Data Processing in Macromolecular Crystallography

Andrew J. Howard
Department of Biological, Chemical, and Physical Sciences
Illinois Institute of Technology Chicago IL 60616 USA
ahoward@harry.csrri.iit.edu

Abstract

Data processing in macromolecular crystallography is the effort by which a user converts a set of raw diffraction data into a list of Bragg reflections with measured intensities. With modern crystallographic hardware the raw data consists of a set of two-dimensional detector images, each collected at a particular orientation of the crystal. Data processing may be broken down into four steps: calibration, determination of the unit cell, measurement of the integrated intensities, and merging and scaling the integrated data. Algorithms for each of these steps have been derived and validated, and are implemented in several commercially available software packages. In recent years these packages have become more flexible and easier to use. Software developed for monochromatic laboratory experiments must be modified or even replaced when the user is performing more exotic experiments, particularly Laue diffraction experiments.

1 Introduction

Macromolecular crystallography is a relatively mature science in that the procedures used, at least for monochromatic experiments, are well-defined and well-validated through thirty-five years of experience in several hundred laboratories. The fundamental experiment is little different from that performed by John Kendrew in the 1950's to solve the myoglobin structure. In this basic experiment, a respectable-sized single crystal of the macromolecule is grown and mounted on a rotation stage in front of an X-ray source; the crystal is exposed to X-rays as it is rotated about an axis, and some sort of detector is placed where it can intercept the diffracted X-ray beams; the scientist determines the indices (hkl) of the Bragg diffraction spots in the pattern; the background-subtracted intensity of the spots is determined; various corrections are applied to these measurements; the intensity measurements of the current sample are compared or combined with those from related experiments to derive phase information; and the intensity measurements are combined with the phase information to provide input to a Fourier transform from which the electron density can be visualized and interpreted.

The data processing procedures used routinely today are somewhat newer than these experimental techniques, but even in this realm there is increasing agreement among practitioners about goals and approaches. This convergence of methodology has arisen not because of intentional collusion among software developers, but rather by a realization that there are optimal ways of addressing the basic problems of data processing. The software packages available today differ in their user interfaces and in their relative applicability to different styles of data collection (e.g., one might be better suited to wide-slice data collection protocols, and another might be better-suited to narrow-slice protocols), but there is a substantial degree of unanimity about what the software needs are and how to address them.

This paper summarizes the procedures required for processing macromolecular diffraction data. I will draw most examples from my own X-GEN software, but some discussion of other packages will be provided as well. Almost all of this discussion centers on processing of data from electronic detectors (both digital and analog) and image plates. Data processing for single-counter diffractometry has been covered in extensive detail in previous IUCr Schools (1) and elsewhere (2). Film processing has also been carefully discussed in other venues (3), but the discussion of "wide-slice" image-plate data processing below applies to screenless oscillation photography with the understanding that some of the errors in film are worse than with image plates. The goal of data processing, as defined here, is to obtain appropriately corrected estimates for the intensities of the Bragg reflections contained in a set of two-dimensional detector images.

2 Historical background

In the 1950's and 1960's, macromolecular crystallographic data were collected either by precession methods onto film or by single-counter diffractometry. The former allowed for simultaneous observation of numerous Bragg spots (usually within a single layer) but suffered from the built-in limitations of film, viz. limited dynamic range, film fog, film shrinkage, and the necessity of a time-consuming development step subsequent to data acquisition. By the 1960's diffractometry was at least partially automated, allowed more accurate measurements than were possible with film, and provided for a wide dynamic range, but has a huge limitation with respect to medium-sized to large proteins: it can only measure one Bragg peak at a time. The accuracy of diffractometry and its ease of use made it the method of choice on the small (typically less than 34 kDa per asymmetric unit) proteins that were the targets of crystallographic studies in the period. Film methods rallied in popularity with the advent of screenless precession (4) and screenless oscillation (5) methods, which allowed more efficient data collection and simpler apparatus, but the limitations of film remained evident. It was clear that users would benefit from the development of a method that would provide the efficiency of film and the accuracy and automaticity of diffractometry. The "best of both worlds" would thus be a method of electronic detection that combined the advantages of both existing techniques.

Thus the crystallographic community exhibited substantial enthusiasm for the first efforts to develop area detectors. The flat multiwire proportional chambers developed at UC San Diego (6) and the University of Virginia (7), the spherical drift chambers developed at MIT (8) and LURE (9) and the SIT tube television detectors developed at Cambridge (10) and Brandeis(11) provided performances approaching the "best of both worlds" goals delineated above. It was recognized from the outset that the rapidity and accuracy of the data collection hardware would provide little help to the user unless substantial resources were devoted to software development, both for data acquisition and control and for data processing. In order for an area detector to be useful there needed to be a way of extracting useful information from the raw data it generated, and that depends on software. Thus most of the groups mentioned above developed complete software packages for processing data from their detectors. In most cases there was little distinction made between data acquisition code and data processing code; the packages were set up to provide for both. Since the coupling between conducting the experiment and deriving intensity information from it was high, there was little need to separate the software into separate components.

In the mid- to late-1980's several of these detector designs were commercialized, and structure-determination laboratories, as distinct from instrument-development groups, began to use them. Shortly thereafter, researchers recognized that the scope and requirements for data processing on all these detector systems were fairly similar, and that it would pay off to write detector-independent data processing code. The concept of separate packages for data acquisition and data processing arose in the same period: under some circumstances users were able to acquire their data with the expectation that the data could be processed using more than one package--an option not available if the data acquisition and data processing are indissolubly linked. Comprehensive packages that accomplish both data acquisition and data processing are still viable, but most users of area detectors employ separate software for the two tasks.

