[20] Processing of X-Ray Diffraction Data Collected in Oscillation Mode

By ZBYSZEK OTWINOWSKI and WLADEK MINOR

Introduction

X-ray data can be collected with 0-, 1-, and 2-dimensional detectors, 0-D (single counter) being the simplest and 2-D the most efficient in terms of measuring diffracted X-rays in all directions. Two-dimensional detectors have been used since 1912 for X-ray diffraction studies. Initially the 2-D detector was made of X-ray sensitive photographic film; now electronic detectors and phospholuminescent films (best known by the trade name IP or Imaging Plate) dominate. To analyze single-crystal diffraction data collected with these detectors, several computer programs have been developed. The 2-D detectors and related software are now used predominantly to measure and integrate diffraction from single crystals of biological macromolecules. However, the usefulness of these systems in small-molecule, high-resolution crystallography is just being recognized and much of the rest of this discussion is applicable to that field also.

Among the computer programs that were used widely during the past 15 years are MOSFLM and related programs,[1,2] XDS,[3,4] OSC[5–8] and its derivative WEIS,[9] BUDDHA,[10] FILME,[11] Denzo,[12] MADNES,[13] the San

[1] A. Leslie, "Data Collection and Processing," Proceedings of the CCP4 Study Weekend, 29–30 January 1993, compiled by L Sawyer, N. Isaac, S. Bailey, pp. 44–51.

[2] A. G. W. Leslie, *in* "Proceedings of the Daresbury Study Weekend at Daresbury Laboratory," 23–24 January 1987, compiled by J. R. Helliwell, P. A. Machin, and M. Z. Papiz, pp. 39–50.

[3] W. Kabsch, *J. Appl. Cryst.* **21,** 67–81 (1988).

[4] W. Kabsch, *J. Appl. Cryst.* **21,** 916–924 (1988).

[5] M. G. Rossmann, *in* "Biological Systems: Structure and Analysis," Proceedings of the Study Weekend at Daresbury Laboratory, 24–25 March, 1984 (G. P. Diakun and C. D. Garner, eds.), pp. 28–40. Science and Engineering Research Council, Daresbury.

[6] M. G. Rossmann, A. G. W. Leslie, S. S. Abdel-Meguid, and T. Tsukihara. *J. Appl. Cryst.* **12,** 570–581 (1979).

[7] M. G. Rossmann, *J. Appl. Cryst.* **12,** 225–238 (1979).

[8] M. G. Rossmann and J. W. Erickson, *J. Appl. Cryst.* **16,** 629–636 (1983).

[9] N. Sakabe, *Nucl. Instrum. Methods* **A303,** 448–463 (1991).

[10] M. Blum, S. C. Harrison, and D. C. Wiley, *J. Appl. Cryst.* **20,** 235–242 (1987).

[11] P. Schwager, K. Bartels, and A. Jones, *J. Appl. Cryst.* **8,** 275–280 (1975).

[12] Z. Otwinowski, *in* "Data Collection and Processing," Proceedings of the CCP4 Study Weekend, 29–30 January 1993, compiled by L. Sawyer, N. Isaac, and S. Bailey, pp. 56–62.

[13] A. Messerschmidt and J. W. Pflugrath, *J. Appl. Crystallogr.* **20,** 306–315 (1987).

Copyright © 1997 by Academic Press, Inc.
All rights of reproduction in any form reserved.

Diego programs,[14] and related programs XENGEN[15] and X-GEN. The theory behind the data-reduction methods is complex enough that a series of European Economic Community workshops were dedicated to this task only.[16,17] The proceedings from these workshops contain the best, although most voluminous, presentation of the theory.

The four most important developments in the data analysis of macromolecular diffraction measurements are autoindexing,[3,13,18–21] profile fitting,[22,23] transformation of data to a reciprocal-space coordinate system,[4,19] and demonstration[7] that a single oscillation image contains all of the information necessary to derive the diffraction intensities from that image. The analysis and reduction of single-crystal diffraction data consists of seven major steps. These include (1) visualization and preliminary analysis of the original, unprocessed detector data, (2) indexing of the diffraction pattern, (3) refinement of the crystal and detector parameters, (4) integration of the diffraction maxima, (5) finding the relative scale factors between measurements, (6) precise refinement of crystal parameters using the whole data set, and (7) merging and statistical analysis of the measurements related by space-group symmetry.

We have developed three programs: Denzo and Scalepack to integrate and scale the data, and Xdisplayf to analyze the process visually. Together, these programs form the HKL and the MAC-Denzo packages. Steps 1 through 4 are carried out by the programs Denzo and Xdisplayf, while steps 5 through 7 are performed by the companion program, Scalepack.[23a]

The programs can estimate Bragg intensities from single-crystal diffraction data that are recorded on position-sensitive X-ray (also potentially neutron-diffraction or electron-diffraction) detectors, for example film, IP scanners, or charge-coupled device (CCD) area detectors. The programs allow for data collection by oscillation, Weissenberg, and precession meth-

[14] A. Howard, C. Nielsen, and Ng. H. Xuong, *Methods Enzymol.* **114,** 452–472 (1985).

[15] A. J. Howard, G. L. Gilliland, B. C. Finzel, T. L. Poulos, D. H. Ohlendorf, and F. R. Salemme, *J. Appl. Cryst.* **20,** 383–387 (1987).

[16] G. Bricogne, *in* "Proceedings of the Daresbury Study Weekend at Daresbury Laboratory," 23–24 January 1987, compiled by J. R. Helliwell, P. A. Machin, and M. Z. Papiz, pp. 120–146 (1987).

