Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Integration, scaling, space-group assignment and post-refinement

Wolfgang Kabsch

Max-Planck-Institut für Medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, 69120 Heidelberg, Germany

Correspondence e-mail: wolfgang.kabsch@mpimf-heidelberg.mpg.de

Received 19 August 2009
Accepted 9 November 2009

A version of this paper will be published as a chapter in the new edition of Volume *F* of *International Tables for Crystallography*.

Important steps in the processing of rotation data are described that are common to most software packages. These programs differ in the details and in the methods implemented to carry out the tasks. Here, the working principles underlying the data-reduction package *XDS* are explained, including the new features of automatic determination of spot size and reflecting range, recognition and assignment of crystal symmetry and a highly efficient algorithm for the determination of correction/scaling factors.

1. Introduction

The key steps in the processing of diffraction data from single crystals involve (i) modelling of the observed reflection positions in the detector plane, (ii) integration of diffraction intensities, (iii) data correction, scaling and post-refinement and (iv) space-group assignment. Much of the theory and many of the methods for carrying out these steps were developed about three decades ago for processing rotation data recorded on film and were subsequently extended in order to fully exploit the capabilities of a variety of electronic area detectors; some CCD (charge-coupled device) and multiwire detectors as well as a new pixel detector specially developed for data collection at synchrotron beamlines allow the recording of finely sliced rotation data because of their fast data read-out. In this article, the principles of the methods are described as employed by the program *XDS* (Kabsch, 2010). These apply equally well to rotation images covering small or large oscillation ranges. A large number of other data-processing systems have been developed which differ in the details of the implementations. Some of these packages were described in chapter 25.2 of Volume *F* of *International Tables for Crystallography* (2001). The theory and practice of processing fine-sliced data have been discussed by Pflugrath (1997).

2. Modelling rotation images

The observed diffraction pattern, *i.e.* the positions of the reflections recorded on the rotation-data images, is controlled by a small set of parameters which must be accurately determined before integration can start. Approximate values for some of these parameters are given by the experimental setup, whereas others may be completely unknown and must be obtained from the rotation images. This is achieved by the automatic location of strong diffraction spots, the extraction of a primitive lattice basis that yields integer indices for the observed reflections and the subsequent refinement of all

parameters to minimize the discrepancies between observed and calculated spot positions in the data images.

2.1. Coordinate systems and parameters

In the rotation method, the incident-beam wavevector $\mathbf{S}_0$ of length $1/\lambda$ (where λ is the wavelength) is fixed while the crystal is rotated around a fixed axis described by a unit vector $\mathbf{m}_2$. $\mathbf{S}_0$ points from the X-ray source towards the crystal. It is assumed that the incident beam and the rotation axis intersect at one point at which the crystal must be located. This point is defined as the origin of a right-handed orthonormal laboratory coordinate system $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3\}$. This fixed but otherwise arbitrary system is used as a reference frame to specify the setup of the diffraction experiment.

Diffraction data are assumed to be recorded on a fixed planar detector. A right-handed orthonormal detector coordinate system $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$ is defined such that a point with coordinates X, Y in the detector plane is represented by the vector $(X - X_0)\mathbf{d}_1 + (Y - Y_0)\mathbf{d}_2 + F\mathbf{d}_3$ with respect to the laboratory coordinate system. The origin X_0, Y_0 of the detector plane is found at a distance $|F|$ from the crystal position. It is assumed that the diffraction data are recorded on adjacent non-overlapping rotation images, each covering a constant oscillation range Δ_φ, with image No. 1 starting at spindle angle φ_0.

Diffraction geometry is conveniently expressed with respect to a right-handed orthonormal goniostat system $\{\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3\}$. It is constructed from the rotation axis and the incident-beam direction such that $\mathbf{m}_1 = (\mathbf{m}_2 \times \mathbf{S}_0)/|\mathbf{m}_2 \times \mathbf{S}_0|$ and $\mathbf{m}_3 = \mathbf{m}_1 \times \mathbf{m}_2$. The origin of the goniostat system is defined to coincide with the origin of the laboratory system.

Finally, a right-handed crystal coordinate system $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ and its reciprocal basis $\{\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*\}$ are defined to represent the unrotated crystal, *i.e.* at rotation angle $\varphi = 0^\circ$, such that any reciprocal-lattice vector can be expressed as $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$, where h, k, l are integers.

As shown in §2.2, the location of all diffraction peaks recorded in the data images can be computed from the parameters $\mathbf{S}_0$, $\mathbf{m}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, X_0, Y_0, F, $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, φ_0 and Δ_φ. In addition, knowledge of the shape and extent of the diffraction spots is required for accurate estimations of their intensities. This can be achieved by a Gaussian model involving two parameters: the standard deviations of the reflecting range, σ_M, and of the beam divergence, σ_D (see §2.3). This leads to an integration region around the spot defined by the parameters δ_M and δ_D, which are typically chosen to be 6–10 times larger than σ_M and σ_D, respectively.

2.2. Spot prediction

Let $\mathbf{p}_0^*$ denote any arbitrary reciprocal-lattice vector if the crystal has not been rotated, *i.e.* at rotation angle $\varphi = 0^\circ$. Depending on the diffraction geometry, $\mathbf{p}_0^*$ may be rotated into a position fulfilling the reflecting condition. The required rotation angle φ and the coordinates X, Y of the diffracted beam at its intersection with the detector plane can be found from $\mathbf{p}_0^*$ as follows.

$\mathbf{p}_0^*$ can be expressed by its components with respect to the orthonormal goniostat system as

$$\mathbf{p}_0^* = \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}_0^*) + \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*) + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}_0^*).$$

Rotation by φ around axis $\mathbf{m}_2$ changes $\mathbf{p}_0^*$ into $\mathbf{p}^*$,

$$\begin{aligned}\mathbf{p}^* &= D(\mathbf{m}_2, \varphi)\mathbf{p}_0^* \\ &= \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*) + [\mathbf{p}_0^* - \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*)]\cos\varphi + \mathbf{m}_2 \times \mathbf{p}_0^* \sin\varphi \\ &= \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}_0^* \cos\varphi + \mathbf{m}_3 \cdot \mathbf{p}_0^* \sin\varphi) + \mathbf{m}_2\mathbf{m}_2 \cdot \mathbf{p}_0^* \\ &\quad + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}_0^* \cos\varphi - \mathbf{m}_1 \cdot \mathbf{p}_0^* \sin\varphi) \\ &= \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}^*) + \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}^*) + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}^*).\end{aligned}$$