Software packages developed during this period included MADNES, a package originally developed by James Pflugrath and Albrecht Messerschmidt at the Max-Planck Institut in Martinsried (12) and subsequently incorporated into a European Community-supported software initiative under the direction of Gerard Bricogne (13). MADNES was originally applied to data from the Enraf-Nonius "FAST" SIT-tube detector system, and has since been used on several other analog and digital detector systems. During the same period I developed *XENGEN* (14), a package for processing data from the Xentronics (subsequently Nicolet and then Siemens) Area Detector. Wolfgang Kabsch developed XDS (15) for processing Xentronics data, and his software was subsequently used on other systems. MADNES, *XENGEN*, and XDS employ explicitly three-dimensional data processing paradigms; reflections are expected to extend over several consecutive detector images, and the images are assumed to cover contiguous ranges of scanning angle. Thus in this "narrow-slicing" paradigm, reflection positions can be calculated as centroids in scanning-angle w as well as detector position (X,Y), and integrated intensities can be calculated by summation over ranges of scanning angle as well as detector position. All of these packages employ some form of three-dimensional profile analysis: the expected three-dimensional profile of each Bragg reflection is determined and compared with the actual background-subtracted profile, and the shape of the model profile is used to help inform the intensity measurement.

The appearance of storage-phosphor detectors or image plates in the 1980's led to further software refinements. Most image-plate users collect data over fairly wide oscillation ranges (> 1 degree per image), so that most Bragg reflections are entirely recorded on a single image. The data resemble those from screenless oscillation photography; indeed, image plates function as "electronic film". Thus software originally developed for film processing were applied with modest modifications to the task of processing image plate data. Corrections for partiality (16) originally developed for film processing were extended to this type of data. Zbyszkek Otwinowski's DENZO (17) is the most successful of these "two-dimensional" or "wide-slice" data processing packages. The two-dimensional profile-fitting techniques developed by Rossmann (16) and others for film work were adapted in DENZO and similar software to the image-plate application.

In the 1990's even the distinction between two-dimensional and three-dimensional data collection was seen to permit further merging of concepts. With some care one can treat two-dimensional profile-fitting as a special case of three-dimensional profile-analysis, and the conceptual differences between wide-slice and narrow-slice data collection can be rendered unimportant. My X-GEN package can be used comfortably for processing both wide-oscillation and narrow-oscillation data, and other recent packages have similar capabilities. Meanwhile the importance of providing graphics support for the underlying algorithmic functions was recognized, and several packages were expanded to incorporate graphics. Thus HKL, the successor to DENZO; X-GEN, the successor to *XENGEN*; D*Trek, the successor to MADNES; and the packages offered by the detector manufacturers all provide graphical user interfaces and (generally) flexible utilities for viewing detector images and visual clues to the quality of the processed data. Increasingly the algorithms described in the next section are employed in all the major software packages. The differences are in the user interfaces, the relative emphases on two- and three-dimensional approaches, and the number of choices the user is allowed or required to make in processing the data.

3 Methods

The steps required in deriving a set of intensity estimates for Bragg reflections are as follows:
(1) calibration of the experimental arrangement;
(2) determining the sample's unit cell;
(3) determining the integrated intensities of the Bragg spots;
(4) merging and scaling the integrated data;
Each of these steps comprises several sub-tasks, as set forth below.

3.1 Calibration

Electronic and image-plate detectors do not offer perfectly uniform responses to the X-ray beams that impinge on them, and the purpose of the calibration is to characterize this response so that the raw detector images can be appropriately massaged before any intensity measurements are made. Detectors vary widely in the degree to which these nonuniformities affect the data, and the corrections assume varying significances depending on the detector type.

The first type of calibration to consider is correction for spatial distortion. Most electronic detectors, with the exception of the San Diego Multiwire Systems instrument, show substantial geometrical distortions. Thus the mapping from pixel position (X,Y) to a Cartesian position (x,y) measured in centimeters from a reference point on the detector face may involve a complex transfer function

$$(x,y) = T(X,Y) \qquad (1)$$

With most electronic detectors the form of the function T is derived from an experiment in which a metal plate is attached to the front of the detector and X-rays are trained upon the detector from a point source. The plate has n small holes (i = 1,2,...n) drilled in it in a known pattern, and the centroids (Xi,Yi) of the spots produced as X-rays travel through the holes onto the detector face are recorded. The Cartesian positions (xi,yi) of the holes is known in advance, so as long as one can identify which centroid corresponds to which Cartesian position, the mapping from (xi,yi) to (Xi,Yi) is well-defined. The transfer function T is simply the inverse of this mapping, appropriately interpolated. In some software packages the function T is calculated as a polynomial expansion to several orders; in others, including X-GEN, a simple lookup table is stored and the interpolation is performed with a two-dimensional spline.

The second type of calibration to consider is for nonuniformity of response to X-rays. With some detectors, particularly SIT-tube systems and the recently introduced charged-coupled device systems, the signal produced on the detector face when an X-ray photon arrives varies substantially from point to point across the face. We can separate the nonuniformity of response into a local component and a regional component. The local component characterizes the difference in response between two pixels close together on the face; the regional component describes variations over large fractions of the detector area. In principle one could correct for both local and regional variations by collecting a long flood-field image of the detector, compensating for variations in the distance from the source to each pixel, and then defining the nonuniformity of response to be proportional to the observed count at each pixel in the flood field. There are two difficulties with this approach. First, the time required to achieve adequate counting statistics in each pixel may be inconveniently long: with typical point sources it may take as long as one day to accumulate enough X-ray photons in each pixel (e.g., 10000 counts per pixel to derive a correction that is accurate to 1%) to derive a useful correction. Second, the local efficiency nonuniformity is, in this approach, convoluted with a nonuniformity in the effective pixel size from pixel to pixel, and the measured count in a pixel may be affected more by this spatial nonuniformity than it is by the response nonuniformity. Both of these problems disappear if we focus only on the regional nonuniformity. In this case the counting statistics even in a brief flood field are adequate if we compute moving averages of the flux over regions of the detector face, and the local variations in pixel size will average out. Therefore a regional correction is reasonably easy to derive. Success in performing local corrections has been achieved in some packages (not including X-GEN), but regional corrections are probably more important in any case.

