PROCESSING DIFFRACTION DATA WITH MOSFLM

ANDREW G.W. LESLIE AND HAROLD R. POWELL

MRC Laboratory of Molecular Biology, Hills Road,
Cambridge CB2 0QH, UK

Abstract: Processing diffraction data falls naturally into three distinct steps: First, determining an initial estimate of the unit cell and orientation of the crystal; second, obtaining refined values for these parameters; and third, integrating the diffraction images. The basic principles underlying autoindexing, parameter refinement, and spot integration by summation integration and profile fitting are described.

Keywords: data processing; profile fitting; autoindexing; postrefinement.

1. Introduction

This chapter will describe in outline the procedure for integrating monochromatic diffraction data from macromolecules. It is assumed that the diffraction images have been collected using the rotation method. Although the procedures will be described with reference to the MOSFLM program, the basic principles involved are common to most, if not all, data integration programs currently in use. More detailed accounts of many aspects of data processing are covered in the proceedings of a recent CCP4 Study Weekend [1].

2. Collecting the images

While the focus of this chapter is on data integration rather than data collection, it is worth emphasizing that successful data integration depends on the choice of appropriate experimental parameters during data collection. It is therefore crucial that the diffraction experiment is correctly designed and executed. A list of the most important issues that need to be considered is given below.

- Is the crystal single? Is the diffraction highly anisotropic? Two diffraction images 90° apart in phi should be examined carefully for evidence of split

R.J. Read and J.L. Sussman (eds.): Evolving Methods for Macromolecular Crystallography, 41–51

spots or the presence of a second lattice. A single image can easily be misleading in this respect.

- Can the image be successfully indexed? Failure of the indexing could indicate the presence of a second lattice. Does the derived cell and orientation account for all the spots on the image (with an appropriate mosaic spread)? Are there lines of weak spots between those predicted (indicative of a pseudocell)? Are there additional spots due to the presence of a satellite crystal?
- Does the crystal really diffract to the edge of the detector? If not, either increase the exposure time or move the detector further away to improve the data quality (signal to noise).
- Is the collimation adequate to resolve adjacent reciprocal lattice spots for the longest cell spacing? If not, move the detector further back or try reducing the incident beam size or, in some circumstances, the beam divergence.
- Is the dynamic range of the detector sufficient to avoid overloaded reflections at low resolution? If not, a rapid pass may be necessary to measure these strong reflections. Ideally, collect this rapid pass first.
- What is the optimum rotation angle per image? Too large a value will result in spatial overlap of spots in adjacent lunes. Too small a value will give a poor duty cycle, as the exposure time becomes comparable with the detector readout time. Very short exposure times (less than ~0.5 s) on modern synchrotron sources can lead to problems with shutter synchronization.
- Ideally, aim for high data multiplicity as this will improve the overall quality of the data by reducing random errors and facilitating outlier identification. If this is not possible, aim for high completeness, possibly by collecting several segments of data rather than a single large rotation. Be conservative in the choice of exposure time, so that the data set is complete before the onset of serious radiation damage.
- Always integrate at least some (and preferably all) the diffraction images during data collection, to check for unforeseen problems and to get a quantitative estimate of data quality. Soon it should be possible to do this automatically.

3. Determining the crystal cell parameters and orientation

The autoindexing algorithms currently in use are extremely powerful and in general it will be possible to determine the unit cell dimensions and crystal orientation from a single diffraction image, providing that the direct beam position, crystal to detector distance, and radiation wavelength are accurately known. Failure of the autoindexing can result from errors in these experimental parameters, the presence of a second lattice, or if only very few spots are available for the autoindexing. In the last case, inclusion of spots from two or more images should lead to success.

