International Tables for Crystallography (2006). Vol. F, Section 11.4.2, pp. 226–227.

11.4. *DENZO* and *SCALEPACK*

BY Z. OTWINOWSKI AND W. MINOR

11.4.1. Introduction

X-ray diffraction data analysis, performed by the *HKL* package (Otwinowski, 1993; Otwinowski & Minor, 1997) or similar programs (Rossmann, 1979; Howard *et al.*, 1985; Blum *et al.*, 1987; Bricogne, 1987; Howard *et al.*, 1987; Leslie, 1987; Messerschmidt & Pflugrath, 1987; Kabsch, 1988; Higashi, 1990; Sakabe, 1991), is used to obtain the following results:

(1) estimates of structure factors and determination of the crystal symmetry;
(2) estimates of the crystal unit-cell parameters;
(3) error estimates of the structure factors and unit cell;
(4) detector calibration; and
(5) detection of hardware malfunctions.

Other results, like indexing of the diffraction pattern, are in most cases only intermediate steps to achieve the above goals. The *HKL* system and other programs also have tools to validate the results by self-consistency checks.

The fundamental stages of data analysis are:
(1) visual inspection of the diffraction images;
(2) (auto)indexing;
(3) diffraction geometry refinement;
(4) integration of the diffraction peaks;
(5) conversion of the data to a common scale;
(6) symmetry determination and merging of symmetry-related reflections; and
(7) statistical summary and estimation of errors.

This order represents the natural flow of data reduction, but quite often these steps are repeated based on information obtained at a later stage.

The three basic questions in collecting diffraction data are:
(1) *whether to collect*;
(2) *what to collect*; and
(3) *how to collect and analyse the data.*

These questions and steps (1)–(7) of data analysis are intimately intertwined.

Data analysis makes specific assumptions which the collected data must, or at least should, satisfy. However, the experimenter can verify whether the data satisfy those assumptions only by data analysis. This circular logic can be broken by an iterative process. On-line data analysis provides immediate feedback during data collection and can remove the guesswork about *whether*, *what* and *how* from the process. The description of data analysis and algorithms that follows will make frequent references to the assumptions about the data and offer guidelines on how to make the experiment fulfil these assumptions.

This article uses the *HKL* package coordinate system to describe data algorithms and analysis. However, as most equations are written in vector notation, they can be easily adapted to conventions used in other programs.

11.4.2. Diffraction from a perfect crystal lattice

X-ray photons can scatter from individual electrons by inelastic and incoherent processes. The coherent scattering by the whole crystal is called diffraction.* Energy conservation, when expressed in photon momentum vectors, is equivalent to

$$\mathbf{S}\cdot\mathbf{S}_0 = \tfrac{1}{2}\mathbf{S}\cdot\mathbf{S}, \tag{11.4.2.1}$$

where $\mathbf{S}$ is the diffraction vector, defined as the change of photon momentum in the scattering process, and $\mathbf{S}_0$ is the vector which has beam direction and length $1/\lambda$. Diffraction from a perfect crystal lattice occurs when diffraction from all repeating crystal elements is in phase, which can be stated in vector algebra as

$$\mathbf{S}\cdot\mathbf{a} = h \tag{11.4.2.2}$$

$$\mathbf{S}\cdot\mathbf{b} = k \tag{11.4.2.3}$$

$$\mathbf{S}\cdot\mathbf{c} = l. \tag{11.4.2.4}$$

In shorter notation, these may be written as $\mathbf{h} = [A]\mathbf{S}$, which is equivalent to $\mathbf{S} = [A]^{-1}\mathbf{h}$, where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are the real-space crystal periodicity vectors,

$$[A] = \begin{pmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{pmatrix}$$

and h, k, l are the integer Miller indices. Often, the orientation matrix is defined in reciprocal space as the inverse of $[A]$.

