International Tables for Crystallography (2006). Vol. F, Section 11.2.4, pp. 212–213.

11.2. Integration of macromolecular diffraction data

BY A. G. W. LESLIE

11.2.1. Introduction

Data integration refers to the process of obtaining estimates of diffracted intensities (and their standard deviations) from the raw images recorded by an X-ray detector. As two-dimensional (2D) area detectors are almost universally used to collect macromolecular diffraction data, only this type of detector will be considered in the following analysis.

When collecting data with a 2D area detector, a decision has to be taken about the magnitude of the angular rotation of the crystal during the recording of each image. Two distinct modes of operation are possible: the rotation per image can be comparable to, or greater than, the angular reflection range of a typical reflection (coarse φ slicing), or it can be much less than the reflection width (fine φ slicing). The latter approach allows the use of three-dimensional profile fitting and, providing that the detector is relatively noise-free, improves the quality of the resulting data by minimizing the contribution of the X-ray background to the total measured intensity. However, there are significant overheads associated with recording, storing and processing the relatively large number of images that are required. Three-dimensional profile fitting is described in Chapter 11.3 and will not be discussed here.

11.2.2. Prerequisites for accurate integration

11.2.2.1. *Crystal parameters*

Only the integration procedure itself will be described in detail in this article. However, in order to obtain the highest quality data possible from a given set of images, there are a number of parameters that need to be determined in advance of, or during, the integration. The most important of these are the unit-cell parameters, which should be determined to an accuracy of a few parts in a thousand (or better). Post-refinement procedures (Winkler *et al.*, 1979; Rossmann *et al.*, 1979), which make use of the estimated φ centroids of observed spots rather than their detector coordinates, generally provide more accurate estimates than methods based on the spot positions. This is because spot positions are affected by residual spatial distortions (after applying appropriate corrections) and the cell parameters are correlated with the crystal-to-detector distance, which is not always accurately known. For either method, it is necessary to include data from widely separated regions of reciprocal space (ideally φ values 90° apart) in order to determine all unit-cell parameters accurately. This is particularly important for lower-symmetry space groups.

The crystal orientation also needs to be known to an accuracy that corresponds to a few per cent of the reflection width. For crystals with low mosaicity (*e.g.* 0.1°) this corresponds to a hundredth of a degree or better. Fortunately, it is a feature of post refinement that the error in determining the orientation is typically a few per cent of the reflection width, and so this condition can generally be met. It is important to allow for movement of the crystal by continuously updating the crystal orientation during integration. This is even true when using cryo-cooled crystals, as the magnetic couplings that attach the pin (holding the crystal) to the goniometer head are not strong enough to prevent small movements, particularly with the high angular rotation rates employed on intense synchrotron beamlines. Non-orthogonality of the incident X-ray beam and the rotation axis (if not allowed for) or an off-centre crystal will also give rise to apparent changes in crystal orientation with spindle rotation.

The crystal mosaicity can be estimated by visual inspection and refined by post refinement. Refined values are quite reliable when the mosaic spread is less than about 0.5°, but become more dependent on the rocking-curve model for the high mosaicities that are often associated with frozen crystals. The presence of diffuse scatter, which appears as haloes around the Bragg diffraction spots, presents further difficulties in determining the correct mosaic spread. When processing coarse-sliced images it is preferable to overestimate the mosaic spread slightly (rather than underestimate it). This will result in an increase in random errors (by adding in the X-ray background from an image on which the spot is not actually present), whereas using too small a value can give systematic errors (by underestimating the number of images on which the spot lies).

11.2.2.2. *Detector parameters*

Detector calibration is essential for high data quality. Both the spatial distortion and the non-uniformity of response of the detector must be accurately known, and it is equally important that these corrections are stable over the timescale of the experiment (and preferably for much longer).

Finally, the crystal-to-detector distance, the detector orientation and the direct-beam position must be refined and continuously updated during integration, using observed spot positions. The crystal-to-detector distance can vary during data collection if the crystal is not exactly centred on the rotation axis, and the direct-beam position can move after a beam refill at a synchrotron. For image-plate detectors with two (or more) plates, the direct-beam position and detector distance often differ slightly for different plates.

With appropriate care, it is normally possible to predict reflection positions on the detector to an accuracy of 20–30 μm, or a fraction of the pixel size, particularly for highly collimated X-ray beams available at synchrotron sources. This level of accuracy is necessary to minimize possible systematic errors, particularly in the case of profile fitting.

11.2.3. Methods of integration

There are two quite distinct procedures available for determining the integrated intensities: summation integration and profile fitting. Summation integration involves simply adding the pixel values for all pixels lying within the area of a spot, and then subtracting the estimated background contribution to the same pixels. Profile fitting (Diamond, 1969; Ford, 1974; Rossmann, 1979) assumes that the actual spot shape or profile is known (in two or three dimensions) and the intensity is derived by finding the scale factor that, when applied to the known (or standard) profile, gives the best fit to the observed spot profile. In practice, profile fitting requires two separate steps: the determination of the standard profiles and the evaluation of the profile-fitting intensities. As will be shown later, profile fitting results in a reduction in the random error associated with weak intensities, but offers no improvement for very high intensities.