[17] EEC Cooperative Workshop on Position-Sensitive Detector Software, (Phase I and II) LURE, Paris, 16 May–7 June 1986, (Phase III) LURE, Paris, 12–19 November 1986.

[18] G. Vriend, and M. G. Rossmann, *J. Appl. Cryst.* **20,** 338–343 (1987).

[19] W. Kabsch, *in* "Data Collection and Processing," Proceedings of the CCP4 Study Weekend, 29–30 January 1993, compiled by L. Sawyer, N. Isaac, and S. Bailey, pp. 63–70.

[20] S. Kim, *J. Appl. Cryst.* **22,** 53–60 (1989).

[21] T. J. Higashi, *J. Appl. Cryst.* **23,** 253–257 (1990).

[22] R. Diamond, *Acta Cryst.* **A25,** 43–55 (1974).

[23] G. Ford, *J. Appl. Cryst.* **7,** 555–564 (1974).

[23a] D. Gewirth, "The HKL Manual—A Description of Programs Denzo, XDisplayf and Scalepack," 4th Ed. Yale University, New Haven, CT, 1994.

ods. The detector can be either flat or cylindrical. The detector readout can be either rectilinear or spiral, although spiral coordinates must be converted to rectilinear before processing. The programs allow for random changes in the position and the sensitivity of the detector between consecutive exposures. The programs Denzo, Xdisplayf, and Scalepack implement most of the ideas discussed at the EEC Cooperative Programming Workshop on Position-Sensitive Detector Software.[16,17] In particular, the programs feature profile fitting, weighted refinement, eigenvalue filtering, and universal definition of detector geometry.

Visualization of Diffraction Space

A diffraction data set forms an image of three-dimensional (3-D) reciprocal space. This 3-D image consists of a series of two-dimensional (2-D) diffraction images, each of them representing a different, curved slice of reciprocal space. In order for the diffraction maxima to be accurately integrated they must appear as separated (nonoverlapping) spots in the individual 2-D images. Unless the data are collected by the precession method, the diffracted image contains a distorted view of reciprocal space. This distortion of the image is a function of the data-collection method, the diffraction geometry, and the characteristics of the detector. For the data reduction to be successful, the distortion of reciprocal space as viewed by the detector has to be accounted for correctly by the program. The distortion of the image of reciprocal space can vary even between images collected on the same detector. This is because the position of the detector, the X-ray wavelength, the oscillation range, pixel size, scanner gain, and the exposure level all affect the detector representation of diffraction space.

One should start data collection and reduction with a careful inspection of the data in their raw (original) form.[24] The zoom option of the program Xdisplayf allows one to examine reflections in pixel-by-pixel detail to check that the diffraction maxima are resolved. Because the program displays the resolution (in angstroms) corresponding to the position of the mouse-driven cursor, the diffraction limit of the crystals can be estimated even without data reduction. The display in high-zoom mode provides digital pixel values, so one can check, among other things, that the exposure level is appropriate.

If problems exist with the detector or other components of the data collection system, the display option helps to discover these before all the data are recorded. The examination of the image may reveal if there are extraneous sources of X-ray background. There are other statistics that can

[24] W. Minor, American Crystallographic Association Abstracts, p. 31.

be provided instantly by Xdisplayf which may indicate, for example, A/D (analog/digital) converter malfunction. If there are many diffraction maxima in the image that form a characteristic pattern of diffraction from a single crystal, then the next step is deducing a crystal lattice that accounts for such a pattern. This step is called indexing.

Indexing

Autoindexing

The HKL package offers two indexing methods: automatic and interactive. The automatic method, applicable in most cases, is fast and simple. The first step in the automatic method is the peak search, which chooses the spots to be used by the autoindexing subroutine. Ideally, the peaks should come from diffraction by a single crystal. The Denzo program accepts peaks for autoindexing only from a single oscillation image. It is important that the oscillation range be small enough (it can even be zero, i.e., a still) so that the lunes (rings of spots, all from one reciprocal plane) are resolved. One should note that requirement of lunes separation is distinct from requirement of spot separation. If lunes overlap, spots may have more than one index consistent with a particular position on the detector. However, the oscillation range should be large enough to have a sufficient number of spots for the program to be able to establish periodicity of the diffraction pattern. This may require at least 0.5° oscillation for a small unit cell protein crystal and 2–3° oscillation in the case of organic small molecule crystals.

The second step in the autoindexing is the mapping of the diffraction maxima identified by the peak search onto reciprocal space. Because the precise angles at which reflections diffract are *a priori* unknown for oscillation data, the center of the oscillation range is used as the best estimate of the angle at which the diffraction occurs.

The autoindexing in Denzo is based on a novel algorithm: a complete search of all possible indices of all reflections that are found by a peak search or are selected manually. When the program finds values (integer numbers) of one index (for example, **h**) for all reflections, this is equivalent to having found one real-space direction of the crystal axis (in this case, a). For this reason such indexing is called *real-space indexing.* Finding one real-space vector is logically equivalent to finding the periodicity of the reciprocal lattice in the direction of this vector. The search for real-space vectors is performed by a fast Fourier transform (FFT) and takes advantage of the fact that finding all values of one index (e.g., **h**) for all reflections is independent of finding all values of another index (e.g., **k**). The Denzo implementation of this method is not dependent on prior knowledge of

the crystal unit cell; however, for efficiency reasons, the search is restricted to a reasonable range of unit cell lengths, obtained by default from the requirement of spot separation.