The condition for crystal diffraction with Miller indices h, k, l is the existence of a (unique) vector $\mathbf{S}$ which is a solution to equations (11.4.2.1)–(11.4.2.4). Equation (11.4.2.1) states the diffraction condition for vector $\mathbf{S}$. Mathematically speaking, the space of the solutions to equations (11.4.2.2)–(11.4.2.4) is called reciprocal space, and vector $\mathbf{S}$ belongs to this space. However, the following presentation does not depend on the properties of reciprocal space. The laboratory coordinate system used has its origin at the position of the crystal. A diffraction peak at the detector position in three-dimensional laboratory space $\mathbf{X} = \{x, y, z\}$ corresponds to vector $\mathbf{S}$:

$$\mathbf{S} = \mathbf{X}/\lambda|\mathbf{X}| + \mathbf{S}_0. \tag{11.4.2.5}$$

Rotation of the crystal around the goniostat axes can be described by vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in equations (11.4.2.2)–(11.4.2.4) as a function of the goniostat angles, and vectors $\mathbf{a}_0, \mathbf{b}_0, \mathbf{c}_0$ represent the crystal orientation at the zero position of the goniostat. These rotations are described by Bricogne (1987):

$$\mathbf{a} = [R_1(\varphi_1)][R_2(\varphi_2)][R_3(\varphi_3)]\mathbf{a}_0, \tag{11.4.2.6}$$

where

$$[R(\varphi)] = \cos(\varphi)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + [1-\cos(\varphi)]\begin{pmatrix} e_xe_x & e_xe_y & e_xe_z \\ e_xe_y & e_ye_y & e_ye_z \\ e_xe_z & e_ye_z & e_ze_z \end{pmatrix} + \sin(\varphi)\begin{pmatrix} 0 & -e_z & e_y \\ e_z & 0 & -e_x \\ -e_y & e_x & 0 \end{pmatrix}, \tag{11.4.2.7}$$

where e_x, e_y, e_z represent the direction cosines of a rotation axis. To complete the description of diffraction geometry, we need a function $X(p, q)$, describing the position in experimental space of each pixel with integer coordinates $\{p, q\}$. This function is detector-specific and describes the detector geometry and distortion. For a planar detector,

$$\mathbf{X}(p,q) = [R_x][R_y][R_z]([R_{2\theta}]([L](K[D(p,q)] - \mathbf{B}) + \mathbf{T}_D) - \mathbf{T}_D) + \mathbf{T}_D, \tag{11.4.2.8}$$

where R_x, R_y, R_z represent the detector misorientation, $R_{2\theta}$ represents rotation around the 2θ (swing) axis, $\mathbf{T}_D$ is the detector

* Owing to the large difference in mass between the crystal and the photon, the energy of the photon is virtually unchanged.

translation from the crystal, operation L represents the axis naming/direction convention used by the detector manufacturer (eight possibilities), K is an operation scaling pixels to millimetres, D is a detector distortion function and $\mathbf{B}$ represents the beam position on the detector surface.

Equations (11.4.2.1)–(11.4.2.8) fully describe the existence and position of the diffraction peaks, which is all that is needed for the autoindexing procedure.

11.4.3. Autoindexing

Among the number of autoindexing algorithms proposed (Vriend & Rossmann, 1987; Kabsch, 1988; Kim, 1989; Higashi, 1990; Leslie, 1993), the method based on periodicity of the reciprocal lattice tends to be the most reliable (Otwinowski & Minor, 1997; Steller *et al.*, 1997).

Autoindexing starts with a peak search, which results in the set of $\{p, q, i\}$ triplets, where i is the number of the image in which the peak with position $\{p, q\}$ was found. The program takes advantage of the fact that for any rotation matrix

$$([R]\mathbf{S}) \cdot ([R]\mathbf{a}) = \mathbf{S} \cdot \mathbf{a}. \qquad (11.4.3.1)$$

When

$$[R] = [R_3(-\varphi_3)][R_2(-\varphi_2)][R_1(-\varphi_1)], \qquad (11.4.3.2)$$

equation (11.4.3.1) applied to equation (11.4.2.2) becomes:

$$([R]\mathbf{S}) \cdot \mathbf{a}_o = h, \qquad (11.4.3.3)$$

where $\mathbf{a}_o$ is a three-dimensional vector with as yet unknown components. Note that the matrix $[R]$ represents crystal rotation when the crystal is in the diffraction condition defined by the existence of the solution to equations (11.4.2.1)–(11.4.2.4), described by vector $\mathbf{S}$. For data collected in the wide oscillation mode* the angle at which diffraction occurs is not known *a priori*; however, it can be approximated by the middle of the oscillation range of the image. Combining the peak position $\{p, q\}$ with equations (11.4.2.5) and (11.4.2.8) provides an estimate of the vector $\mathbf{S}$. So, we expect that equation (11.4.3.3) and similar equations for k and l are approximately (owing to approximation and experimental errors) satisfied. The purpose of autoindexing is to determine the unknown vectors $\mathbf{a}_0, \mathbf{b}_0, \mathbf{c}_0$ and the $\{h, k, l\}$ triplet for each peak. To accomplish this, three equations (11.4.3.3) for each peak must be solved. *DENZO* introduced a method based on the observation that the maxima of the function