The incident-beam and diffracted-beam wavevectors, $\mathbf{S}_0$ and $\mathbf{S}$, have their termini on the Ewald sphere and satisfy the Laue equations

$$\mathbf{S} = \mathbf{S}_0 + \mathbf{p}^*, \quad \mathbf{S}^2 = \mathbf{S}_0^2 \implies \mathbf{p}^{*2} = -2\mathbf{S}_0 \cdot \mathbf{p}^* = \mathbf{p}_0^{*2}.$$

If $\rho = [\mathbf{p}_0^{*2} - (\mathbf{p}_0^* \cdot \mathbf{m}_2)^2]^{1/2}$ denotes the distance of $\mathbf{p}_0^*$ from the rotation axis, solutions for $\mathbf{p}^*$ and φ can be obtained in terms of $\mathbf{p}_0^*$ as

$$\begin{aligned}\mathbf{p}^* \cdot \mathbf{m}_3 &= [-\mathbf{p}_0^{*2}/2 - (\mathbf{p}_0^* \cdot \mathbf{m}_2)(\mathbf{S}_0 \cdot \mathbf{m}_2)]/\mathbf{S}_0 \cdot \mathbf{m}_3 \\ \mathbf{p}^* \cdot \mathbf{m}_2 &= \mathbf{p}_0^* \cdot \mathbf{m}_2 \\ \mathbf{p}^* \cdot \mathbf{m}_1 &= \pm[\rho^2 - (\mathbf{p}^* \cdot \mathbf{m}_3)^2]^{1/2} \\ \cos\varphi &= [(\mathbf{p}^* \cdot \mathbf{m}_1)(\mathbf{p}_0^* \cdot \mathbf{m}_1) + (\mathbf{p}^* \cdot \mathbf{m}_3)(\mathbf{p}_0^* \cdot \mathbf{m}_3)]/\rho^2 \\ \sin\varphi &= [(\mathbf{p}^* \cdot \mathbf{m}_1)(\mathbf{p}_0^* \cdot \mathbf{m}_3) - (\mathbf{p}^* \cdot \mathbf{m}_3)(\mathbf{p}_0^* \cdot \mathbf{m}_1)]/\rho^2.\end{aligned}$$

In general there are two solutions according to the sign of $\mathbf{p}^* \cdot \mathbf{m}_1$. If $\rho^2 < (\mathbf{p}^* \cdot \mathbf{m}_3)^2$ or $\mathbf{p}_0^{*2} > 4\mathbf{S}_0^2$ the Laue equations have no solution and the reciprocal-lattice point $\mathbf{p}_0^*$ is in the 'blind' region.

If $F\mathbf{S} \cdot \mathbf{d}_3 > 0$ the diffracted beam intersects the detector plane at the point

$$\begin{aligned}F\mathbf{S}/\mathbf{S} \cdot \mathbf{d}_3 &= (F\mathbf{S} \cdot \mathbf{d}_1/\mathbf{S} \cdot \mathbf{d}_3)\mathbf{d}_1 + (F\mathbf{S} \cdot \mathbf{d}_2/\mathbf{S} \cdot \mathbf{d}_3)\mathbf{d}_2 + F\mathbf{d}_3 \\ &= (X - X_0)\mathbf{d}_1 + (Y - Y_0)\mathbf{d}_2 + F\mathbf{d}_3,\end{aligned}$$

which leads to a diffraction spot recorded at detector coordinates

$$\begin{aligned}X &= X_0 + F\mathbf{S} \cdot \mathbf{d}_1/\mathbf{S} \cdot \mathbf{d}_3, \\ Y &= Y_0 + F\mathbf{S} \cdot \mathbf{d}_2/\mathbf{S} \cdot \mathbf{d}_3.\end{aligned}$$

2.3. Standard spot shape

A reciprocal-lattice point crosses the Ewald sphere by the shortest route only if the crystal happens to be rotated about an axis perpendicular to both the diffracted-beam and incident-beam wavevectors, the 'β axis' $\mathbf{e}_1 = \mathbf{S} \times \mathbf{S}_0/|\mathbf{S} \times \mathbf{S}_0|$, as introduced by Schutt & Winkler (1977). Rotation around the fixed axis $\mathbf{m}_2$, as enforced by the rotation camera, thus leads to an increase in the length of the shortest path by the factor $1/|\mathbf{e}_1 \cdot \mathbf{m}_2|$. This motivated the introduction of a coordinate system $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, specific for each reflection, which has its origin on the surface of the Ewald sphere at the terminus of the diffracted beam wavevector $\mathbf{S}$,

$$\mathbf{e}_1 = \mathbf{S} \times \mathbf{S}_0/|\mathbf{S} \times \mathbf{S}_0|,$$
$$\mathbf{e}_2 = \mathbf{S} \times \mathbf{e}_1/|\mathbf{S} \times \mathbf{e}_1|,$$
$$\mathbf{e}_3 = (\mathbf{S} + \mathbf{S}_0)/|\mathbf{S} + \mathbf{S}_0|.$$

The unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are tangential to the Ewald sphere, while $\mathbf{e}_3$ is perpendicular to $\mathbf{e}_1$ and $\mathbf{p}^* = \mathbf{S} - \mathbf{S}_0$. The shape of a reflection, as represented with respect to $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, then no longer contains geometrical distortions resulting from the fixed rotation axis of the camera and the oblique incidence of the diffracted beam on a flat detector. Instead, all reflections appear as if they had followed the shortest path through the Ewald sphere and had been recorded on the surface of the sphere.

A detector pixel at X', Y' in the neighbourhood of the reflection centre X, Y, when the crystal is rotated by φ' instead of φ, is mapped to the profile coordinates ε_1, ε_2, ε_3 by the following procedure:

$$\mathbf{S}' = \frac{(X' - X_0)\mathbf{d}_1 + (Y' - Y_0)\mathbf{d}_2 + F\mathbf{d}_3}{\lambda \cdot [(X' - X_0)^2 + (Y' - Y_0)^2 + F^2)]^{1/2}}$$
$$\varepsilon_1 = \mathbf{e}_1 \cdot (\mathbf{S}' - \mathbf{S})180/(|\mathbf{S}|\pi)$$
$$\varepsilon_2 = \mathbf{e}_2 \cdot (\mathbf{S}' - \mathbf{S})180/(|\mathbf{S}|\pi)$$
$$\varepsilon_3 = \mathbf{e}_3 \cdot [D(\mathbf{m}_2, \varphi' - \varphi)\mathbf{p}^* - \mathbf{p}^*]180/(|\mathbf{p}^*|\pi) \simeq \zeta \cdot (\varphi' - \varphi)$$
$$\zeta = \mathbf{m}_2 \cdot \mathbf{e}_1.$$