A third type of calibration is a determination of the active area of the detector. In most experiments some fraction of the pixels on the detector face are unavailable for data integration, either because there is no way for X-rays to produce counts in a region or because an occlusion blocks the path from the sample to the region. On the Siemens Multiwire or MAR Research detectors, for example, the detector face is actually round, but the electronic recording area is rectangular; thus the corners are outside the active area. In most experiments the shadow of the beamstop and the hanger from which it is suspended appear on the detector image, so those areas are unavailable for data collection. In some experiments the shadows of goniostat motors can occlude portions of the detector face for at least portions of data runs. The easiest way to detect these inactive areas of the detector is to obtain one or more detector images, remove the Bragg reflections from them, and then determine where the low counts (counts well below the mean) appear. Thus if after removing reflections from a series of images a summed "background" image C(X,Y) is obtained, we can determine the mean background:

$$<C> = \Sigma\, C(X,Y) / N \qquad (2)$$

where N is the number of pixels summed. Then a pixel (X,Y) is defined as inactive if

$$C(X,Y) < Z * <C> \qquad (3)$$

In X-GEN, Z is a user-adjustable constant, but it generally lies around 0.4. X-GEN and some other packages provide for manual adjustment of the active area with a command-driven or graphical interface. These manual adjustments can permit elimination of regions where the count is unusually high rather than low, e.g. when a deliberately leaky beamstop is employed; they also allow for special experiments, such as experiments in which a portion of the detector needs to be excluded from consideration.

3.2 Determining the Sample's Unit Cell

In the early days of area detectors users tended to study the same protein for many months, and their knowledge of the unit cell and the orientation of the crystal was sufficiently intimate that an auto-indexing capability was seen as unnecessary. But as more laboratories began using area detectors, and more attention to automation and convenience was given, the need for auto-indexing capabilities arose. Even before auto-indexing became routine, though, software needed to be used to refine manually determined orientation matrices and unit cell values. These required methods of gathering lists of bright spots to be used as input to the refinement. These same lists of bright spots were used in auto-indexing, once the latter was implemented in software packages like *XENGEN*, MADNES, and DENZO.

In order to obtain the data necessary for auto-indexing and refinement, we must obtain a list of centroids in detector position (X,Y) and in scanning angle ω for some reflections. In DENZO and HKL, as in other film-based packages, the centroids are obtained very precisely in (X,Y) and the expectation is that indexing will be accomplished with only one image. The potential ambiguity associated with all the scanning angle values being equal is avoided because the curvature of reciprocal space is sufficient to provide all three basis vectors for the complete unit cell. In X-GEN and other three-dimensional packages, full three-dimensional centroids are obtained and used in calculating diffraction vectors all referenced to a common origin in reciprocal space. If the sample goniostat includes more than one rotatable axis, the sample data can even be taken from more than one range of the non-scanned goniostat angles. Thus an orientation matrix computed over an omega range with chi and phi set to particular values can be made even more accurate if an additional set of data collected at different chi and phi values is included in the determination.

These facilities for finding bright spots function in fairly similar ways. Typically the detector is divided into moderate-sized regions, and estimates of the background in each region are obtained from the mean, the median, or even the mode of the counts in the region. Then places where the count is substantially above that background are identified. Groups of neighboring pixels that are all above background are assumed to be parts of the same reflection; an isolated bright pixel is assumed to be an error and is ignored. The centroid in (X,Y) is then obtained for these groups of pixels. In packages like MADNES and X-GEN, the two-dimensional centroid is extended to three dimensions by examining the neighborhood in omega of any (X,Y) centroid and re-doing the centroid calculation in a three-dimensional summation box. Spots with oddly-shaped profiles are filtered out, as are spots that are too close to excluded regions of the detector or too close to the rotation axis.

Methods for auto-indexing developed for small-molecule crystallography depend on recognizing the integer-like behavior ("graininess") of the diffraction vectors associated with the individual reflections. Thus if we define $\mathbf{s} = (s_x, s_y, s_z)$ as the diffraction vector associated with the reflection $\mathbf{h} = (h,k,l)$, then if A is the reciprocal-space unit cell matrix for the sample, then

$$A = \begin{pmatrix} a^*_x & b^*_x & c^*_x \\ a^*_y & b^*_y & c^*_y \\ a^*_z & b^*_z & c^*_z \end{pmatrix} \qquad (4)$$

and

$$\mathbf{s} = \lambda\, A * \mathbf{h}, \qquad (5)$$

where λ is the wavelength at which the data were collected. Thus in small-molecule methods the diffraction vectors s for a set of reflections are calculated, and the indices for a few (three or four) of these reflections are guessed. If we use three reflections ($\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$), then the orientation matrix A* associated with any guess of the indices can be computed by converting the equation above into a 3x3 matrix equation:

$$S = \lambda\, A * H, \qquad (6)$$

where $H = (\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$ and $S = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3)$. This matrix equation can be inverted if H is nonsingular to obtain A, and the results can be used to index all other reflections. If the indices of the resulting reflections are close to integers, then A* is probably correct, and we can proceed to refinement. If the indices are not close to integers, a different assignment of the guess H can be chosen until a success arises.

This method has difficulty with macromolecular data because the density of spots in reciprocal space is high. The low-resolution spots, for which the integerness of the reflections would be easy to recognize, are often unavailable because they are obscured by the beamstop or because their profiles are too broad in omega or (X,Y). Thus a modest modification of the method is used, in which the diffraction vectors s are replaced with the differences between nearby diffraction vectors $\mathbf{u} = \mathbf{s} - \mathbf{s}'$ and three small difference vectors are chosen as references: $S = (\mathbf{u}1\ \mathbf{u}2\ \mathbf{u}3)$. This has the advantage that small errors in the reflection centroids, as expressed in their diffraction vectors, tend to be subtractive rather than additive or small distances on the detector face. Thus the **u** vectors are actually more accurate than the **s** vectors, and the chance of finding a valid solution is substantial. Of course the resulting orientation matrix A* may be correct even if all of the computed indices are off by constant offset vector dh; this can arise in the common situation in which the user does not know where the beam center position is, and estimates it incorrectly so that all the k indices (for example) are one unit too high or too low. These problems generally can be discovered after the fact. This method of examining differences between diffraction vectors rather than the diffraction vectors is known as the "difference vector" method, and it is implemented in most macromolecular packages. Some offer a choice of a traditional diffraction-vector or "Sparks" algorithm or the difference-vector algorithm.