The most robust autoindexing algorithms employ a Fourier transform approach [2]. The general principle behind Fourier-based autoindexing can be understood as follows. Figure 1 indicates a situation where a "still" image (i.e., zero oscillation angle) has been taken with the crystal in an orientation such that a principle zone axis lies along the X-ray beam direction. The planes of reciprocal lattice points normal to this zone axis intersect the Ewald sphere in a series of concentric circles, centered on the direct beam position. In the diffraction image, a series of concentric lunes will be seen. Using the Ewald sphere construction, all of the spots on the detector can be mapped back to the reciprocal lattice points that gave rise

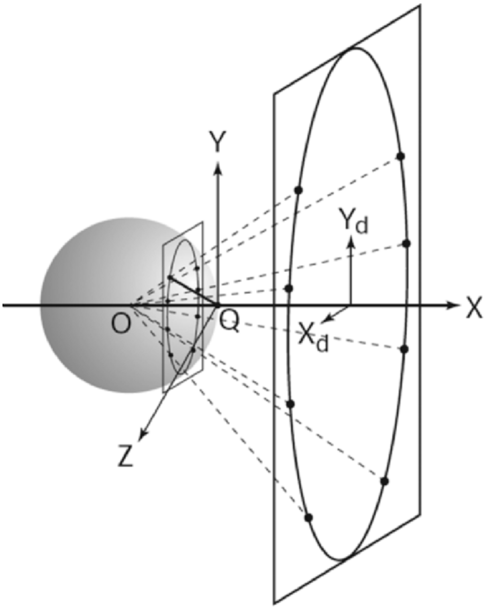

Figure 1. The Ewald sphere construction showing the geometry of diffraction for the case in which a principle zone axis lies along the X-ray beam direction. X, Y, Z define the laboratory coordinate frame. X-rays are parallel to the X-axis, and the phi rotation axis is along Z. The crystal sits at position O. A single plane of the reciprocal lattice normal to the X-ray beam is shown. Within this plane, reciprocal lattice points that lie on the Ewald sphere and are therefore in a diffracting condition are shown as dots. The resulting spots on the detector face are also shown. Knowing the detector geometry and the spot coordinates, the positions of the reciprocal lattice points can be calculated. A representative scattering vector (from the origin of the reciprocal lattice (marked Q) to the reciprocal lattice point) is shown.

to those spots (there is a small error involved in doing this, as the actual phi value for each reflection is not known). Now consider what happens when all of these "scattering vectors" are projected onto the zone axis. All the spots lying within the same lune will give rise to a projected vector of the same length. Thus, the projected scattering vectors for all the spots on the image will fall into clusters, where the separation between each cluster corresponds to the vector between adjacent reciprocal lattice planes. The Fourier transform of the projected clusters will form a series of regularly spaced spikes (Figure 2), where the distance between adjacent spikes corresponds to the real cell spacing along the principal zone axis direction. Now consider projecting the scattering vectors along a direction at an angle of (say) 20° to the true zone axis direction. In this case, spots in the same lune will project to give vectors of *different* lengths and so the Fourier transform of the projected scattering vectors will *not* have the clear set of maxima shown in Figure 2.

In practice [2, 3], the direction of the projection axis is varied in small angular steps (e.g., 2°) for the complete hemisphere of directions and in each case the Fourier transform of the projected scattering vectors is calculated.