$$\sum_i \cos[2\pi([R]\mathbf{S}_i) \cdot \mathbf{a}_0] \qquad (11.4.3.4)$$

are the approximate solutions to this set of equations (11.4.3.3). To speed up the search for all significant maxima, a two-step process is used. The first step is the search for maxima of function (11.4.3.3) on a three-dimensional uniform grid, made very fast owing to the use of a fast Fourier transform (FFT) to evaluate function (11.4.3.4). Function (11.4.3.4) is identical to structure-factor calculations in the space group $P1$, which allows the use of the crystallographic FFT. Because the maxima at the grid points (*HKL* uses a $96 \times 96 \times 96$ grid) only approximate the maxima of function (11.4.3.4), the vectors resulting from a grid search are optimized by the Newton method. Function (11.4.3.4) has maxima not only for basic periodic vectors $\mathbf{a}_0, \mathbf{b}_0$ and $\mathbf{c}_0$, but also for any integer linear combination of them. Any set of three such vectors with a minimal nonzero determinant can be used to describe the crystal lattice. Steller *et al.* (1997) describe the algorithm that finds the most reliable set of three vectors. This set needs to be converted to the one conventionally used by crystallographers, as defined in *IT* A (1995).

To generate the conventional solution, two steps are used. Step 1 finds the reduced primitive triclinic cell. *IT* A provides the algorithm for this step. Subsequently, step 2 finds conventional cells in Bravais lattices of higher symmetry.

11.4.3.1. *Lattice symmetry*

The relationship between a higher-symmetry cell and the reduced primitive triclinic cell can be described by

$$[A] = [M][P], \qquad (11.4.3.5)$$

where $[A]$ and $[P]$ are matrices of the type $\{a_0, b_0, c_0\}$, with $[P]$ representing the reduced triclinic primitive cell, and $[M]$ is one of the 44 matrices listed in *IT* A.† If $[A]$ is generated using equation (11.4.3.5) from an experimentally determined $[P]$, owing to experimental errors it will not exactly satisfy the symmetry restraints. *DENZO* introduced a novel index that helps evaluate the significance of this violation of symmetry. This index is based on the observation that from $[A]$ one can deduce the value of the unit cell, apply symmetry restraints to the unit cell and calculate any matrix $[A']$ for the unit cell that satisfies these symmetry restraints. If $[A]$ satisfies symmetry restraints, the matrix $[U]$, where

$$[U] = [A][A']^{-1}, \qquad (11.4.3.6)$$

will be unitary and

$$[U]^T - [U]^{-1} = 0. \qquad (11.4.3.7)$$

The index of distortion printed by *DENZO* is

$$\left\{\sum_i \sum_j \left([U]_{ij}^T - [U]_{ij}^{-1}\right)^2\right\}^{1/2} \Big/ 6, \qquad (11.4.3.8)$$

where i and j are indices of the 3×3 matrix $[U]$.

The value of this index increases as additional symmetry restraints are imposed, starting from zero for a triclinic cell. Autoindexing in *DENZO* always finishes with a table of distortion indices for 14 possible Bravais lattices, but does not automatically make any lattice choice.

11.4.3.2. *Lattice pseudosymmetry*

The cell-reduction procedure cannot determine lattice symmetry, since it cannot distinguish true lattice symmetry from a lattice accidentally having higher symmetry within experimental error (*e.g.* a monoclinic lattice with $\beta \simeq 90°$ is approximately orthorhombic). If one is not certain about the lattice symmetry, the safe choice is to assume space group $P1$, with a primitive triclinic lattice for the crystal, and to check the table again after the refinement of diffraction-geometry parameters. A reliable symmetry analysis can be done only by comparing intensities of symmetry-related reflections, which is done later in *SCALEPACK* or another scaling program.

11.4.3.3. *Data-collection requirements*

The total oscillation range has to cover a sufficient number of spots to establish periodicity of the diffraction pattern in three dimensions. It is important that the oscillation range of each image is small enough so that the lunes (rings of spots, all from one reciprocal plane) are resolved. One should note that the requirement

* This is when the crystal oscillation angle during the measurement of a single diffraction pattern is larger than the angular reflection width.

† It should be noted that No. 17 contains an error (Kabsch, 1993).

references