Problems with auto-indexing in macromolecular crystallography are generally traceable to one of four sources: (a) incorrect choice of reference spots, either because the centroids are mis-measured or because two neighboring reflections are merged by the software to create a bogus reflection somewhere between the two; (b) translational offsets that allow the difference vector algorithm to succeed but do not produce correct indexings of the spots because of a constant offset; and (c) the presence of spots that belong to other lattices, typically from a small crystallite of the same macromolecule that is oriented a few degrees away from the principal crystal. Problem (a) can usually be resolved by adjusting parameters in the software to prevent miscalculations of the centroid. Problem (b) can be overcome after the difference-vector auto-indexing is complete, often by trial and error. Problem (c) can be addressed by using only a higher-resolution subset of the data, wherein the smaller crystallite contributes no reflections. In an extreme case of this method, I once indexed a severely cracked subtilisin sample in which two lattices, one diffracting to 1.8 Å and one only to 2.4 Å, were both diffracting. By indexing the data from 2.2 Å to 1.8 Å, I was able to determine the primary crystallite's unit cell. Then by eliminating from the set of sample data all reflections that indeed could be indexed as integers by the primary orientation matrix, I was able to index the other crystallite as well. We then proceeded to integrate the data from both crystals and successfully refined the structure.

Recognition of a correct solution in auto-indexing depends on the integerness of the computed (hkl) values for the observed reflections or difference vectors. The most obvious way to recognize integerness is by a low value of an integerness residual:

$$\Sigma\, (h - [h])^2 + (k - [k])^2 + (l - [l])^2 \qquad (7)$$

where [x] means "integer closest to x". Bricogne has suggested a more sophisticated metric, namely, minimization of a Fourier expression

$$\Sigma \exp(-2\pi ih)^* \,\Sigma \exp(-2\pi ik)) * \Sigma \exp(-2\pi il) \ (8)$$

This formulation has been implemented in recent versions of MADNES and offers some advantages over the previous formulations.

Once the auto-indexing effort has succeeded, the user is presented with an approximate set of unit cell lengths and angles. The unit cell obtained from the auto-indexing may not be the Delauny reduced cell; it may not even be the crystallographically conventional unit cell. Several packages, including XDS and DENZO, provide tools for recognizing and installing the crystallographically conventional cell rather than the cell that the auto-indexing effort happens to find. In some cases the probabilities that the cell and symmetry operators are correct are presented to the user. In other packages, including X-GEN, the user must supply some crystallographic know-how to recognize the proper unit cell.

The unit cell will be expressed as unit cell lengths and angles, plus three angular values that describe the orientation of the cell with respect to the laboratory. These latter angles are expressed differently in different packages. In *XENGEN* and X-GEN the Euler angles of a fictitious goniostat, sometimes aligned with the real goniostat and sometimes aligned with such that the fictitious (ω=0,χ=0,ϕ=0) position corresponds with the start of the data run, are calculated such that a view down the Z axis of the fictitious goniostat with X vertical and Y horizontal will display the real-space a axis along X and the real-space b along Y (assuming γ = 90•). In several other packages a nominal zero position of the crystal is determined and a set of offset angles (RotX, RotY, RotZ) are calculated from the orientation matrix. Since both of these methods involve, in effect, finding the eigenvectors of a 3x3 rotation matrix, their behavior is computationally very similar. Each has some advantages in visualizing the relationship between the sample and the laboratory axes.

Following the auto-indexing and the cell conversion, if any, the user will need to improve the unit cell estimates by refinement. Typically the refinement takes the form of a minimization of differences between observed and calculated values of the spot positions in (X,Y) and in scanning angle, or by minimizing the non-integerness of the reflection indices. These minimizations can be performed by analytical calculation of first derivatives of the target functions with respect to the refinable parameters and conventional linear least-squares, or by a non-derivative based method like Simplex.

Thus in the conventional least-squares technique for integerness, we assume that the integer closest to an index is actually its correct value, and write

$$\mathbf{h} = R * \mathbf{s} \qquad (9)$$

where $\mathbf{h}$ is the (hkl) vector of the reflection, $\mathbf{s}$ is its scattering vector, and R is the real-space unit cell matrix. Note that $R^{-1} = \lambda A$ by our earlier discussion. The diffraction vector can be calculated from the (X,Y) position of the spot on the detector, together with its scanning-angle position, provided that we know the position of the detector in space and the transfer function T described above. Then if in fact the integer version of $\mathbf{h}$, [$\mathbf{h}$], is correct, then

$$\mathbf{h} - [\mathbf{h}] = \mathrm{err}(\mathbf{h}) = \Sigma_i\ (\delta(p_i) * \partial(R*\mathbf{s})/\partial p_i \qquad (10)$$

where the parameters pi are the refinable parameters and $\partial(R*\mathbf{s})/\partial p_i$ is the partial derivative of $R*\mathbf{s}$ with respect to the parameter p_i. The unit cell parameters a,b,c,α,β,γ, and the rotation angles only affect R, whereas the detector parameters (see below) only affect the mapping from observed (X,Y,ω) positions to diffraction vectors, so they only affect $\mathbf{s}$. By the chain rule, we may separate these in simple ways. Both the linear least-squares and the Simplex options are available in X-GEN, where the integerness residual can be minimized by conventional least-squares, and the scanning-angle and (X,Y) residuals are minimized by Simplex. The latter methods are slower but tend to avoid false minima. More sophisticated approaches allow for rescaling of the least-squares normal matrix so that metrics can remain balanced, or they provide for eigenvalue filtering to prevent non-physical excursions of highly correlated parameters. In practice the simpler methods can avoid these same problems if the user recognizes the correlations and “turns off” simultaneous refinement of highly correlated parameters.