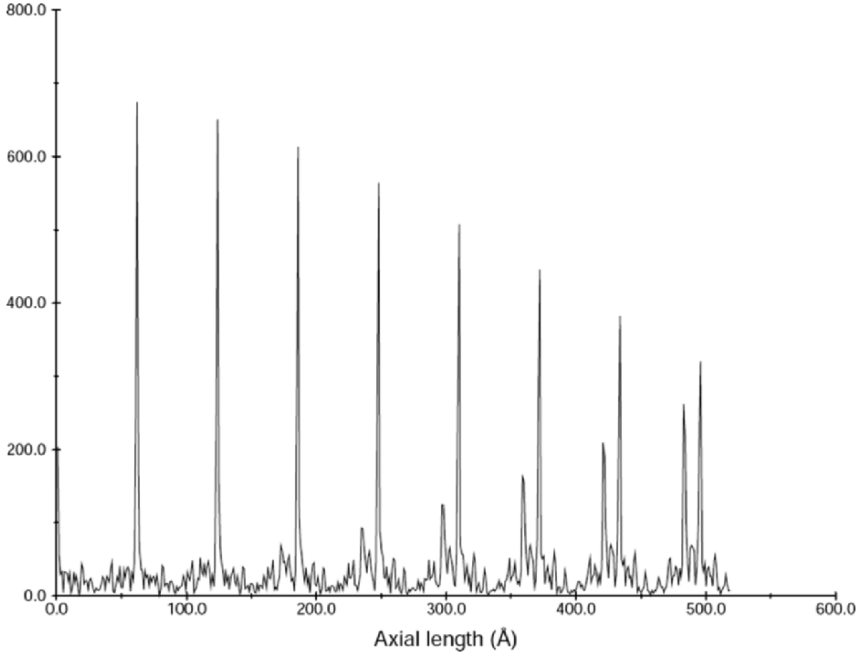

Figure 2. The Fourier transform of the projected scattering vectors for the case shown in Figure 1 will consist of a number of regularly spaced discrete maxima, where the spacing between adjacent peaks reflects the real cell spacing along the zone axis direction.

Then three directions are chosen from this list that have large maxima in the Fourier transform and reasonably large interaxial angles. These will define three principle zone axes and their repeats, thus defining a unit cell with which it should be possible to index all spots in the diffraction image. In general, the resulting unit cell will be a triclinic one that will not reflect the true symmetry of the lattice. The final stage is therefore to find the reduced cell from the chosen cell and then evaluate a "goodness of fit" to the 44 possible lattice types [4, 5]. The user is presented with a list of possible solutions, each with a corresponding quality index and, in general, the solution with the highest Bravais lattice symmetry that still has a good quality index will be chosen. It is important to realize that there is no information available at this stage on the *true* crystal symmetry, which can only be determined from the diffraction intensities. The spot positions only give information about the lattice symmetry, which can be higher than the true crystal symmetry. This is particularly important when considering the strategy for data collection. An incorrect assumption about the crystal symmetry may lead to the choice of a total rotation angle that is too small to collect all the unique data. For example, there are numerous examples of monoclinic crystals with a β angle very close to 90°. If the symmetry is incorrectly assumed to be orthorhombic and only 90° of data are collected rotating around the *b*-axis, then the resulting data will be very incomplete.

4. Parameter refinement

Once an orientation matrix and cell parameters have been derived from the autoindexing, these parameters (and others) are refined further using different algorithms. The parameters to be refined can be conveniently grouped into three classes:

- Crystal parameters: cell parameters, crystal orientation, and mosaic spread (isotropic or anisotropic)
- Detector parameters: the detector position and orientation and (if appropriate) distortion parameters (e.g., the radial and tangential offsets for the Mar image plate scanner)
- Beam parameters: the orientation of the primary beam and beam divergence (isotropic or anisotropic)

There are two complementary sources of information that can be used in the refinement; the spot coordinates measured on the detector, and the spot coordinates in phi. The latter can be measured empirically if the oscillation angle is much smaller than the reflection width, or can be estimated from the way in which the intensity for partially recorded reflections is distributed over the two (or more) images on which the reflection is recorded if the oscillation angle is comparable to, or greater than, the reflection width.

4.1. REFINEMENT USING SPOT COORDINATES MEASURED ON THE DETECTOR

The parameters are refined by least squares minimization of a positional residual:

$$\Omega_1 = \sum_i w_{ix}\left(X_i^{\text{calc}} - X_i^{\text{obs}}\right)^2 + w_{iy}\left(Y_i^{\text{calc}} - Y_i^{\text{obs}}\right)^2 \tag{1}$$

where X and Y are the spot coordinates on the detector, and w_{ix} and w_{iy} are appropriate weights.

Note that it is not possible to refine changes in crystal orientation around the rotation axis using this residual, as this parameter has no effect on the spot positions. Other parameters, such as cell dimensions and crystal to detector distance, may also be highly correlated (depending on the maximum Bragg angle).