ζ corrects for the increased path length of the reflection through the Ewald sphere and is closely related to the reciprocal Lorentz correction factor

$$L^{-1} = \frac{|\mathbf{m}_2 \cdot (\mathbf{S} \times \mathbf{S}_0)|}{(|\mathbf{S}| \cdot |\mathbf{S}_0|)} = |\zeta \cdot \sin \measuredangle(\mathbf{S}, \mathbf{S}_0)|.$$

Because of crystal mosaicity and beam divergence, the intensity of a reflection is smeared around the diffraction maximum. The fraction of total reflection intensity found in the volume element $\mathrm{d}\varepsilon_1\mathrm{d}\varepsilon_2\mathrm{d}\varepsilon_3$ at ε_1, ε_2, ε_3 can be approximated by Gaussian functions:

$$\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)\mathrm{d}\varepsilon_1\mathrm{d}\varepsilon_2\mathrm{d}\varepsilon_3 = \frac{\exp(-\varepsilon_1^2/2\sigma_\mathrm{D}^2)}{(2\pi)^{1/2}\sigma_\mathrm{D}}\,\mathrm{d}\varepsilon_1 \cdot \frac{\exp(-\varepsilon_2^2/2\sigma_\mathrm{D}^2)}{(2\pi)^{1/2}\sigma_\mathrm{D}}\,\mathrm{d}\varepsilon_2 \cdot \frac{\exp(-\varepsilon_3^2/2\sigma_\mathrm{M}^2)}{(2\pi)^{1/2}\sigma_\mathrm{M}}\,\mathrm{d}\varepsilon_3.$$

2.4. Spot centroids and partiality

The intensity of a reflection can be completely recorded on one image or distributed among several adjacent images. The fraction R_j of total intensity recorded on image j, the 'partiality' of the reflection, can be derived from the distribution function $\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ as

$$R_j = \int_{-\infty}^{\infty} \mathrm{d}\varepsilon_1 \int_{-\infty}^{\infty} \mathrm{d}\varepsilon_2 \int_{\zeta[\varphi_0+(j-1)\Delta_\varphi-\varphi]}^{\zeta(\varphi_0+j\Delta_\varphi-\varphi)} \mathrm{d}\varepsilon_3 \omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$$
$$= \frac{1}{(2\pi)^{1/2}\sigma_\mathrm{M}/|\zeta|} \int_{\varphi_0+(j-1)\Delta_\varphi}^{\varphi_0+j\Delta_\varphi} \exp[-(\varphi' - \varphi)^2/2(\sigma_\mathrm{M}/|\zeta|)^2]\,\mathrm{d}\varphi'$$
$$= \big(\mathrm{erf}[|\zeta|(\varphi_0 + j\Delta_\varphi - \varphi)/(2)^{1/2}\sigma_\mathrm{M}] - \mathrm{erf}\{|\zeta|[\varphi_0 + (j-1)\Delta_\varphi - \varphi]/(2)^{1/2}\sigma_\mathrm{M}\}\big)/2.$$

The integral is evaluated by using a numerical approximation of the error function, erf (Abramowitz & Stegun, 1972).

While the spot centroids in the detector plane are usually good estimates for the detector position of the diffraction maximum, the angular centroid about the rotation axis,

$$Z = \varphi_0 + \Delta_\varphi \cdot \sum_{j=-\infty}^{\infty} (j - \tfrac{1}{2})R_j \simeq \varphi,$$

can be a rather poor guess for the true φ angle of the maximum. Its accuracy depends strongly on the value of φ and the size of the oscillation range $\Delta\varphi$ relative to the mosaicity σ_M of the crystal. For a reflection fully recorded on image j, the value $Z = \varphi_0 + (j - \frac{1}{2})\cdot\Delta_\varphi$ will always be obtained, which is correct only if φ accidentally happens to be close to the centre of the rotation range of the image. In contrast, the φ angle of a partial reflection recorded on images j and $j + 1$ is closely approximated by $Z = \varphi_0 + [j + (R_{j+1} - R_j)/2]\cdot\Delta_\varphi$. If many images contribute to the spot intensity, $Z(\varphi)$ is always an excellent approximation to the ideal angular position φ when the Laue equations are satisfied; in fact, in the limiting case of infinitely fine-sliced data it can be shown that $\lim_{\Delta\varphi\rightarrow 0} Z(\varphi) = \varphi$.

Most refinement routines minimize the discrepancies between the predicted φ angles and their approximations obtained from the observed Z centroids and must therefore carefully distinguish between fully and partially recorded reflections. However, this distinction is unnecessary if the observed Z centroids are instead compared with their analytic forms, because the sensitivity of the centroid positions to the diffraction parameters is correctly weighted in either case (see §2.8).

2.5. Localizing diffraction spots

Often, some of the parameters controlling the diffraction experiment are either completely unknown or available only at a crude approximation. Accurate values for all parameters must be obtained from the recorded data, *i.e.* from a list of the coordinates of strong spots occurring in the images. As implemented in *XDS*, this list is obtained from all or a subset of the data images by the following procedure. Firstly, each pixel value is compared with the mean value and standard deviation of surrounding pixels in the same image and classified as a strong pixel if its value exceeds the mean by a given multiple (typically 3–5) of the standard deviation. Values of the strong pixels and their location addresses and image running numbers are saved in a file. After the scan, a hash table of sufficient size is allocated to accommodate the strong pixels from the file together with their addresses (for a discussion of the hash technique, see Wirth, 1976). As several strong pixels may belong to the same spot, they are labelled with a unique spot number so that any two such pixels which can be connected by direct strong neighbours in two or three dimensions (if there are adjacent images) belong to the same spot (equivalence class). The labelling is achieved by the highly efficient algorithm for the recording of equivalence classes developed by Rem (see Dijkstra, 1976). On termina-

tion, a list X'_i, Y'_i, Z'_i ($i = 1, \ldots, n$) of the centroids of n strong spots is available.

