysis is implemented in the program *SCALEIT* (Collaborative Computational Project, Number 4, 1994).

$$\Delta(\text{real}) = (F_1 - F_2) / [\sigma(F_1)^2 + \sigma(F_2)^2]^{1/2}. \quad (2.2.3.3)$$

2.2.4. Quality indicators for substructure determination

The first step of a crystal structure determination after processing and analysing the diffraction data is to use the observed isomorphous and/or anomalous differences in order to determine the isomorphous or anomalous substructure. Since such substructures typically consist of rather few atoms, which are relatively far apart, direct-methods approaches have been adopted from small-molecule crystallography and have largely superseded pure Patterson-based methods. The most widely used computer programs for substructure determination are *SHELXD* (Schneider & Sheldrick, 2002; Sheldrick, 2008), *SnB* (Weeks *et al.*, 1993, 1994) and *HySS* (Grosse-Kunstleve & Adams, 2003).

Correlation coefficient $CC(\text{all})$. $CC(\text{all})$ measures the correlation of the E values (analogous to normalized structure factors; Hauptman & Karle, 1953) derived from the observed isomorphous or anomalous differences and those calculated from the substructure model. $CC(\text{all})$ is another application of the Pearson linear correlation coefficient [see equation (2.2.2.13)] and is calculated for all reflections. In the equation, x and y are the E values derived from the observed and calculated differences. $CC(\text{all})$ is typically used to assess the successful determination of the isomorphous or anomalous substructure in a macromolecule. A value of $CC(\text{all}) \geq 0.3$ often indicates that the substructure has been correctly identified (Schneider & Sheldrick, 2002).

Correlation coefficient $CC(\text{weak})$. Like $CC(\text{all})$, $CC(\text{weak})$ is the Pearson linear correlation coefficient [see equation (2.2.2.13)]