4.2. REFINEMENT USING PHI COORDINATES

In this case, the residual to be minimized is given by:

$$\Omega_2 = \sum_i w_i\left[\left(R_i^{\text{calc}} - R_i^{\text{obs}}\right)/d_i^*\right]2 \tag{2}$$

where R_i^{calc} and R_i^{obs} are the calculated and observed distances of the reciprocal lattice point d_i^* from the center of the Ewald sphere (OP and OP′ in Figure 3) and again w_i is a weighting term. R_i^{calc} is determined from the current values for the cell parameters and crystal orientation. R_i^{obs} is obtained from the Φ centroid if fine Φ slices have been used. For coarse Φ slices, the position in phi of partially recorded reflections is estimated from the degree of partiality of the reflection (i.e., the way in which the total intensity is distributed between the two (or more) abutting images). This latter approach, known as postrefinement [6, 7] because it depends on knowing the integrated intensities, requires a model for the rocking curve, and permits refinement of either crystal mosaicity or beam divergence.

The effective radius of the reciprocal lattice point (see Figure 3) is given by

$$\varepsilon = \frac{\gamma d^*}{2}\cos\theta \tag{3}$$

where γ is the combined mosaic spread and beam divergence, d^* is the reciprocal lattice spacing and θ is the Bragg angle. The distance of the reciprocal lattice point from the Ewald sphere, Δr, is then given by

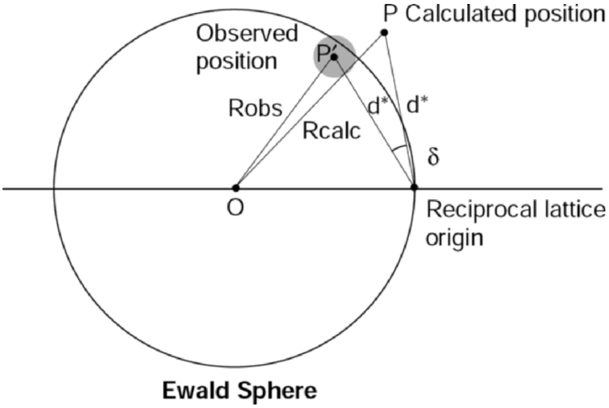

Figure 3. The large circle represents a section through the Ewald sphere, while the small shaded circle represents the position of a reciprocal lattice point at the end of an oscillation. A fraction of the total intensity corresponding to the volume of the reciprocal lattice point that has already passed through the Ewald sphere will be recorded on the current image. The remaining intensity will be recorded on the following image as P′ rotates clockwise.

$$P = 1/2\left[1 + \sin\left(\pi \Delta r / 2\varepsilon\right)\right] \tag{4}$$

$$\text{where}\, P = \frac{I_1}{I_1 + I_2} \tag{5}$$

and I_1, I_2 are the intensities recorded on the two abutting images (assuming the reflection only spans two images). Knowing P from the measured intensities, Δr can be calculated from equation 4, and thus R^{obs} can be determined. Rocking curve models other than the simple sine model in equation 4 have also been used. Because ε depends on the combined mosaic spread and beam divergence, this parameter can also be refined. (For fine Φ slices the mosaic spread or beam divergence is estimated from the observed reflection width in Φ.)

4.3. REFINEMENT STRATEGY

The refinement strategy can depend on how the data has been collected. If fine Φ slices have been used, accurate Φ centroids and coordinates (X, Y) are available for most strong reflections (excluding those very close to the rotation axis) and both residuals (Ω_1, Ω_2) can be minimized simultaneously using a suitable selection of reflections (strong and evenly distributed over the detector and in Φ). Problems arising due to correlations of different parameters can be avoided either by fixing some parameters or by the use of eigenvalue filtering. These problems can be particularly serious for low resolution data, where there is a strong correlation between crystal to detector distance and the cell parameters, or for an offset detector where there is a high correlation between the detector swing angle and the (horizontal) primary beam coordinate. If only a narrow Φ range of reflections is used in the refinement then some unit cell parameters will be poorly defined and may be correlated with the crystal setting angles, and there will also be a strong correlation between the detector orientation around the X-ray beam and the crystal setting angle around the beam. In such circumstances, the refined parameters may assume physically unrealistic values, but this will not necessarily affect the accuracy of the prediction of reflection positions and widths.