The parameters available for refinement are the cell lengths and angles; the orientation angles (Eulerian or rotation); and some parameters describing the position and orientation of the detector. In MADNES and D*Trek these parameters are all treated in a comprehensive vectorial notation (19), but in most other packages the physical parameters (e.g. the distance from the sample to the detector) are called out more explicitly. In X-GEN five detector parameters are used: the sample-to-detector distance; two offsets describing the translation of the detector relative to an origin in its plane; a rotation of the detector axes with respect to the sample goniostat’s rotation axis; and the “two-theta” angle between a normal to the detector face through the sample and the direct beam. At least two other angular parameters could in principle be refined: the angle between the direct beam and the rotation axis, and the angle between the normal to the detector face and the rotation axis. In most experiments these angles are close to 90• degrees, and the errors introduced in the problem by their deviation from 90• are small except over wide scanning ranges. Since we expect to refine orientation parameters during integration “on the fly” (see below), this causes few problems.

3.3 Determining integrated intensities

Once the sample’s unit cell and orientation are determined and refined, the user can measure intensities for the Bragg spots expected within the range of detector images for which data are available. This task is the heart of the processing effort because the most important result of data processing--the measurement of the I(hkl) values--arises from it. During data processing, moreover, some of the operational parameters relevant to the integration must be tracked so that we can follow time-dependent or scanning-angle-dependent changes in the state of the experiment. Thus the integration step is probably the most algorithmically complex as well as the most important. The sub-steps involved in reflection integration therefore include (a) pre-determining the positions of the spots to be integrated; (b) defining the range of pixels over which the summation or profile analysis is to be performed; (c) defining the two- or three-dimensional model profile with which each reflection's own background-subtracted profile is to be compared; (d) estimating the background under each pixel in the relevant range; (e) performing the sums necessary for the integration; (f) applying pre-determined multiplicative corrections to the sum; (g) updating dynamic estimates, including the model profiles, the background estimates, and the unit cell and orientational parameters.

The positions (X,Y,ω) of the reflections can be pre-determined, so that the summations that lead to the intensity determinations can extend only over the neighborhoods of the reflections. The count values (or analog-to-digital-unit values, on analog detectors) for pixels that lie between the Bragg reflections are not necessarily ignored: they are used in estimating backgrounds under neighboring reflections. But the pixels distant from the spots escape the intense scrutiny to which the pixels near the predicted spot positions are subjected. The pre-calculation of the spot centroid positions involves straightforward diffraction geometry for oscillation data (18). For Weissenberg and other non-standard data collection schemes the geometry is more complex, but remains tractable. HKL, for example, provides for Weissenberg data.

For each reflection to be measured, the data processing program must identify the range of pixels, surrounding the centroid, over which the summation or profile analysis will be performed. The simplest approach to this task is to choose a rectangular ;arallelpiped in (X,Y,ω) large enough to include all the pixels immediately surrounding the centroid. Thus if we expect that spots will begin two images before the peak, end two images after the peak, and extend for four pixels to the left, right, above, and below the reflection, then a summing box that is 11x11x7 pixels will be sufficient to "catch" all the pixels required for integration. This approach has at least two disadvantages. First, by including pixels in the corners of the parallelpiped, outside the area we actually expect the reflection to inhabit, we degrade the signal-to-noise ratio of the intensity measurement. Second, we increase the likelihood of incorporating pixels that belong to other reflections in the summation box of the current reflection.Therefore a better approach is to identify the actual perimeter of the reflection by analyzing the three-dimensional or two-dimensional profiles of previously-analyzed reflections and use only a very limited number of pixels outside this perimeter to allow for mismeasurement of the spot centroid. Once the range of pixels has been defined, we can begin to use the data within the range.

Any processing program that performs profile analysis must include provision for defining the two- or three-dimensional model profile that is expected to apply to a given reflection. These profiles could in principle be derived from an *a priori* calculation based on a detailed knowledge of the properties of the crystal, the beam, and the sample goniostat. Such a theoretical approach has been avoided in *XENGEN*, X-GEN, XDS, MADNES, and most other packages that employ profile analysis. Instead, these programs provide for division of the detector into a modest number of (perhaps overlapping) regions, and the centered profiles of a group of bright reflections in each region are added up to produce a model profile appropriate to that region. The model applied to any reflection is either taken as the normalized model for the region in which it resides or a weighted average of that model with those of neighboring regions. In *XENGEN* and X-GEN it has proven helpful with strong reflections to further modify the model by comparing its second moments in X,Y, and ω with those of the observed profile. Large deviations of the model's moments from the corresponding moments of the observed profile suggest that the profile is globally too flat or too peaked in the direction indicated. An exponential sharpening or flattening function is then applied to the model to increase its match to the observed profile without completely obliterating the shape of the model. The way the model is actually used in integration is discussed below.

The background under a given intensity measurement is often a large fraction of the raw count in its component pixels. To obtain accurate intensity measurements it is absolutely crucial that a properly computed and experimentally appropriate background-estimation scheme be employed. Weak reflections, whose intensities play such a crucial role in high-resolution studies, are particularly sensitive to errors in background measurement. X-ray background in the conventional sense and legitimate intensity are only two of the sources of "counts" that appear in a pixel. With analog detectors, particularly SIT-tube and CCD detectors, analog-to-digital increments appear in each pixel's histogramming counter even in the absence of incident X-rays due to electronic or optical events in the phosphor and the components between the phosphor and the counter. Such "dark current" is mostly thermal, and can be reduced and rendered more time-independent by cooling theoptical train, but it will still need to be removed from the observed count values "before" the software begins to treat X-ray background. X-ray background, arising from Compton scattering and elastic but non-Bragg scattering in and around the sample, can be reduced at the detector by careful attention to the environment of the sample, but again it cannot be eliminated. It can also be reduced somewhat by moving the detector farther back: most of the sources of background fall off on a pixel-by-pixel basis as the square of the sample-to-detector distance, whereas the number of pixels that must be summed to measure the intensity grows somewhat less rapidly than the square. But whatever remains in a given experiment must be accurately estimated.