2.6. Basis extraction

Any reciprocal-lattice vector can be written in the form $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$, where h, k, l are integers and $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ are reciprocal basis vectors of the lattice. The basis vectors which describe the orientation, metric and symmetry of the crystal, as well as the reflection indices h, k, l, have to be determined from the list of strong diffraction spots X'_i, Y'_i, Z'_i ($i = 1, \ldots, n$). Ideally, each spot corresponds to a reciprocal-lattice vector $\mathbf{p}_0^*$ which satisfies the Laue equations after a crystal rotation by φ. Substituting the observed value Z' for the unknown φ angle (see §2.4), $\mathbf{p}_0^*$ is found from the observed spot coordinates as

$$\mathbf{p}_0^* = D(\mathbf{m}_2, -Z')(\mathbf{S}' - \mathbf{S}_0)$$
$$\mathbf{S}' = \frac{(X' - X_0)\mathbf{d}_1 + (Y' - Y_0)\mathbf{d}_2 + F\mathbf{d}_3}{\lambda \cdot [(X' - X_0)^2 + (Y' - Y_0)^2 + F^2]^{1/2}}.$$

Unfortunately, the reciprocal-lattice vectors $\mathbf{p}_{0i}^*$ ($i = 1, \ldots, n$) derived from the above list of strong diffraction spots often contain a number of 'aliens' (spots arising from fluctuations in the background, from ice or from satellite crystals) and a robust method has to be used which is still capable of recognizing the dominant lattice. One approach, suggested by Bricogne (1986) and implemented in a number of variants (Otwinowski & Minor, 1997; Steller *et al.*, 1997), is to identify a lattice basis as the three shortest linear independent vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, each at a maximum of the Fourier transform $\sum_{i=1}^{n} \cos(2\pi\mathbf{b} \cdot \mathbf{p}_{0i}^*)$. Alternatively, a reciprocal basis for the dominant lattice can be determined from short differences between the reciprocal-lattice vectors (Howard, 1986; Kabsch, 1988*a*). As implemented in *XDS*, a lattice basis is found by the following procedure.

The list of given reciprocal-lattice points $\mathbf{p}_{0i}^*$ ($i = 1, \ldots, n$) is first reduced to a small number m of low-resolution difference-vector clusters $\mathbf{v}_\mu^*$ ($\mu = 1, \ldots, m$). f_μ is the population of a difference-vector cluster $\mathbf{v}_\mu^*$; that is, the number of times the difference between any two reciprocal-lattice vectors $\mathbf{p}_{0i}^* - \mathbf{p}_{0j}^*$ is approximately equal to $\mathbf{v}_\mu^*$. In a second step, three linear independent vectors $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ are selected among all possible triplets of difference-vector clusters that maximize the function Q,

$$Q(\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*) = \sum_{\mu=1}^{m} f_\mu q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu),$$
$$q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu) = \exp\Bigg(-2 \sum_{k=1}^{3} \{[\max(|\xi_k^\mu - h_k^\mu| - \varepsilon, 0)/\varepsilon]^2 + [\max(|h_k^\mu| - \delta, 0)]^2\} \Bigg),$$

where

$$\xi_k^\mu = \mathbf{v}_\mu^* \cdot \mathbf{b}_k, \quad \mathbf{v}_\mu^* = \sum_{k=1}^{3} \xi_k^\mu \mathbf{b}_k^*, \quad \mathbf{b}_k \cdot \mathbf{b}_l^* = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases}$$

and h_k^μ is the nearest integer to ξ_k^μ. The absolute maximum of Q is assumed if all difference vectors can be expressed as small integral multiples of the best triplet. Deviations from this ideal situation are quantified by the quality measure q. The value of q declines sharply if the expansion coefficients ξ_k^μ deviate by more than ε from their nearest integers h_k^μ or if the indices are absolutely larger than δ. The constraint on the allowed range of indices prevents the selection of a spurious triplet of very short difference-vector clusters which might be present in the set. Excellent results have been obtained using $\varepsilon = 0.05$ and $\delta = 5$. The best vector triplet thus found is refined against the observed difference-vector clusters. Finally, a reduced cell is derived from the refined reciprocal-base vector triplet (see §6).

2.7. Indexing

Once a basis $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ of the lattice is available, integral indices h_i, k_i, l_i must be assigned to each reciprocal-lattice vector $\mathbf{p}_{0i}^*$ ($i = 1, \ldots, n$). Using the integers nearest to $\mathbf{p}_{0i}^* \cdot \mathbf{b}_k$ ($k = 1, 2, 3$) as indices of the reciprocal-lattice vectors $\mathbf{p}_{0i}^*$ could easily lead to a misindexing of longer vectors because of inaccuracies in the basis vectors $\mathbf{b}_k$ and the initial values of the parameters describing the instrumental setup. A more robust solution of the indexing problem is provided by the local indexing method, which assigns only small index differences $h_i - h_j, k_i - k_j, l_i - l_j$ between pairs of neighbouring reciprocal-lattice vectors (Kabsch, 1993).

The reciprocal-lattice points can be considered as nodes of a tree. The tree connects the n points to each other with the connections as its branches. The length ℓ_{ij} of a possible branch between nodes i and j is defined here as

$$\ell_{ij} = 1 - \exp\Bigg(-2 \sum_{k=1}^{3} \{[\max(|\xi_k^{ij} - h_k^{ij}| - \varepsilon, 0)/\varepsilon]^2 + [\max(|h_k^{ij}| - \delta, 0)]^2\} \Bigg),$$

where

$$\xi_k^{ij} = (\mathbf{p}_{0i}^* - \mathbf{p}_{0j}^*) \cdot \mathbf{b}_k,$$

h_k^{ij} is the nearest integer of ξ_k^{ij} and $k = 1, 2, 3$. Reliable index differences are indicated by short branches; in fact, ℓ_{ij} is 0 if none of the indices h_k^{ij} is absolutely larger than δ and the ξ_k^{ij} are integer values to within ε. Typical values are $\varepsilon = 0.05$ and $\delta = 5$. Defining the length of a tree as the sum of the lengths of its branches, a shortest tree among all n^{n-2} possible trees is determined using the elegant algorithm described by Dijkstra (1976). Starting with arbitrary indices 0, 0, 0 for the root node, the local indexing method then consists of traversing the shortest tree and thereby assigning each node the indices of its predecessor plus the small index differences between the two nodes.

During traversal of the tree, each node is also given a subtree number. Starting with subtree number 1 for the root node, each successor node is given the same subtree number as its predecessor if the length of the connecting branch is below a minimal length $\ell_{\min}$. Otherwise, its subtree number is incremented by 1. Thus, all nodes in the same subtree have internally consistent reflection indices. Defining the size of a

subtree by the number of its nodes, 'aliens' are usually found in small subtrees. Finally, a constant index offset is determined such that the centroids of the observed reciprocal-lattice points $\mathbf{p}^*_{0i}$ belonging to the largest subtree and their corresponding grid vectors $\sum_{k=1}^{3} h_k^i \mathbf{b}_k^*$ are as close as possible. This offset is added to the indices of each reciprocal-lattice point.