When the data is collected with coarse Φ slices, only fully recorded reflections will give accurate spot positions (X, Y), and accurate Φ centroids can only be determined for partially recorded reflections. In MOSFLM, the two residuals are currently minimized independently. Only the detector parameters are refined when minimizing the positional residual, and only cell, orientation and optionally beam parameters are refined against the angular residual. This approach does have the advantage that the accuracy of the refined cell parameters does not depend on the accuracy of the crystal to detector distance or direct beam position, providing these are known sufficiently well to allow correct indexing of the reflections.

5. Integration of the images

Once accurate values for the crystal cell parameters and orientation have been obtained, the images can be integrated. Stated in the simplest way, this procedure involves predicting the position in the digitized image of each Bragg reflection present on that image, and then estimating its intensity (after subtracting the X-ray background) and an error estimate of the intensity. In practice, this apparently simple task is quite complex.

5.1. PREDICTING REFLECTION POSITIONS

A knowledge of the crystal cell and orientation will allow the prediction of spot positions on a “virtual detector,” i.e., a detector whose position and orientation are exactly known. These positions must then be mapped onto the digitized image, and this mapping must take into account any spatial distortions introduced by the detector, either using a predetermined calibration table or by refining the distortion parameters for each image.

5.2. DEFINING THE PEAK/BACKGROUND MASK

Because it is physically impossible to measure the X-ray background actually under the diffraction spot (which strictly is what is required to obtain the background subtracted intensity) the background is measured in a region around the spot either in two dimensions (X, Y, the detector coordinates) for coarse Φ slices or in three dimensions (X, Y, and Φ) for fine Φ slices. A background plane is fitted to these background pixels, and this plane is then used to estimate the background under the spot. To do this it is necessary to define a pixel mask, which, when centered on the predicted position of the spot, will define which pixels are to be considered as part of the peak and which are to be used to determine the background (see Figure 4).

This mask can be defined by the user after visual inspection of the spot shapes, but MOSFLM will automatically optimize the peak/background definition. It is clearly important that pixels are not misclassified, as this can lead to systematic errors in the integrated intensity. The presence of strong diffuse scattering, which is quite commonly observed with synchrotron data, can lead to difficulties in differentiating between peak and background pixels. Unfortunately, there is no simple way of dealing with this problem.

5.3. SUMMATION INTEGRATION AND PROFILE FITTING

Having determined the background plane, the simplest way to obtain an estimate of the integrated intensity is to sum the pixel values of all pixels in the peak area of the mask, and then subtract the sum of the background values calculated from the background plane for the same pixels. This is known as summation integration and for spots where the background level is very low

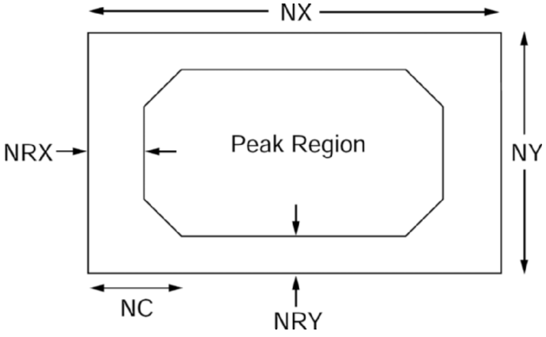

Figure 4. The peak/background mask definition used in MOSFLM. The overall mask size (in pixels) is defined by NX and NY, and the differentiation between peak and background pixels is defined by a background rim in X and Y (NRX, NRY pixels) and a corner cutoff (NC pixels).

compared to the intensity of the spot this will give as accurate an estimate of the intensity as it is possible to get. (In such cases, the accuracy is determined by counting statistics, so for a total count of N photons the standard deviation is $\sqrt{N}$).