With multiwire detectors the properties of the instrument and the nature of the experiment dictate that the background under any pixel be estimated as an average over other images of the counts at the current pixel position. The images to be averaged should be those in which the pixel does not lie within any reflection's profile. Thus in this scheme

$$B(X,Y,\omega) = \Sigma_{\omega'}\, w_{\omega'}\, C(X,Y,\omega') \,/\, \Sigma_{\omega'}\, w_{\omega'} \qquad (11)$$

where the images ω' being summed are those in which (X,Y) lies outside all reflections and the weight factors $w_{\omega'}$ vary so that the images closest to the current image ω carry higher weight than those farther away. A version of this approach, pioneered by Xuong in San Diego, computes a running average for each pixel:

$$B(X,Y,\omega) = \begin{cases} (1\text{-}z)B(X,Y,\omega\text{-}1) + zC(X,Y,\omega) \\ \\ B(X,Y,\omega\text{-}1) \end{cases} \qquad (12)$$

where the lower choice applies if (X,Y,ω) is contained in a reflection and the upper choice if it is not. In Xuong's original implementation z is a constant (1/16); in XENGEN and X-GEN this approach is one of the background schemes available, and the value of z is an adjustable parameter. The advantage of this technique is that it provides a simple way of defining the background at any stage of the integration, and the background associated with a given reflection is dominated by the ten or so images that precede the current one. Furthermore, unexpected excursions in the count value can be filtered out. If

$$|\, C(X,Y,\omega) - B(X,Y,\omega\text{-}1) \,| > N\sigma(B), \qquad (13)$$

we can choose not to update. The most common circumstance where this arises is the presence of an unpredicted reflection, e.g. a spot arising from a different crystallite. The value of N in this inequality is set to 4 in X-GEN but could be made user-adjustable.

The appropriateness of this "updating" technique derives from the fact that multiwire detector pixels can vary widely in their effective pixel area and the fact that the stepsize between images on these detectors is usually small. Thus the background under a pixel (X,Y,w) may be distinctly different from that under a neighboring pixel, whereas the background at (X,Y) will change only slowly from image to image.

This "updating" approach to background estimation is inappropriate to image-plate data and other film-like data collection schemes. There the stepsize is usually large, so the background under a pixel may change substantially from one image to the next; further, the backgrounds under adjacent pixels within a single image are likely to be very similar, since the effective area of neighboring pixels on image plates (and film) are equal. Thus for these systems a backgrounding scheme that averages over pixels within an image is more appropriate. The technique advocated by Rossmann (16) involves computing a least-squares plane:

$$B(X,Y) = a_0 + a_1X + a_2Y \qquad (14)$$

where a_0 , a_1, and a_2 are estimated by least-squares calculations extended over pixels near to but outside the current reflection in the current image. This algorithm is implemented in X-GEN and is the default for image-plate systems.

There are intermediate cases where it is unclear whether averaging across images within a pixel or averaging within an image across pixels would be better. Charged-coupled devices are a case in point. They often have a substantial degree of geometrical nonlinearity and nonuniformity of response. These properties would argue for averaging across images within a pixel; but will that be the right answer if a broad stepsize between images is used? X-GEN allows the user to select which approach is employed with a given set of data, so there is an opportunity to test this question with real instruments and real data. To my knowledge no such tests have been performed.

With an adequate estimate of the background in hand, the intensity can be calculated in various ways. The simplest would be to sum the background-subtracted counts for reflection j in the pre-selected pixels of the profile:

$$I_j = \Sigma_i\, g_{ij} \qquad (15)$$

where i indexes the pixels (X,Y,ω) and

$$g_{ij} \equiv C_i - B_i = C(X,Y,\omega) - B(X,Y,\omega), \qquad (16)$$

but this takes no advantage of our knowledge of the expected profile shape. In profile-analysis, this simple summation is replaced by the following analysis. If the *normalized* model profile is expressed as $f_i = f(X,Y,\omega)$, then we expect that it will resemble the observed profile apart from an overall scale factor k_j, so that the quantity $k_j * f_i - g_{ij}$ will be small throughout the reflection. Thus

$$k_j * \Sigma_i f_i \approx \Sigma_i g_{ij} = I_j \qquad (17)$$

but since the model f_i is normalized the sum on the left is unity and

$$k_j = I_j, \qquad (18)$$

i.e. the scale factor relating the observed profile to the model is in fact the intensity. Thus if we can compute k by least squares, we will arrive at an estimate of I_j. We do the least-squares calculation in a typical way:

$$\min \Sigma_i [(k_j f_i - g_{ij}) u_{ij}]^2 \qquad (19)$$

with solution

$$I_j = k_j = \Sigma_i (f_i g_{ij}) u_{ij}{}^2 / \Sigma_i (f_i u_{ij})^2 \qquad (20)$$

Note that for uniform weights u_{ij} and a "top-hat" model f_i, i.e. $f_i = 1 / M$ (M = number of pixels), this equation for k_j reduces to

$$k_j = \Sigma_i g_{ij} \qquad (21)$$

so that simple summation becomes a special case of profile-fitting. In X-GEN both the simply-summed and the profile-fitted intensities are computed, and the user is free to choose either for further computations. The summation intensity is rendered more accurate by the recognition that if some portion of the profile g_{ij} is unavailable due to overlap with another spot or occlusion by the beamstop, the intensity estimated from the model profile for those unavailable pixels can be substituted for direct observations. This can be done by replacing eqn. (15) with

$$I_j = [\Sigma_i g_{ij}] / [\Sigma_i f_{ij}] \qquad (22)$$

where the sum extends over all available pixels. Thus if all pixels are available the denominator is unity and eqn. (22) reduces to eqn. (15); if pixels are missing their contributions are implicitly included in (22) because the denominator will be less than one. This in fact is the only use to which profile analysis is put in XDS and MADNES; in X-GEN it is used for the summation intensities, and the full-blown analysis implied by eqn. (20) is provided as an alternative. In X-GEN the weight factors u_{ij} in eqn. (20) are unity; other authors (13) have proposed alternatives.

Note that the profile-analysis approach automatically gives us a way to gradually modify the model profiles fi to accommodate small changes in the models. In this approach, we compute the model profile pixel-by-pixel for a group of reflections j = 1, 2, ... n:

$$\min \Sigma_j [(k_j f_i - g_{ij}) u_{ij})^2 \qquad (23)$$

with solution

$$f_i = \Sigma_j (k_j g_{ij}) u_{ij}{}^2 / \Sigma_i (k_j u_{ij})^2 \qquad (24)$$

As with background updating, the usable values of f_i are computed by weighted averages of this "new" value of f_i and the previous one.