2.8. Refinement

For a fixed detector, the diffraction pattern depends on the parameters $\mathbf{S}_0$, $\mathbf{m}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, X_0, Y_0, F. Starting values for the parameters can be obtained by the procedures described above, which do not rely on prior knowledge of the crystal orientation, space-group symmetry or unit-cell metric. Better estimates of the parameter values, as required for the subsequent integration step, can be obtained by the method of least squares from the list of n observed indexed reflection centroids h_i, k_i, l_i, X_i', Y_i', Z_i' ($i = 1, \ldots, n$). In this method, the parameters are chosen to minimize a weighted sum of squares of the residuals

$$E = w_X \sum_{i=1}^{n} (\Delta_X^i)^2 + w_Y \sum_{i=1}^{n} (\Delta_Y^i)^2 + w_Z \sum_{i=1}^{n} (\Delta_Z^i)^2.$$

The residuals between the calculated (X_i, Y_i, Z_i) and observed spot centroids are

$$\begin{aligned}
\Delta_X^i &= X_i - X_i' = X_0 + F\mathbf{S}_i \cdot \mathbf{d}_1 / \mathbf{S}_i \cdot \mathbf{d}_3 - X_i' \\
\Delta_Y^i &= Y_i - Y_i' = Y_0 + F\mathbf{S}_i \cdot \mathbf{d}_2 / \mathbf{S}_i \cdot \mathbf{d}_3 - Y_i' \\
\Delta_Z^i &= Z_i - Z_i' = \varphi_0 + \Delta_\varphi \sum_{j=-\infty}^{\infty} (j - \tfrac{1}{2}) R_j^i - Z_i'.
\end{aligned}$$

Let s_μ ($\mu = 1, \ldots, k$) denote the k independent parameters for which initial estimates are available. Expanding the residuals to first order in the parameter changes δs_μ gives

$$\Delta(s_\mu + \delta s_\mu) \simeq \Delta(s_\mu) + \sum_{\mu=1}^{k} \frac{\partial \Delta}{\partial s_\mu} \delta s_\mu.$$

The parameters should be changed in such a way as to minimize $E(\delta s_\mu)$, which implies $\partial E / \partial \delta s_\mu = 0$ for $\mu = 1, \ldots, k$. The δs_μ are found as the solution of the k normal equations

$$\begin{aligned}
\sum_{\mu'=1}^{k} \left(w_X \sum_{i=1}^{n} \frac{\partial \Delta_X^i}{\partial s_\mu} \frac{\partial \Delta_X^i}{\partial s_{\mu'}} + w_Y \sum_{i=1}^{n} \frac{\partial \Delta_Y^i}{\partial s_\mu} \frac{\partial \Delta_Y^i}{\partial s_{\mu'}} + w_Z \sum_{i=1}^{n} \frac{\partial \Delta_Z^i}{\partial s_\mu} \frac{\partial \Delta_Z^i}{\partial s_{\mu'}} \right) \delta s_{\mu'} \\
= -\left(w_X \sum_{i=1}^{n} \Delta_X^i \frac{\partial \Delta_X^i}{\partial s_\mu} + w_Y \sum_{i=1}^{n} \Delta_Y^i \frac{\partial \Delta_Y^i}{\partial s_\mu} + w_Z \sum_{i=1}^{n} \Delta_Z^i \frac{\partial \Delta_Z^i}{\partial s_\mu} \right).
\end{aligned}$$

The parameters are corrected by δs_μ and a new cycle of refinement is started until a minimum of E is reached. The weights

$$w_X = 1 / \sum_{i=1}^{n} (\Delta_X^i)^2, \quad w_Y = 1 / \sum_{i=1}^{n} (\Delta_Y^i)^2, \quad w_Z = 1 / \sum_{i=1}^{n} (\Delta_Z^i)^2$$

are calculated with the current guess for s_μ at the beginning of each cycle.

The derivatives appearing in the normal equations can be worked out from the definitions given in §§2.2 and 2.4 and only the form of the gradient of the Z residuals is shown. Assuming $\sigma_i = \sigma_M / |\zeta_i|$ ($i = 1, \ldots, n$) is constant for each reflection, the gradients of the Z residuals are obtained from the chain rule and the relation $\mathrm{d}\,\mathrm{erf}(z)/\mathrm{d}z = (2/\pi^{1/2}) \exp(-z^2)$.

$$\begin{aligned}
\frac{\partial \Delta_Z^i}{\partial s_\mu} &= \frac{\partial \Delta_Z^i}{\partial \varphi_i} \frac{\partial \varphi_i}{\partial s_\mu} \\
\frac{\partial \Delta_Z^i}{\partial \varphi_i} &= \frac{\Delta_\varphi}{(2\pi)^{1/2} \sigma_i} \sum_{j=-\infty}^{\infty} \exp[-(\varphi_0 + j\Delta_\varphi - \varphi_i)^2 / 2\sigma_i^2] \\
\frac{\partial \varphi_i}{\partial s_\mu} &= \cos\varphi_i \frac{\partial \sin\varphi_i}{\partial s_\mu} - \sin\varphi_i \frac{\partial \cos\varphi_i}{\partial s_\mu}.
\end{aligned}$$

Obviously, $\partial \Delta_Z^i / \partial s_\mu$ is small for a fully recorded reflection because of the small values of all exponentials appearing in $\partial \Delta_Z^i / \partial \varphi_i$. In contrast, the gradient for a partial reflection that is equally recorded on two adjacent images is most sensitive to parameter variations because one of the exponentials assumes its maximum value. In the limiting case of infinitely fine-sliced data it can be shown that $\lim_{\Delta\varphi \to 0} \partial \Delta_Z^i / \partial \varphi_i = 1$. Thus, the refinement scheme based on observed Z centroids, as described here and implemented in *XDS*, is applicable to fine-sliced data and also to data recorded with a large oscillation range.

3. Integration

Assuming that the diffraction parameters have been refined successfully as described above, the intensity of a reflection is distributed in the neighbourhood of the predicted location of the diffraction peak among detector pixels of one or several adjacent rotation images. Accurate integration requires several steps: determination of a reflection mask, estimation of the background, generation of reference profiles and integration by profile fitting.

The intensity distribution of a reflection can be modelled analytically or derived from the observed profiles of neighbouring strong spots. For the rotation method, the profile shape depends strongly on the specific path of the reflection through the Ewald sphere and on variations in the angle of incidence of the diffracted beam on a flat detector. These geometrical distortions can be eliminated by mapping the reflections onto the coordinate system defined in §2.3, which simplifies the task of modelling the expected intensity distribution as all reflection profiles become similar.