After the search for real-space vectors is completed, the program finds the three linearly independent vectors, with minimal determinant (unit cell volume), that would index all (or, more precisely, almost all) of the observed peaks. These three vectors form a basis, but are unlikely to form a standard basis for a description of the unit cell. The process of converting a basis into a standard basis is called *cell reduction.* The program follows the definitions in the International Tables for Crystallography[25,*] and finds the best cells for all of the 14 Bravais lattices.

The transformation of the primitive cell to a higher symmetry cell may require some distortion of the best triclinic lattice that fits the peak-search list. Because of experimental errors, the fit is never perfect for the correct crystal lattice. Sometimes the observed reflections can fit a higher symmetry lattice than one defined by space-group symmetry. Such condition is called lattice (or metric tensor) pseudo symmetry. If this happens, the lattice determination and assignment of lattice symmetry may get complicated. The procedure in such case is to index the data in the lowest symmetry lattice that does not introduce a wrong lattice symmetry (the triclinic lattice is always a safe choice), and look for the symmetry of the intensity pattern during the scaling of symmetry-related reflections. Denzo calculates the distortion index for all 14 of the Bravais lattices. It is up to the user to define the lattice and space-group symmetry, since the program, at this stage of the calculation, cannot distinguish lattice symmetry from pseudo-symmetry.

Reliability

Autoindexing by the HKL programs is very reliable. The authors are not aware of a single failure of autoindexing with known detector geometry and a diffraction image satisfying the assumptions described previously. Autoindexing worked also on a significant fraction of data where these assumptions were violated. In practice, problems in autoindexing (and subsequent refinement) are mostly due to simple experimental mistakes.

The real-space indexing method finds the best assignment of indices to all reflections simultaneously. Therefore, a small percentage of incorrectly identified diffraction maxima does not affect the method. The method is

[25] International Tables for Crystallography, Vol. A, pp. 738–749. Kluwer Academic Publishers, Dordrecht, 1989.

* The definition as implemented in Denzo differs from the practice of some labs when a crystal has either a primitive monoclinic or an orthorhombic lattice.

insensitive to how many short difference vectors can be created from the peak-search list, and this is one of the reasons why it is a more reliable method than the traditional ones[3,13,18–21] based on direct indexing of difference vectors.

Failure of Autoindexing

Autoindexing is based on the assumption that the diffraction spots are correctly mapped from detector coordinates to diffraction (reciprocal) space. The origin of the diffraction space is defined by the position of the direct beam on the detector. A substantial error in the beam position can shift the indexing of the diffraction pattern by an integer vector. Such misindexing can be totally self-consistent until the stage when symmetry-related reflections are compared. For any assumed (starting) value of the beam position, the origin of the diffraction space during indexing will be shifted to the nearest grid point of the best primitive lattice. An initial error in the direct beam position by 0.48 times the distance between reflections will lead to correct indexing, while an error of 0.52 times the same distance will cause a misindexing of the diffraction pattern by one index. Misindexing by one is never corrected by subsequent refinement of the crystal and detector parameters. Misindexing often produces poor agreement between the predicted and the observed positions of the reflections, but for some crystal orientations, the agreement between the predicted and the observed positions can be equally good for both correctly indexed and misindexed cases. This property of the diffraction geometry creates a potential trap for the unwary crystallographer.

Errors in detector orientation will produce distorted mapping from detector to reciprocal space. A wrong specimen-to-detector distance will enlarge (or decrease) the apparent reciprocal-space separation between Bragg peaks. This error will not be uniform in three directions; in the first approximation, along the beam direction, the error will be proportional to the square root of the error in the distance; in the other two directions the error will have linear dependence.

In most cases a significantly wrong crystal-to-detector distance (say an error of 10%) will not make the autoindexing step fail immediately; however, the calculated unit cell will be quite wrong. The length of the unit cell along the beam direction will be 5% shorter, and in the perpendicular directions, 10% shorter. If the crystal is diagonally oriented (no principal axis along the beam direction), then the apparent angles between axes will violate the lattice symmetry. Similarly, incorrectly defined angles between the detector and the beam will result in wrong angles between crystal axes.

Therefore, the interpretation of the lattice symmetry is dependent, to some extent, on how precisely the detector parameters are known *a priori.*

Most failures of autoindexing happen because of incorrect detector parameters input to Denzo. Autoindexing can also fail when more than one crystal contributes to the diffraction image. Sometimes, editing of weaker reflections and resolution cuts can make one crystal dominate the peak-search list enough for the autoindexing method to succeed. If crystals have similar orientation, sometimes using only very low-resolution data can be the right method. In case of twinned crystals, autoindexing sometimes finds a superlattice that finds integer indexes for both crystals. In such a case Denzo solves the problem of finding the best 3-D lattice that goes through all of the observed peaks. Unfortunately, for a twinned crystal this is a mathematically correct solution to a wrong problem.

Sometimes the crystal asymmetric unit may have molecules related by an approximate translation by a fraction (typically one-half) of a unit cell edge or diagonal. The resulting diffraction pattern will have odd-index reflections much weaker than even-index reflections. Autoindexing may find one of the two possible solutions, the choice depending on whether odd reflections are weak enough to be assumed (within experimental error) systematically absent or not. This depends on what is the fraction of odd reflections in the peak search. If there are only a few odd reflections, then most of the peak-search result may be explained with a smaller real-space unit cell. To prevent autoindexing from finding such a smaller cell, one should enhance the fraction of odd reflection used in autoindexing: for example, by changing peak-search criteria, or by using only high-resolution reflections in autoindexing. If one still cannot index odd reflections, then one should consider ignoring them all together. In such a case, one can solve the structure in a smaller unit cell, and the resulting structural error will not be very significant if odd reflections are much weaker than even ones.