For weaker reflections, it is possible to get a more accurate estimate of the integrated intensity by using a procedure known as profile fitting [8–11]. In this procedure, it is assumed that the shape or profile (in two or three dimensions) of the spots is known. The background plane is determined in the same way as for summation integration, but the intensity is derived by determining the scale factor which, when applied to the *known* spot profile, gives the best fit to the *observed* spot profile. This scale factor is then proportional to the profile fitted intensity for the reflection. In practice, the fitting is done by least squares methods, to minimize the residual

$$R = \sum_{\substack{\text{peak}\\\text{pixels}}} w_i \left(X_i - KP_i \right)^2 \qquad (6)$$

where

X_i is the background subtracted intensity at pixel i

P_i is the value of the standard profile at the corresponding pixel

w_i is a weight, derived from the expected variance of X_i

K is the scale factor to be determined.

The improvement gained by profile fitting depends on the spot intensity relative to background and the spot shape, but typically it can provide a reduction in variance by a factor of 2 (1.4 in the standard deviation) for weak reflections. This is a significant gain, and all modern software packages employ profile fitting, although the implementation differs in detail.

The procedure assumes that the *true* reflection profile is known. In practice, this is determined from the observed reflection profiles of a number of reflections in the immediate vicinity of the reflection being integrated. An appropriate weighted sum of the individual profiles is used to form the “true” or standard profile. The reflection shape will vary with position on the detector (due to changes in obliquity of incidence and other factors) and it is important to allow for this. MOSFLM determines a “standard” profile for several defined areas and then calculate the best profile for each reflection as a weighted mean of the closest “standard” profiles.

Profile fitting is a powerful technique for reducing the random error in weak diffraction data, but equally an error in determining the standard profiles will lead to systematic errors in all measured intensities. Modern software packages go to some lengths to minimize the magnitude of the systematic errors introduced by the use of nonideal standard profiles.

5.4. STANDARD DEVIATION ESTIMATES

It is important to obtain reasonable estimates of the standard deviations of the integrated intensities, since these are used as weights when merging multiple observations, and in subsequent steps of the structure determination (e.g., identification of heavy atom derivatives, heavy atom parameter refinement, and model refinement). For summation integration, a standard deviation can be obtained based on Poisson statistics, while for profile fitted intensities the goodness of fit of the scaled standard profile to the true reflection profile can be used. These will generally underestimate the true errors, as they take no account of systematic errors arising from effects such as absorption, beam instability, detector nonlinearity, or errors in nonuniformity corrections. The standard deviation estimates should therefore be modified when the data is merged, making use of the *observed* agreement between multiple observations.

References

1. Multiple contributions (1999) *Acta Crystallographica*, **D10**: 1631–1772.
2. Steller, I. et al. (1997) An algorithm for automatic indexing of oscillation images using Fourier analysis. *Journal of Applied Crystallography*, **30**: 1036–1040.
3. Powell, H.R. (1999) The Rossmann Fourier autoindexing algorithm in MOSFLM. *Acta Crystallographica*, **D10**: 1690–1695.
4. Burzlaff, H. et al. (1992) *International Tables for Crystallography*, vol. A. Edited by T. Hahn. Dordrecht: Kluwer Academic, pp. 737–749.
5. Kabsch, W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of Applied Crystallography*, **26**: 795–800.
6. Winkler, F.K. et al. (1979) The oscillation method for crystals with very large unit cells. *Acta Crystallographica*, **A35**: 901–911.
7. Rossmann, M.G. et al. (1979) Processing and post-refinement of oscillation camera data. *Journal of Applied Crystallography*, **12**: 570–581.
8. Diamond, R. (1969) Profile analysis in single crystal diffractometry. *Acta Crystallographica*, **A25**: 43–54.
9. Ford, G.C. (1974) Intensity determination by profile fitting applied to precession photographs. *Journal of Applied Crystallography*, **7**: 555–564.
10. Rossmann, M.G. (1979) Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. *Journal of Applied Crystallography*, **12**: 225–238.
11. Leslie, A.G.W. (1999) Integration of macromolecular diffraction data. *Acta Crystallographica*, **D10**: 1696–1702.