3.1. Reflection mask

The parameters σ_M and σ_D of the Gaussian model (see §2.3) used to describe reflection shape can be determined automatically from one or more data images by the following procedure.

(i) Identify and mark strong pixels in the data image.

(ii) Assign the indices of the nearest reflection to each strong pixel.

(iii) Sort the strong pixels by the assigned reflection indices such that pixels with the same indices follow each other in the list.

(iv) For each strong reflection find the rectangular box that encloses all of the strong pixels belonging to the reflection.

(v) Increase the box slightly and use all pixels within the box that are not strong for background determination.

(vi) Subtract the background and determine the centroid and variance s^2 of the intensity-weighted diffracted-beam directions $\lambda\mathbf{S}'$ associated with each strong pixel belonging to the spot (see §2.3).

(vii) Reject the spot if the centroid position deviates too much from the calculated spot location.

(viii) Calculate φ and ζ for the accepted reflection and save the three values φ, ζ and s^2 in a list.

The standard deviation of the beam divergence is obtained directly from this list of n reflections as

$$\sigma_{\rm D}^2 = \frac{1}{n}\sum_{j=1}^{n} s_j^2.$$

Determination of the standard deviation of the reflecting range, the mosaicity $\sigma_{\rm M}$, requires additional considerations. For each of the n reflections from the list above, let τ denote the angular difference between the rotation angle φ at its Bragg maximum and the centre of the oscillation angles covered by the image. The fraction of observed reflection intensity is (see §2.4)

$$R(\tau;\sigma_{\rm M}/\zeta) = \frac{1}{2\Delta_\varphi}\left[\operatorname{erf}\left(\frac{\tau+\Delta_\varphi/2}{2^{1/2}\sigma_{\rm M}/\zeta}\right) - \operatorname{erf}\left(\frac{\tau-\Delta_\varphi/2}{2^{1/2}\sigma_{\rm M}/\zeta}\right)\right].$$

For a given $\sigma_{\rm M}/\zeta$ the function $R(\tau;\sigma_{\rm M}/\zeta)$ assumes its maximum at $\tau = 0$ and declines as $|\tau|$ increases. The decline depends strongly on the mosaicity $\sigma_{\rm M}$ and on the path length of the reflection through the Ewald sphere, which is accounted for by the factor $1/\zeta$. For a large mosaicity $R(\tau;\sigma_{\rm M}/\zeta)$ declines slowly, which explains why for such crystals many reflections with large $|\tau|$ values can be observed on a data image. Clearly, the list of strong spots located by the automatic procedure described above contains information about the mosaicity of the crystal. The problem of finding $\sigma_{\rm M}$ from this list can be solved if one considers each τ value as a random variable drawn from a probability distribution $R(\tau;\sigma_{\rm M}/\zeta)$ with population parameter $\sigma_{\rm M}/\zeta$. The mosaicity $\sigma_{\rm M}$ can then be estimated so that it maximizes the likelihood (joint probability)

$$L(\sigma_{\rm M}) = R(\tau_1;\sigma_{\rm M}/\zeta_1)\cdot R(\tau_2;\sigma_{\rm M}/\zeta_2)\cdots R(\tau_n;\sigma_{\rm M}/\zeta_n).$$

The parameters $\sigma_{\rm D}$ and $\sigma_{\rm M}$ are mainly used to specify the integration region around the spot defined by the parameters $\delta_{\rm M}$ and $\delta_{\rm D}$, which are typically chosen to be 6–10 times larger than $\sigma_{\rm M}$ and $\sigma_{\rm D}$, respectively (see §2.1). The reflection mask thus comprises all image pixels that satisfy

$$|\varepsilon_1| \le \delta_{\rm D}/2, \quad |\varepsilon_2| \le \delta_{\rm D}/2, \quad |\varepsilon_3| \le \delta_{\rm M}/2$$

when mapped to the profile coordinate system $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ defined in §2.3. In addition, pixels are excluded from the mask if they are closer to the predicted Bragg peak of an intruding reflection from the neighbourhood.

3.2. Background

The region around a spot is assumed to have been chosen to be large enough to include a sufficient number of pixels which can be used for determination of the background. Background determination, as implemented in *XDS*, begins by sorting all pixels belonging to a reflection by increasing intensity. For weak or absent reflections, these values should represent a random sample drawn from a normal distribution. If this is not the case, the pixel with the largest intensity is removed until the sampling distribution of the remaining smaller items satisfies the expected distribution. This method will also exclude pixels with unexpected high values, such as ice reflections. The background, determined as the mean value of the accepted pixels, is systematically overestimated for strong spots because of some residual intensity extending into the accepted background pixels. This residual intensity is estimated from the expected distribution $\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ defined in §2.3 and removed from the final background value.

3.3. Standard profiles

Reflection profiles are represented on the Ewald sphere within a domain D_0 comprising $2n_1 + 1$, $2n_2 + 1$ and $2n_3 + 1$ equidistant gridpoints along $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$, respectively. The sampling distances between adjacent grid points are then $\Delta_1 = \delta_{\rm D}/(2n_1 + 1)$, $\Delta_2 = \delta_{\rm D}/(2n_2 + 1)$ and $\Delta_3 = \delta_{\rm M}/(2n_3 + 1)$. Thus, grid coordinate ν_3 ($\nu_3 = -n_3, \ldots, n_3$) covers the set of rotation angles

$$\Gamma_{\nu_3} = \{\varphi' | (\nu_3 - \tfrac{1}{2})\Delta_3 \le (\varphi' - \varphi)\cdot\zeta \le (\nu_3 + \tfrac{1}{2})\Delta_3\}.$$

Contributions to the spot intensity come from one or several adjacent data images ($j = j_1, \ldots, j_2$), each covering the set of rotation angles

$$\Gamma_j = \{\varphi' | \varphi_0 + (j-1)\Delta_\varphi \le \varphi' \le \varphi_0 + j\Delta_\varphi\}.$$

Assuming Gaussian profiles along $\mathbf{e}_3$ for all reflections (see §2.3), the fraction of counts (after subtraction of the background) contributed by data frame j to grid coordinate ν_3 is

$$f_{\nu_3 j} \simeq \frac{\int\limits_{\Gamma_j\cap\Gamma_{\nu_3}} \exp[-(\varphi'-\varphi)^2/2\sigma^2]\,{\rm d}\varphi'}{\int\limits_{\Gamma_j} \exp[-(\varphi'-\varphi)^2/2\sigma^2]\,{\rm d}\varphi'},$$

where $\sigma = \sigma_{\rm M}/|\zeta|$. The integrals can be expressed in terms of the error function, for which efficient numerical approximations are available (Abramowitz & Stegun, 1972). Finally, each pixel in data image j belonging to the reflection is subdivided into 5×5 areas of equal size and $f_{\nu_3 j}/25$ of the pixel signal is added to the profile value at grid coordinates ν_1, ν_3, ν_3 corresponding to each subdivision.