11.2.4. The measurement box

X-ray scattering from air, the sample holder and the specimen itself gives rise to a general background in the images which has to be subtracted in order to obtain the Bragg intensities. Ideally, the background should be measured for the same pixels used to record the Bragg diffraction spot, but this is not usually practical and the background is determined using pixels immediately adjacent to the spot. In practice, the pixels to be used for the determination of

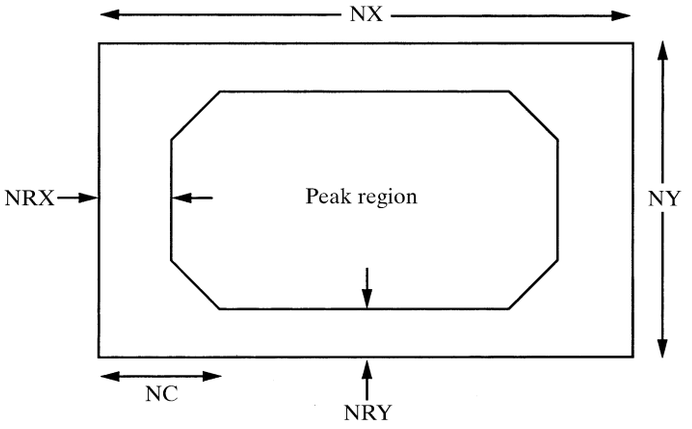

Fig. 11.2.4.1. The measurement-box definition used in *MOSFLM*. The measurement box has overall dimensions of NX by NY pixels (both odd integers). The separation between peak and background pixels is defined by the widths of the background rims (NRX and NRY) and the corner cutoff (NC). The size of the peak region is optimized separately for each of the standard profiles.

the background (background pixels) and those to be used for evaluating the intensity (peak pixels) are defined using a 'measurement box'. This is a rectangular box of pixels centred on the predicted spot position. Each pixel within the box is classified as being a background or a peak pixel (or neither). This mask can either be defined by the user, or the classification can be made automatically by the program. An example of a possible measurement-box definition is given in Fig. 11.2.4.1. The background parameters NRX, NRY and NC can be optimized automatically by maximizing the ratio of the intensity divided by its standard deviation, in a manner analogous to that described by Lehmann & Larsen (1974). It is generally assumed that the background can be adequately modelled as a plane, and the plane constants are determined using the background pixels. This allows the background to be estimated for the peak pixels, so that the background-corrected intensity can be calculated.

11.2.5. Integration by simple summation

11.2.5.1. *Determination of the best background plane*

The background plane constants a, b, c are determined by minimizing

$$R_1 = \sum_{i=1}^{n} w_i(\rho_i - ap_i - bq_i - c)^2, \qquad (11.2.5.1)$$

where ρ_i is the total counts at the pixel with coordinates $(p_i q_i)$ with respect to the centre of the measurement box, and the summation is over the n background pixels. w_i is a weight which should ideally be the inverse of the variance of ρ_i. Assuming that the variance is determined by counting statistics, this gives

$$w_i = 1/GE(\rho_i), \qquad (11.2.5.2)$$

where G is the gain of detector, which converts pixel counts to equivalent X-ray photons, and $E(\rho_i)$ is the expectation value of the background counts ρ_i. In practice, the variation in background across the measurement box is usually sufficiently small that all weights can be considered to be equal.

This gives the following equations for a, b and c, as given in Rossmann (1979),

$$\begin{pmatrix} \sum p^2 & \sum pq & \sum p \\ \sum pq & \sum q^2 & \sum q \\ \sum p & \sum q & n \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum p\rho \\ \sum q\rho \\ \sum \rho \end{pmatrix}, \qquad (11.2.5.3)$$

where all summations are over the n background pixels.

11.2.5.1.1. *Outlier rejection*

It is not unusual for the diffraction pattern to display features other than the Bragg diffraction spots from the crystal of interest. Possible causes are the presence of a satellite crystal or twin component, white-radiation streaks, cosmic rays or zingers. In order to minimize their effect on the determination of the background plane constants, the following outlier rejection algorithm is employed:

(1) Determine the background plane constants using a fraction (say 80%) of the background pixels, selecting those with the *lowest* pixel values.

(2) Evaluate the fit of all background pixels to this plane, rejecting those that deviate by more than three standard deviations.

(3) Re-determine the background plane using all accepted pixels.

(4) Re-evaluate the fit of all accepted pixels and reject outliers. If any new outliers are found, re-determine the plane constants.

The rationale for using a subset of the pixels with the lowest pixel values in step (1) is that the presence of zingers or cosmic rays, or a strongly diffracting satellite crystal, can distort the initial calculation of the background plane so much that it becomes difficult to identify the true outliers. Such features will normally only affect a small percentage of the background pixels and will invariably give higher than expected pixel counts. Selecting a subset with the lowest pixel values will facilitate identification of the true outliers. The initial bias in the resulting plane constant c due to this procedure will be corrected in step (3). Poisson statistics are used to evaluate the standard deviations used in outlier rejection, and the standard deviation used in step (2) is increased to allow for the choice of background pixels in step (1).

11.2.5.2. *Evaluating the integrated intensity and standard deviation*

The summation integration intensity I_s is given by

$$I_s = \sum_{i=1}^{m}(\rho_i - ap_i - bq_i - c), \qquad (11.2.5.4)$$

where the summation is over the m pixels in the peak region of the measurement box. If the peak region has mm symmetry, this simplifies to

$$I_s = \sum_{i=1}^{m}(\rho_i - c). \qquad (11.2.5.5)$$

To evaluate the standard deviation, this can be written as

$$I_s = \sum_{i=1}^{m}\rho_i - (m/n)\sum_{j=1}^{n}\rho_j, \qquad (11.2.5.6)$$

where the second summation is over the n background pixels.

The variance in I_s is

$$\sigma^2_{I_s} = \sum_{i=1}^{m}\sigma_i^2 + (m/n)^2\sum_{j=1}^{n}\sigma_j^2. \qquad (11.2.5.7)$$

From Poisson statistics this becomes

references