A variety of multiplicative corrections must be applied to the intensities measured either by profile fitting or summation. The most obvious are the Lorentz correction, which measures the velocity of the Bragg spot through the finite thickness of the Ewald sphere, and the polarization correction, which is a physical property of diffraction geometry. An additional correction that can be applied is a correction for absorption of the diffracted beam by the medium between the sample and the detector; if this medium is helium or vacuum the correction is probably unnecessary, but if the medium is air the difference in path length between spots measured near the detector's normal to the sample and the those measured far from that normal may be substantial. Thus the actual intensity is computed not from eqns. (20) or (22) but rather from

$$I_j = (C_j/p_j)\Sigma_i (1/L_{ij}) f_i g_{ij} u_{ij}{}^2 / \Sigma_i (f_i u_{ij})^2 \qquad (25)$$

or from

$$I_j = (C_j/p_j)[\Sigma_i (1/L_{ij}) g_{ij}] / [\Sigma_i f_{ij}] \qquad (26)$$

where C_j is the path-absorption correction for reflection j, p_j is the polarization correction for reflection j, and $1/L_{ij}$ is the Lorentz correction for pixel i of reflection j. The Lorentz factor can change by as much as 20% across the extent of a Bragg reflection that is close to the rotation axis, so it can be dangerous to assume that the Lorentz factor for the entire reflection can be treated as equivalent.

For simple diffraction geometries and conventional X-ray sources these corrections are straightforward. The path absorption is simply

$$C_j = \exp(-\alpha x_j) \qquad (27)$$

where α is the absorption coefficient of the medium (~0.01 cm^{-1} for air with 8KeV X-rays) and x is the distance from the sample to the point on the detector where the spot is visualized. For simple rotation geometries (with the direct beam along z and the rotation axis along x) the Lorentz factor is also simple:

$$L_{ij} = |s_{ijy}| \qquad (28)$$

where s_{ijy} is the y component of the diffraction vector of the spot, measured at pixel i of reflection j. Precession, Weissenberg, and other geometries produce more complicated Lorentz formulas. For conventional X-ray sources without single-crystal monochromators the input X-rays are unpolarized, so the only polarization to correct for is that produced by the sample itself:

$$p_j = (1 + \cos^2 2\theta_j) / 2 \qquad (29)$$

With a single-crystal monochromator the X-rays become polarized by the monochromator crystal itself, so the polarization takes on other forms. These are described in detail by Azaroff (20). For synchrotrons the X-rays emanating from the source are already polarized before any beamline optics come into play, so the polarization correction becomes a complex formula involving the initial beam's properties, thetype and geometry of monochromator, and the sample's own effects.

Macromolecular crystallographic experiments extend over fairly long time intervals (minutes to weeks) and wide ranges of scanning angles, even within a single data run. The parameters that characterize both the sample and the experimental arrangement may change over the course of the run. Thus the unit cell parameters may change by a few tenths of a percent over a data run; the orientation angles may change by a degree or more; the model profiles for the spots may change, either due to time or because different projections of the lattice produce differently-shaped profiles; and the background measurements may vary distinctly as a function of scanning angle. Also, if the software does not explicitly correct for all forms of geometrical misalignment in the experimental system, the parameters that can compensate for these misalignments may assume values that provide for good matches between predicted and observed reflection locations over a narrow range of scanning angle, but the compensations fail at scanning angles far removed from the starting point. For all these reasons it is useful to update the operating parameters of the integration effort as it proceeds.

Some of these updates have already been discussed. Updating the background estimates is straightforward. The backgrounding method embodied in eqn. (12) has updating built into it, whereas the Rossmann-style background algorithm of eqn. (14) retains no memory of previous images. Updating the model profiles using eqn. (24) has also been described.

Updating the unit cell, orientation, and detector parameters is a bit more complex. X-GEN and other packages have facilities for refining these parameters based on differences between observed and predicted spot positions in (X,Y,ω) as data processing proceeds, and they are generally successful in tracking slow changes in sample orientation or unit cell parameters. In X-GEN these on-the-fly refinements are accomplished with a Simplex algorithm that minimizes a residual that includes weighted contributions from errors in (X,Y), errors in ω, and non-integerness of the spot indices of the observed centroids:

$$\min w_x E(x) + w_y E(y) + w_\omega E(\omega) + w_h E(h) \quad (30)$$

with

$$w_x + w_y + w_\omega + w_h = 1, \quad (31)$$

$$E(x) = \Sigma (X_{jo} - X_{jc})^2, \quad (32)$$

$$E(y) = \Sigma (Y_{jo} - Y_{jc})^2, \quad (33)$$

$$E(\omega) = \Sigma (\omega_{jo} - \omega_{jc})^2, \quad (34)$$

$$E(h) = \Sigma (h_{jc} - [h_{jc}])^2 + (k_{jc} - [k_{jc}])^2 + (l_{jc} - [l_{jc}])^2, \quad (35)$$

where X_{jo}, etc. are the observed values for observation j and X_{jc}, etc. are the computed values. The user can choose which parameters are to be refined in this scheme. In principle only the orientational parameters and perhaps the unit cell lengths angles should be allowed to vary, since the detector parameters should not shift during the run, but in practice due to unmodeled instrumental misadjustments (see above) it is often useful to allow the detector's translational offsets to vary during refinement as well. The refinements serve the calculational purpose of ensuring that the calculated centroids track the observed centroids, and they also provide a diagnostic role: if non-plausible shifts arise (e.g. large changes in the detector's translational offsets, or 2% changes in unit cell), it is generally a sign that something is wrong with the data and requires further attention.

A large, discontinuous shift (e.g., a sudden two-degree rotation of the sample about the rotation axis) will result in a complete mismatch between the predicted pattern and the observed pattern, so the refinement would never get any bright reflections to work from; in this case the on-the-fly refinement will fail. One could envision implementing an on-the-fly auto-indexing capability that would take over in these extreme cases, but to my knowledge no current package does this. Discontinuous changes currently require manual intervention.