This complicated procedure leads to more uniform intensity profiles for all reflections than using their untransformed shape. This simplifies the task of modelling the expected intensity distribution needed for integration by profile fitting. As implemented in *XDS*, reference profiles are learnt every 5° of crystal rotation at nine positions on the detector, each covering an equal area of the detector face. In the learning phase, profile boxes of the strong reflections are normalized and added to their nearest reference profile boxes. The contributions are weighted according to the distance from the location of the reference profile. Each grid point within the

average profile boxes is classified as signal if it is above 2% of the peak maximum. Finally, each profile is scaled such that the sum of its signal pixels normalizes to one. The analytical expression $\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ defined in §2.3 for the expected intensity distribution is only a rough initial approximation, which is now replaced by the empirical reference profiles.

3.4. Intensity estimation

If an expected intensity distribution $\{p_i | i \in D_0\}$ of the observed profile is given in a domain D_0, the reflection intensity I can be estimated as

$$I = \frac{\sum_{i\in D}(c_i - b_i)p_i/v_i}{\sum_{i\in D} p_i^2/v_i},$$

which minimizes the function

$$\psi(I) = \sum_{i\in D}(c_i - I\cdot p_i - b_i)^2/v_i, \quad \sum_{i\in D_0} p_i = 1.$$

b_i, c_i, v_i ($i \in D$) are the background, contents and variance of pixels observed in a subdomain $D \subseteq D_0$ of the expected distribution. The background b_i underneath a diffraction spot is often assumed to be a constant which is estimated from the neighbourhood around the reflection. Determination of reflection intensities by profile fitting has a long tradition (Diamond, 1969; Ford, 1974; Kabsch, 1988*b*; Otwinowski, 1993). Implementations of the method differ mainly in their assumptions about the variances v_i. Ford used constant variances, which work well for films, which have a high intrinsic background. In *XDS*, which was originally designed for a multwire detector, $v_i \propto p_i$ was assumed, which results in a straight summation of background-subtracted counts within the expected profile region, $I = \sum_{i\in D}(c_i - b_i)/\sum_{i\in D} p_i$. This particular simple formula is very satisfactory for the low background typical of these detectors. For the general case, however, better results can be obtained by using $v_i = b_i + Ip_i$ for the pixel variances as shown by Otwinowski and implemented in *DENZO* and in later versions of *XDS*. Starting with $v_i = b_i$, the intensity is now found by an iterative process which is terminated if the new intensity estimate becomes negative or does not change within a small tolerance, which is usually reached after three cycles. It can be shown that the solution thus obtained is unique.

4. Scaling

The integrated intensities of the reflections need to be corrected by various factors arising from the following

(i) changes in the intensity of the incident beam and variations in the illuminated crystal volume,

(ii) absorption of incident and diffracted beams,

(iii) radiation damage,

(iv) variations in detector sensitivity within the detector plane and

(v) different crystal sizes and crystalline order if the data are from several crystals.

The combined effect manifests itself in correlations of the intensity of a reflection with details of its measurement, such as time (or image number) and location in the detector plane. Usually, many statistically independent observations of symmetry-related reflections are recorded in the rotation images taken from one or several similar crystals of the same compound. The squared structure-factor amplitudes of equivalent reflections should be equal and many scaling procedures (see, for example Evans, 2006; Otwinowski *et al.*, 2003; Kabsch, 1988*b*) exploit this *a priori* knowledge to determine a correction factor for each observed intensity. However, the scaling programs differ in the details of their scaling models, *i.e.* the parametrization and methods used for determination of the correction factors. Below, the approach is described as implemented recently in the programs *XDS* and *XSCALE* (Kabsch, 2010).

If more than one data set is included, these are first put on approximately the same scale by the factor $K\cdot\exp[B\cdot(2\sin\theta/\lambda^2)]$ involving two parameters, K and B, for each data set. The parameter values are assigned so that the resulting correction factors fit best to the observed intensity ratios of common reflections in each pair of data sets.

For the more detailed corrections, three types of two-dimensional functions are used in succession to remove correlations of the intensity of a reflection with (i) image number and resolution, (ii) location in the detector plane and (iii) image number and 13 detector surface regions. To correct for non-uniform detector response such as edge effects at the boundaries of multisegment detectors, the use of smooth analytical correction functions was avoided. Instead, the correction functions are sampled at a finite set of grid regions covering all of the function's definition range. The grid regions are chosen automatically to be as small as possible without overfitting the data so that each sampling region contains more than a specified minimum number of reflections (default 50). Thus, the correction function G is represented by a possibly large number of reciprocal factors G_l, where the subscript l denotes the grid regions.

The correction factors G_l are found in a cyclic prodedure starting with $G_l = 1$. In each cycle, G_l is updated by a factor g_l. The target function for refinement is based on an observational equation for each reflection

$$\psi_{hl} = (I_{hl} - g_l G_l I_h)/\sigma_{hl}$$

as introduced by Hamilton *et al.* (1965). The subscript h represents the unique reflection indices and hl denotes symmetry-related reflections to h that need to be corrected by the reciprocal scaling factor g_l associated with grid point l; I_{hl} and σ_{hl} are their weighted mean and standard deviation, respectively. This standard deviation is considered to be infinitely large if no such reflection was measured, which amounts to omitting the observational equation altogether. The factors g_l and the 'true' intensities I_h are found at the minimum of the function

$$\Psi = \sum_{hl}\psi_{hl}^2 + \sum_l (g_l G_l - 1)^2/\sigma^2.$$

The first sum on the right side is a homogeneous function of g_l of degree zero so that the g_l would only be determined up to an arbitrary factor. The second sum on the right side is used to weakly restrain the scaling factors to one; a reasonable value is $\sigma = 0.05$. Minimization of Ψ leads to updates g_l in terms of the 'true' intensities I_h which again depend on g_l,

$$g_l = \frac{\sum_h I_h(I_{hl}/G_l)/(\sigma_{hl}/G_l)^2 + G_l/\sigma^2}{\sum_h I_h^2/(\sigma_{hl}/G_l)^2 + G_l^2/\sigma^2}$$

$$I_h = \frac{\sum_l g_l(I_{hl}/G_l)/(\sigma_{hl}/G_l)^2}{\sum_l g_l^2/(\sigma_{hl}/G_l)^2}$$

$$I_h^o = \frac{\sum_l (I_{hl}/G_l)/(\sigma_{hl}/G_l)^2}{\sum_l 1/(\sigma_{hl}/G_l)^2}.$$

The new update factors g_l are obtained by using 'true' intensities I_h^o from the previous cycle instead of the current I_h as defined above. At the end of the cycle, the old correction factors G_l are updated by multiplication with the new g_l. This cyclic procedure typically converges in less than six cycles.