Autoindexing in Denzo always finds a standard lattice; the crystallographer may prefer a nonstandard choice, for example, to make the lattice similar to one in a different space group. Reindexing in Scalepack or manual indexing in Denzo accommodates such needs.

Interactive Indexing

Because there is no general algorithm to index a diffraction image from multiple crystals, one has to rely on the ability of the brain to sort out complex patterns as an alternative to autoindexing in such cases. The approximate orientation can be determined by an iterative trial-and-error

process where the predicted pattern is being adjusted, keeping the diffraction image constant.

The crystal orientation can be defined relative to any principal or higher-order zone perpendicular to the X-ray beam. This flexibility helps the interactive indexing when only a higher-order zone is visible in the diffraction pattern. This is particularly useful in centered space groups where it may be easier to orient a diagonal zone, rather than a major one. Manipulation of the predicted diffraction patterns also can be used to simulate diffraction experiments.

The simulation can help set a proper data collection strategy in order to avoid later problems in data reduction. Using the program for simulation of diffraction patterns can also be a tool for teaching crystallography.

Refinement of Crystal and Detector Parameters

The integration of reflections requires knowledge of their index and position. The weak reflections can be found only by prediction based on the information obtained from strong reflections. The autoindexing step provides only the approximate orientation of the crystal, and the result may be imprecise if the initial values of the detector parameters are poorly known. The least-squares refinement process is used to improve the prediction. The parameters describing the measurement process either have to be known *a priori* or have to be estimated from diffraction data by a manual or automatic refinement procedure. Depending on the particulars of the experiment, the same parameters (e.g., crystal to detector distance) are more precisely known *a priori*, or are better estimated from the data. Denzo allows for the choice of fixing or refining each of the parameters separately. This flexibility is handy under special circumstances; using it well requires considerable knowledge of diffraction experiments. Fortunately, the "fit all" option and detector-specific default values seem to be reliable under most conditions.

The crystal and detector orientation parameters require refinement for each processed image. The refinement can be simple, for a series of images collected with an on-line detector, or more complex, if the detector orientation is only crudely known and varies from image to image, as in the case of off-line scanners. The refinement is controlled by the user and can consist of several steps. In each step the user defines the resolution limits and the order and number of parameters to be fitted. Both detector and crystal parameters can be fitted simultaneously by the fast-converging least squares method. The refinement is done separately for each image to allow for the processing of data even when the crystal (or the detector) slips considerably during data collection.

Occasionally the refinement can be unstable because of a high correlation among some parameters. High correlation makes it possible for the errors in one parameter to compensate partially for the errors in other parameters. If the compensation is 100%, the parameter would be undefined, but the error compensation by other parameters would make the predicted *pattern* correct. In such cases eigenvalue filtering (the same method as Singular Value Decomposition, described in Ref. 26) is employed to remove the most correlated components from the refinement and make it numerically stable. Eigenvalue filtering works reliably when starting parameters are close to correct values, but may fail to correct large errors in the input parameters if the correlation is close to, but not exactly, 100%. Once the whole data set is integrated, the global refinement (sometimes called postrefinement)[6,29,30] can refine crystal parameters (unit cell and orientation) more precisely and without correlation with detector parameters. The unit cell used in further calculations should come from the global refinement (in Scalepack) and not from Denzo refinement.

The detector and crystal parameters are refined by a least squares method that minimizes the deviation of the reflection centroids from their predicted positions. Such refinement by itself is seriously deficient when applied to a single oscillation image, since one crystal rotation parameter is undefined (rotation about the spindle does not change the *position* of spots on the detector) and the others are highly correlated and/or poorly defined. To overcome this problem, another term (partiality refinement) is added within Denzo, in which the intensity of the partially recorded reflections is compared to the predicted partiality multiplied by an average intensity in the same resolution range. The formula for the residual (difference between expected and predicted value) is the same as in the postrefinement, however at this stage the error of the predicted fully recorded intensity is much larger, equal to the expected intensity. Nonetheless, the concomitant positional and partiality refinement used in Denzo is both stable and very accurate. The power of this method is in proper weighting (by estimated error) of two very different terms—one describing positional

[26] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, "Numerical Recipies, The Art of Scientific Computing," Cambridge University Press, Cambridge, 1989.

[27] Deleted in proof.

[28] Deleted in proof.

[29] P. R. Evans, *in* "Proceedings of the Daresbury Study Weekend at Daresbury Laboratory," 23–24 January 1987, compiled by J. R. Helliwell, P. A. Machin, and M. Z. Papiz, pp. 58–66 (1987).

[30] P. Evans, "Proceedings of the CCP4 Study Weekend," 29–30 January 1993, compiled by L. Sawyer, N. Isaac, and S. Bailey, pp. 114–123 (1993).

differences and the other describing intensity differences. The method leads to a reduced correlation between the detector and crystal parameters. An additional benefit is the uniform treatment of both detector and crystal variables in the whole refinement process.

A correct understanding of the detector geometry is essential to accurate positional refinement. Unfortunately, most detectors deviate from perfect flat or cylindrical geometry. These deviations are detector specific. The primary sources of error include misalignment of the detector position sensors (MAR, R-AXIS), nonplanarity of the film or IP during exposure or scanning, inaccuracy of the wire placement and distortions of the position readout in multiwire proportional counters (MWPCs), and optical distortion (which can also be due to a magnetic field acting upon the image intensifier) in the TV or CCD-based detectors. If the detector distortion can be parameterized, then these parameters should be added to the refinement. For example, in the case of the spiral scanners there are two parameters describing the end position of the scanning head. In the perfectly adjusted scanner these parameters would be zero. In practice, however, they may deviate from zero by as much as 1 mm. Such misalignment parameters can correlate very strongly with other detector and crystal parameters. If the program does not have the ability to describe detector distortions, then the other parameters such as the unit cell and crystal-to-detector distance will be systematically wrong.