3.4. Merging and Scaling the Integrated Data

Even having made multiplicative corrections to the raw diffraction intensities, we find a certain amount of massaging is necessary to actually use the intensity measurements derived from the integration step. The user is likely to need a single intensity estimate for *all* symmetry-related observations of a given reflection, rather than separate measurements for each observation; so observations must be grouped and formatted internally so that appropriate means can be computed. Corrections for decay and sample absorption generally are applied after integration, and faulty observations are deleted from the data.

Grouping is a straightforward operation, provided that the symmetry of the crystal is known before it begins. In that case one can simply reformat and sort the integrated data so that symmetry-related observations appear together. In most packages, including X-GEN, Friedel mates retain their identity, so that anomalous analyses can be readily carried out. It is also useful to retain enough information about the original reflection that the raw reflection indices and the diffraction-vector properties of the reflection can be extracted if necessary.

Corrections for systematic error in macromolecular crystallography are rarely based on morphological analyses of the sample. Methods that rely on differential intensity measurements as the crystal is rotated around the diffraction vector of a reference reflection (21) can be employed with single-counter diffractometers, but they are inconvenient with area detectors and are rarely used. Consequently most software packages rely on the high degree of redundancy obtained in area-detector data collection to generate a systematic-error model, and this redundancy-based model is used to correct the intensity measurements.

References

(1) R.A. Sparks, “Data Collection with Diffractometers” and E.J. Gabe and Y. Le Page, “Raw Data to Integrated Data Set”, both in: *Computational Crystallography* (D.Sayre, ed.) Clarendon Press, Oxford, 1982, pp. 1-18, 41-55..

(2) H.W. Wyckoff, “Diffractometry” and R.J. Fletterich and J. Sygusch, “Measuring X-Ray Diffraction Data from Large Proteins with X-Ray Diffractometry”, both in: *Methods in Enzymology*, Vol. **114** (H.W. Wyckoff, C.H.W. Hirs, and S.N. Timasheff, eds.) Academic Press, Orlando, 1985, pp. 330-396.

(3) S.C. Harrison, F.K. Winkler, C.E. Schutt, and R. Durbin, “Oscillation Method with Large Unit Cells” and M.G. Rossmann, “Determining the Intensity of Bragg Reflections from Bragg Photographs”, both in: *Methods in Enzymology*, Vol. **114**, op. cit., pp. 211-280.

(4) Ng.-h. Xuong and S.T. Freer, *Acta Crystallogr.* **B**27: 2380-2387, 1971.

(5) U.W. Arndt, J.N. Champness, R.P. Phizackerley, and A.J. Wonacott, *J. Appl. Crystallogr.* **6**: 457, 1973.

(6) R. Hamlin, C. Cork, A. Howard, C. Nielsen, W. Vernon, D. Matthews, Ng.-h. Xuong, and V. Perez-Mendez, “Characteristics of a Flat Multiwire Detector for Protein Crystallography” *J. Appl. Crystallogr.* **14**: 85-93, 1991.

(7) S.E. Sobottka, R.J. Chandross,G.G. Cornick, R.H. Kretsinger, and R.G. Rains, “Design and Performance of the Multiwire Area X-ray Diffractometer at the University of Virginia” *J. Appl. Crystallogr.* **23:** 199-208, 1990.

(8) C. Bolon, J. Crawford, M. Deutsch, and G. Quigley, *IEEE Trans. Nucl. Sci.* **NS-28**, No. 1, 1981.

(9) R. Kahn, R. Fourme, B. Caudron, R. Bosshard, R. Benoit, R. Bouclier, G. Charpak, J. C. Santiard, and F. Sauli, *Nucl. Instrum. Methods* **172:** 337, 1980.

(10) U.W. Arndt, “Television Area Detector Diffractometers” in: *Methods in Enzymology* ,Vol **114**, *op cit.*, pp. 472-485.

(11) K. Kalata, “A General Purpose, Computer-Configurable Television Area Detector for X-Ray Diffraction Applications” in: *Methods in Enzymology* ,Vol **114**, *op cit.*, pp. 486-510.

(12) A. Messerschmidt and J.W. Pflugrath, “Crystal Orientation and X-ray Pattern Prediction Routines for Area-Detector Diffractometer Systems in Macromolecular Crystallography”, *J. Appl.Crystallogr.* **20:** 306, 1987.

(13) G. Bricogne, ed. *Proceedings of the EEC Cooperative Workshop on Position-Sensitive Detector Software, Phases I & II and Phase III.* CNRS/LURE, Orsay, 1986.

(14) A.J. Howard, G.L. Gilliland, B.C. Finzel, T.L. Poulos, D.H. Ohlendorf, and F.R. Salemme, “Use of an Imaging Proportional Counter in Macromolecular Crystallography” *J. Appl. Crystallogr.* **20:** 383-387, 1987.

(15) W. Kabsch, “Evaluation of Single-Crystal X-ray Diffraction Data from a Position-Sensitive Detector”, *J. Appl. Crystallogr.* **21:**916-924, 1988.

(16) M.G. Rossmann, “Processing Oscillation Diffraction Data for Very Large Unit Cells with an Automatic Convolution Technique and Profile Fitting” J. Appl. :Crystallogr. 12: 225-238, 1979.

(17) Z. Otwinowski, publication data unavailable.

(18) Ng.-h. Xuong, S.T. Freer, R. Hamlin, C. Nielsen, and W. Vernon, “The Electronic Stationary Picture Method for High-Speed Measurement of Reflection Intensities from Crystals with Large Unit Cells” *Acta Crystallogr.* **A34:** 289-296, 1978.

(19) D.J. Thomas, “Modern Equations of Diffractometry: Goniometry” *Acta Crystallogr.*

A46: 321-343, 1990, and D.J. Thomas, “Modern Equations of Diffractometry: Diffraction Geometry” *Acta Crystallogr.* **A48:** 134-158, 1992.

(20) L.V. Azaroff, *Acta Crystallogr.* **8:** 701-704, 1955.

(21) A.C.T. North, D.C. Phillips, and F.S. Mathews, *Acta Crystallogr.* **A24:** 351, 1968.