With film and IPs handled manually in cassettes, as at many synchrotrons, the biggest problem lies in keeping the detector flat during exposure and subsequent scanning. In the manual systems, it is much harder to model the possible departures from ideal flat or cylindrical geometry, and Denzo, like most programs, makes limited attempts to correct such distortions. Nonideal film or IP geometry is one of the main factors behind the variable quality of data collected with the manual systems.

Integration of Diffraction Maxima

Profile Fitting

The accurate prediction of spot positions is necessary to achieve a precise integration of Bragg peaks. The most important need for accurate prediction of the spot positions arises from the application of profile fitting. Profile fitting is a two-step process. First, the profile is predicted based on the profiles of the other reflections within a chosen radius. The predicted profile in Denzo is an average of profiles shifted by the predicted separation between the spots, so that they are put on top of each other. If the predicted

positions are in error, then the average profile will be broadened and/or displaced from the actual profile of the reflection.

In the second step, the information from the actual and the predicted profile is combined by the following process:

The observed profile M_i is a sum of the Bragg peak and background. The estimate of M_i, given by P_i, is expressed by the formula

$$P_i = B_i + \text{constant}\,(p_i) \tag{1}$$

where B_i is the predicted value of the background and p_i is the predicted profile. Profile fitting minimizes the function:

$$\sum_i \frac{(M_i - P_i)^2}{V_i} \tag{2}$$

with respect to the constant, where V_i is the variance (σ^2) of M_i. V_i is a function of the expected signal in a pixel, which in the case of a counting detector is P_i. The index i represents all pixels in a two-dimensional profile, however the same formulation of profile fitting applies to one- and three-dimensional profiles. The predicted profile can be normalized arbitrarily; the most natural definition of normalization is that the sum of p_i is equal to 1. Such a choice makes the constant in Eq. (1) the fitted intensity I, i.e., I is a constant.

The solution to the profile fitting can be expressed by an alternative, but mathematically equivalent, approach, presented as follows.

Each pixel provides an estimate of the spot intensity I equal to $(M_i - B_i)/p_i$ with variance $V_i/p_i{}^2$. A profile-fitted intensity is then simply a weighted average of all observations:

$$I = \frac{\sum \frac{p_i^2}{V_i} \frac{(M_i - B_i)}{p_i}}{\sum \frac{p_i^2}{V_i}} = \frac{\sum \frac{p_i(M_i - B_i)}{V_i}}{\sum \frac{p_i^2}{V_i}} \tag{3}$$

This approach [without an explicit solution presented in Eq. (1)] was first published by Diamond in 1969 for the one-dimensional case. However, in 1974, Ford proposed a simplified formula where V_i is constant. This was based on the mistaken idea that the variance of the optical density value of the exposed film is independent of the degree of X-ray exposure. Equation (3) thus became simpler:

$$I = \frac{\Sigma p_i(M_i - B_i)}{\Sigma p_i^2} \tag{4}$$

Many of the subsequent programs followed the formulation of Ford rather than that of Diamond, even when applied to data collected with

proportional counters or IPs. The unweighted formula proposed by Ford works quite well where the peak spot intensity is not much higher than the background intensity. This situation arises more often with data collected on film, which has a high intrinsic background and low saturation, or when the crystals have low scattering power due to a very large unit cell, high solvent content, or disorder. The unweighted profile fitting improves the accuracy of the weak reflections compared to a straight summation, but at the cost of reducing the accuracy of the strong ones. This observation did lead in the past to a partial solution based on taking a weighted average between profile-fitted and summed intensities, where the weight is a function of the reflection intensity. The weighted formula [Eq. (3)] used in Denzo does not deteriorate the accuracy of strong, low-resolution reflections. Thus, the observed problem with the unweighted formula is in the lack of weighting.

Errors of Profile Fitting

The profile fitting increases the precision (decreases the statistical error) of the measurement, but it may introduce an error due to lack of accuracy of the predicted profiles. Denzo applies the averaging of profiles in detector coordinates and, unlike other programs that use profile fitting method, averages profiles separately for each spot. This approach has two main advantages: first, only nearby spots are chosen for averaging, ones that should have the most similar profiles. Second, Denzo avoids interpolation in the profile prediction step; instead it shifts the contributing profiles by vectors that make the smallest possible pixel-truncation error. These translation vectors precisely center the predicted profile on the reflection to be fitted, and the error introduced by these shifts is smaller than that due to interpolation used in some other programs.

The prediction of profile shape is never exact, because errors in the positional refinement, averaging of different shapes, truncation of pixel shifts or interpolation, etc. The resulting error of the fitted intensity was analyzed by Diamond[22] in the case of one-dimensional Gaussian profiles and an unweighted profile-fitting formula. The important parameters are w, the root mean square (rms) width of the actual profile; f, the rms width of the predicted profile; and d, the displacement of the predicted profile from the actual profile. Define the relative change in the square of the reflection width as

$$D^2 = \frac{f^2 - w^2}{w^2} \tag{5}$$

Diamond calculated that the fitted intensity will be wrong by a factor of

$$\left[1 + \frac{D^2}{2 + D^2}\right]^{1/2}$$

The averaging of profiles adds a value of $r^{2/3}$, where r is the detector pixel size, to the value of f^2. Averaging will increase the profile-fitted intensity of most reflections by a constant multiplicative factor, which has little effect on crystallographic procedures. The interpolation broadens the profile by a factor dependent on the position of the predicted reflection relative to the pixel boundaries. The interpolation will also increase f^2 by a number between zero and $r^{2/3}$. The interpolation method will increase the profile-fitted intensities on average by the same factor, but will also add random noise to the reduced data that is not present in the Denzo method.

Other Aspects of Spot Integration

There are other, often subtle, ways in which errors in spot positions can lead to serious integration errors. In many experiments the detector is placed as close as possible to the crystal while keeping the diffraction spots separated. In such cases the reflections are barely separated, and even small errors in the spot prediction would make integration and background measurement areas of a reflection intrude upon the adjacent peaks, and thus lead to an inaccurate estimation of the peaks' intensities.

Errors in the prediction of spot positions also affect the statistical error (precision) of the summed intensities. If the predictions do not match the peak position exactly, one has to enlarge the expected spot area in order to sum the intensity of the whole spot. This enlargement of the predicted spot area increases the total background to be subtracted. A larger background has a larger variance, and this adds to the measurement variance. Autocentering of the spot area can compensate for errors in the prediction, but this works well only for strong spots. It would seriously bias the calculated intensity if applied individually to every spot. Some programs do autocentering by averaging the local deviations between the observed and the predicted positions. While this is not done explicitly in Denzo, the profile prediction algorithm used in the program has a similar effect.

To calculate the diffraction intensity, the background under the Bragg peak has to be estimated and then subtracted from the reflection profile. The standard method used to estimate the background value is to calculate an average detector signal in the neighborhood of a specific reflection. In Denzo it is assumed that the background is a linear function of the detector coordinates. Robust statistics (as discussed in Ref. 26) is applied to remove the contribution of pixels that deviate more than 3 sigma from the best fit

to the background function. If too many background pixels are flagged as outliers from background function, the whole reflection is removed from the integration. Denzo ignores pixels in three other cases: when they have been flagged as no measurement by an auxiliary program, when they have a special value (e.g., zero in the case of *R* axis or MAR), or when they are in the spot area (based on the predicted, rather than the measured position) of an adjacent reflection.

A correction for the nonlinear response function of the detector to the photon flux is applied internally in Denzo so that it can read the original data without the need for any prior transformations, with the exception of the data from spiral scanners. Pixel values can represent two special cases: *no measurement* or *detector overload.* Overloaded pixels are assumed to be close to the center of gravity of the diffraction spots, and as such they are used in determining the spot centroids. Pixels that are either overloaded or had no measurement are ignored in calculating the spot intensity by the profile-fitting method, but the existence of such pixels in the spot area is flagged by a negative sign applied to the sigma estimate. Profile-fitted intensities seem to be reliable independent of the existence of such pixels in the spot area.

Scaling and Merging

The scaling and merging of different data sets and global refinement of crystal parameters (postrefinement) is performed by the program Scalepack. The scaling algorithm is one described by Fox and Holmes.[30a] Scalepack differs in the definition of the estimated error of measurement. In Scalepack, unlike in other procedures, the estimated error is enlarged by a fraction of expected, rather than observed, intensity. The Scalepack method reduces the bias existing in other programs toward reflections with integrated intensity below the average.

Global Refinement: Postrefinement

Owing to correlation between crystal and detector parameters, the values of unit cell parameters refined from a single image may be quite imprecise. This lack of precision is of little significance to the process of integration, as long as the predicted positions are on target. There is no contradiction here, because at some crystal/detector orientations the positions of reflections may depend only weakly on a value of a particular crystal parameter. At the end of the data-reduction process one would wish to get precise unit cell values. This is done in the procedure referred to as a global refinement or postrefinement.[6,29,30] The implementation of this

[30a] G. C. Fox and K. C. Holmes, *Acta Cryst.* **20,** 886–891 (1966).

method in the program Scalepack allows for separate refinement of the orientation of each image, but with the same unit-cell parameter values being used for the whole data set. In each batch of data (typically one image) a different unit-cell parameter may be poorly determined. However, in a typical data set there are enough orientations to determine precisely all unit cell lengths and angles. The global refinement is also more precise than the processing of a single image in determination of the crystal mosaicity and orientation.

Experimental Feedback

Every element of the data collection process must function close to its optimum in order for one to solve a macromolecular structure. The sheer amount of data collected makes computer programs an inevitable intermediary between the researcher and the experiment. The HKL package provides several levels of insight into the data at each stage of the measurement and data analysis process: Scalepack, which provides statistics for the full data set; Denzo, which provides numerical analysis of one oscillation image; and Xdisplayf, which presents data visually, up to a single-pixel level.

Different problems manifest themselves most clearly at different levels of analysis. The traditional method of judging the success of the experiment by the final statistics (e.g., from Scalepack) is not sufficient, since it does not show if the experiment was done optimally. The biggest problem with final statistics is that they do not differentiate well between the sources of the problems, and often come too late to fix them. Therefore, the experimenter must be aware of how the detector, the X-ray beam, the crystal, and the procedure all contribute to the final data quality and how each of them can make the experiment a failure.

Detector

Detector problems are best diagnosed by collecting data with benchmark, high-quality crystals (e.g., tetragonal lysozyme). There is no particular advantage of lysozyme crystals, with the possible exception of how easy it is to grow them, and a larger unit cell crystal would be preferable (e.g., tetragonal chymotrypsin).

One should expect very high data quality from test crystals. The resulting anomalous difference Fourier map should identify all the sulfur atoms in the protein. The detector-parameters refinement should produce a very small spread (tens of microns, hundredths of a degree) from one image to another. Such a test may require the mounting of a test crystal in a way that avoids slippage and minimizes absorption. R-merge statistics in the

range 2–3%, based on high redundancy (four-fold or higher) and high resolution (2 to A or better) should be expected. Only very few (less than 0.1%) outliers should be found during merging.

Results worse than the above indicate a problem with the test crystal or with the experimental setup. Preferably, the test crystal should be kept at 100 K to avoid radiation damage. Problems with the test crystal may mask detector problems. For instance, slippage of the test crystal makes it very difficult to notice a spindle motor backlash or malfunctioning of the X-ray shutter.

Many macromolecular crystallography labs have not developed stringent benchmarks of acceptable performance. The most frequent problem with such lack of rigor is the acceptance of many outliers in the test data. Outlier rejection in merging of symmetry-related data is a valid statistical procedure, but it should be applied with great caution. The definition of an outlier is "a large but sporadic fluctuation in the data," for example, due to a cosmic-radiation hit. A small number, less than 0.01%, of outliers is something to be expected, even in a well-functioning system. However, the practice of many labs has been to accept a much larger number of outliers, even as high as 10%. Many serious problems may be masked out by such a liberal outlier rejection. It should be emphasized that outlier rejection always improves consistency (including consistency indexes, e.g., *R*-merge), but not necessarily the correctness of the merged data.

It is dangerous to accept results from a test with a significant number of reflections flagged as outliers, even if the *R*-merge statistics seem to be good. This is almost a sure sign of a sporadic problem, and unless the problem is well understood, it may not be sporadic when one collects data to solve a crystal structure. One way to attempt to understand the nature of outliers is to locate them in the dector space in order to identify the problem. The clustering of outliers in one area of the detector may indicate a damaged surface; if most outliers are partials, it may indicate a problem with spindle backlash or shutter control. The zoom mode may be used to display the area around the outliers to identify the source of a problem: for example, the existence of a satellite crystal, or single pixel spikes due to electronic failure. Sometimes a histogram of the pixel intensities may suddenly stop below maximum valid pixel value, indicating a saturation of the data acquisition hardware/software.

X-Ray Beam

The main properties of the X-ray beam that need to be checked are stability, focus, angular spread, and wavelength in the case of MAD experi-

ments.[31,32] Large fluctuations in beam intensity show in a variable background intensity and variable scale factors during scaling. The quality of the beam focus is immediately visible in the spot profile of low-resolution reflections. Angular spread of the beam contributes to reflection width, and it may introduce overlaps between reflections for the crystals of very long unit cells. The beam properties (except stability) are best analyzed by the inspection of images. The beam parameters are less significant for crystals with large mosaicity.

Experimental Procedure

In the traditional approach, one collects data first and then starts analyzing the result. This strategy has a risk that there may be a gross inefficiency in the setup of the experiment: the data set may be incomplete, the reflections may overlap, the zones may overlap, a large percentage of the reflections may be overloaded, etc. At that stage the only solution is to repeat the experiment, which may be difficult with unique crystals or with experiments that require a synchrotron source.

Data collection is best performed as a highly interactive process. Immediate data processing, which the authors encourage, provides fast feedback during data collection. Most macromolecular crystallographic projects go through iterative stages of improving crystal quality and data-collection strategy. Typically, most of the data collection time and effort is spent before the optimal point is reached. Then, if data collection is going well, there is a pressure to use the expensive detector and X-ray beam resource efficiently. The three basic questions are whether to collect, what to collect, and how to collect.

The first question is if the data are worth collecting. Quick scaling of a partial data set collected in the first minutes may eliminate the need to collect a full set of nonderivative data. Observing many diffraction spots in an image encourages one to collect a full data set, however a high number of spots may be due to high mosaicity, making such a data set unprocessable. One image is enough to index it, estimate mosaicity, and notice how severe is the problem with overlaps between the reflections. If the Bragg peaks are not resolved, there is no point in collecting such data, however many spots one sees in the image.

The second question is what range of data to collect. Typically, one wants to collect up to the resolution limit. The resolution limit is defined

[31] J. L. Smith, *Curr. Opin. Struct. Biol.* **1,** 1002–1011 (1991).

[32] R. Fourme and W. A. Hendrickson, *in* "Synchrotron Radiation and Biophysics" (S. S. Hasnain, ed.), pp. 156–175. Ellis Horwood Limited, Chichester, 1990.

by the ratio of average intensity to sigma (noise) being about 2. The safest way to establish it is by the processing of a test image, rather than by guessing. One has to note that some space groups have inherent ambiguities in indexing, which only scaling of the initial image to the previously collected data can resolve. Otherwise, one risks recollecting already-measured reflections rather than filling in the missing data.

The third question is how to collect data.[32a] The detector should be placed as far as possible from the specimen consistent with the desired resolution limit. Long unit cells, large mosaicity, or large oscillation range all affect spot separation and potential overlaps. Some overlaps are immediately visible—the ones arising from a long unit cell axis in the plane of the detector. At high resolution, because of weakness of the spots, the overlaps may be less obvious. The simulation of a diffraction pattern, based on indexing of the first image and a proposed data-collection protocol, is the right tool to define the sufficiently short oscillation range and the correct detector placement. There is no particular need to collect fully recorded reflections, so the optimal oscillation range is typically narrow, even equal to a fraction of crystal mosaicity.

Data Reduction

One must be continually vigilant during all stages of data reduction to assure that the process is going well or to detect and diagnose problems. Useful statistics are produced at each stage; Xdisplayf allows one to visualize the data instantly in their original form, and it can be also set up to view the progress of data reduction. The displaying of raw data makes it possible for one to grasp the significance of complex patterns that would be hard to analyze numerically. This allows for a quick assessment of problems in the collected data.

There are two classes of data-reduction problems, one that results in location of reflection masks not corresponding to the positions of the Bragg peaks, and another in which the problems do not displace the predicted positions of the reflections. Misprediction is visibly obvious and is disastrous; it may be due to forcing a wrong space group symmetry, misindexing, or serious detector malfunction. Sometimes data scale poorly and produce many outliers, however the predicted positions agree perfectly with the peaks and no detector or diffraction artifacts are visible. This problem may be a simple mistake in data processing, like using a wrong file format, or a nonuniform exposure during crystal oscillation. The nonuniform exposure may be caused by spindle motor backlash, shutter malfunction (opening

[32a] Z. Danter, *Methods Enzymol.* **276,** [21], 1997 (this volume).

too early or too late), ionization chamber electronics failure (if used), decay or variation of the X-ray beam intensity (if ionization chamber is not used), variable speed of the spindle motor, etc. Nonuniform exposure is best diagnosed by exclusion of other problems that may affect data quality. Graphical feedback is providing confidence that the problem cannot be at the indexing/integration stage.

Large variations in absorption of X-rays by the crystal will make data scale poorly and will produce visible variation of the background, however it will not affect positional agreement. The variation in the absorption can be avoided easily by a proper mounting of the crystal. The correction for absorption is a whole field in itself.[33–38]

Summary

Macromolecular crystallography is an iterative process. Rarely do the first crystals provide all the necessary data to solve the biological problem being studied. Each step benefits from experience learned in previous steps. To monitor the progress, the HKL package provides two tools: (i) Statistics, both weighted (χ^2) and unweighted (R-merge), are provided. The Bayesian reasoning and multicomponent error model facilitates the obtaining of proper error estimates[28,39]; and (ii) visualization of the process plays a double role: it helps the operator to confirm that the process of data reduction, including the resulting statistics, is correct, and it allows one to evaluate problems for which there are no good statistical criteria. Visualization also provides confidence that the point of diminishing returns in data collection and reduction has been reached. At that point the effort should be directed to solving the structure.

The methods presented here have been applied to solve a large variety of problems, from inorganic molecules with 5 Å unit cell to rotavirus of 700 Å diameter crystallized in 700 × 1000 × 1400 Å cell.[40] Overall quality

[33] D. Stuart and N. Walker, *Acta Cryst.* **A35,** 925–933 (1979).

[34] N. Walker and D. Stuart, *Acta Cryst.* **A35,** 158–166 (1983).

[35] C. E. Schutt and P. R. Evans, **A41,** 568–570 (1985).

[36] C. Katayama, *Acta Cryst.* **A42,** 19–23 (1986).

[37] D. Stuart, *in* "Proceedings of the Daresbury Study Weekend at Daresbury Laboratory," 23–24 January 1987, compiled by J. R. Helliwell, P. A. Machin, and M. Z. Papiz, pp. 25–38 (1987).

[38] F. J. Takusagawa, *J. Appl. Cryst.* **20,** 243–245 (1987).

[39] D. Schwarzenbach, S. C. Abrahams, H. D. Flack, W. Gonschorek, T. Hahn, K. Huml, R. E. Marsh, E. Prince, B. E. Robertson, J. S. Rollet, and A. J. L. Wilson, *Acta Cryst.* **A45,** 63–75 (1989).

[40] B. Temple and S. C. Harrison, Personal communication.

of the method has been tested by many researchers by successful application of the programs to MAD structure determinations.

Acknowledgments

We would like to thank Dan Gewirth, Halina Czarnocka, and Bob Sweet for help with editing, Paul Sigler for encouragement in this project, and Michael Rossmann for providing the initial stimulus.

[21] Data Collection Strategy

By Z. DAUTER

The best way to proceed during X-ray diffraction data collection depends on qualitative factors, such as crystal quality and availability, type of X-ray source and detector, and time available, and quantitative ones, such as cell parameters, resolution limit, and crystal symmetry. There are certain rules to help one in producing a data set that is complete and accurate, and extends to as high resolution as possible. Often it is impossible to satisfy all these requirements simultaneously, and in most cases the actual data set collected is the result of a compromise. It is worth remembering that all subsequent steps of crystal structure analysis depend on the quality of data collected in the first instance; phasing, Fourier map interpretation, and refinement will proceed more smoothly if the data are good. To help define the parameters to use to set up data collection, in this chapter we discuss determination of the outer resolution limit, a precise description of the behavior of the reciprocal lattice during rotation photography, the effect of crystal mosaicity, the myth of the blind region, and finding ways in which crystal symmetry can help.

The least well-defined criterion in data collection is perhaps the resolution limit of diffraction. In principle, as long as the ratio of average intensity to the associated estimated error is higher than 1.0, the data contain some information. However, there may be only a few reflections having meaningful intensity among many reflections weaker than their associated errors. Therefore, extending the resolution limit may effectively introduce more noise than signal to the system, whether it is the Fourier transform or a least-squares matrix. A useful rule is to restrict the resolution to the point below which more than half of the intensities are higher than 2σ. This assumes that the errors of the measured intensities are estimated correctly. In most programs used for intensity integration from 2-D detector images,

Copyright © 1997 by Academic Press, Inc.
All rights of reproduction in any form reserved.