The approach described here has been implemented in *XDS* and *XSCALE* and has been successfully used for more than two years. In contrast to the 'shortest path' eigenvector method of Fox & Holmes (1966), which is very efficient for a relatively small number of variables, the computations here require a time that is proportional to the number of reflections used for scaling and thus quickly lead to a solution even when a very large number of correction factors from many data sets are involved.

5. Post-refinement

The number of fully recorded reflections on each single image rapidly declines for small oscillation ranges and the complete intensities of the partially recorded reflections have to be estimated. This presented a serious obstacle in early structural work on virus crystals, as the crystal had to be replaced after each exposure on account of radiation damage. A solution to this problem, the 'post-refinement' technique, was found by Schutt, Winkler and Harrison and variants of this powerful method have been incorporated into most data-reduction programs (for a detailed discussion, see Harrison *et al.*, 1985; Rossmann, 1985). The method derives complete intensities of reflections that are only partially recorded on an image from accurate estimates for the fractions of observed intensity: the 'partiality'. The partiality of each reflection can always be calculated as a function of orientation, unit-cell metric, mosaic spread of the crystal and model intensity distributions. The accuracy of the estimated full reflection intensity obviously then strongly depends on a precise knowledge of the parameters describing the diffraction experiment. Usually, symmetry-related fully recorded reflections can be found for many of the partial reflections and the list of such pairs of intensity observations can be used to refine the required parameters using a least-squares procedure. Clearly, this refinement is carried out after all images have been processed, which explains why the procedure is called 'post-refinement'.

Adjustments of the diffraction parameters s_μ ($\mu = 1, \ldots, k$) are determined by minimization of the function E, which is defined as the weighted sum of squared residuals between calculated and observed partial intensites.

$$E = \sum_{hj} w_{hj}(\Delta_{hj})^2$$

$$\Delta_{hj} = R_j(\varphi_{hj})g_j I_h - I_{hj}$$

$$w_{hj} = 1/\{\sigma^2(I_{hj}) + [R_j(\varphi_{hj})g_j]^2\sigma^2(I_h)\}.$$

Here, I_{hj} is the intensity recorded on image j of a partial reflection with indices summarized as hj, I_h is the mean of the observed intensities of all fully recorded reflections symmetry-equivalent to hj, g_j is the inverse scaling factor of image j, φ_{hj} is the calculated spindle angle of reflection hj at diffraction and R_j is the computed fraction of total intensity recorded on image j.

Expansion of the residuals Δ_{hj} to first order in the parameter changes δs_μ and minimization of $E(\delta s_\mu)$ leads to the k normal equations

$$\sum_{\mu'=1}^{k}\left(\sum_{hj} w_{hj}\frac{\partial\Delta_{hj}}{\partial s_\mu}\frac{\partial\Delta_{hj}}{\partial s_{\mu'}}\right)\delta s_{\mu'} = -\sum_{hj} w_{hj}\Delta_{hj}\frac{\partial\Delta_{hj}}{\partial s_\mu}.$$

Often, the normal matrix is ill-conditioned since changes in some unit-cell parameters or small rotations of the crystal about the incident X-ray beam do not significantly affect the calculated partiality R_j. To take care of these difficulties, the system of equations is rescaled to yield unit diagonal elements for the normal matrix and the correction vector δs_μ is filtered by projection into a subspace defined by the eigenvectors of the normal matrix with sufficiently large eigenvalues (Diamond, 1966).

The parameters are corrected by the filtered δs_μ and a new cycle of refinement is started until a minimum of E is reached. The weights, residuals and their gradients are calculated using the current values for s_μ and g_j at the beginning of each cycle. The derivatives

$$\frac{\partial\Delta_{hj}}{\partial s_\mu} = g_j I_h\left(\frac{\partial R_j}{\partial\varphi_{hj}}\frac{\partial\varphi_{hj}}{\partial s_\mu} + \frac{\partial R_j}{\partial\sigma_M}\frac{\partial\sigma_M}{\partial s_\mu} + \frac{\partial R_j}{\partial|\zeta_{hj}|}\frac{\partial|\zeta_{hj}|}{\partial s_\mu}\right)$$

appearing in the normal equations can be worked out from the definitions given in §§2.2 and 2.4 (to simplify the following equations, the subscript hj is omitted). The fraction R_j of total intensity can be expressed in terms of the error function (see §2.4) as

$$R_j = [\mathrm{erf}(z_1) - \mathrm{erf}(z_2)]/2$$

$$z_1 = |\zeta|(\varphi_0 + j\Delta_\varphi - \varphi)/2^{1/2}\sigma_M$$

$$z_2 = |\zeta|[\varphi_0 + (j-1)\Delta_\varphi - \varphi]/2^{1/2}\sigma_M.$$

Using the relation $\mathrm{d}\,\mathrm{erf}(z)/\mathrm{d}z = (2/\pi^{1/2})\exp(-z^2)$, the derivatives of R_j are

Table 1
Rating of lattice types implied by a given reduced cell.

Lattice type		Quality index	Conventional unit-cell parameters (Å, °)						Reindexing transformation
			a	*b*	*c*	α	β	γ	
44	*aP*	0.0	159.3	159.4	160.4	90.1	90.1	90.1	$11\bar{1}0/1\bar{1}10/\bar{1}\bar{1}\bar{1}0$
31	*aP*	0.4	159.3	159.4	160.4	90.1	89.9	89.9	1000/0100/0010
34	*mP*	1.4	159.3	160.4	159.4	90.1	90.1	90.1	$\bar{1}000/00\bar{1}0/0\bar{1}00$
14	*mC*	1.4	225.1	225.6	160.4	90.0	90.1	89.9	$1100/\bar{1}100/0010$
33	*mP*	1.5	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
35	*mP*	2.0	159.4	159.3	160.4	90.1	90.1	90.1	$0\bar{1}00/\bar{1}000/00\bar{1}0$
13	*oC*	2.3	225.1	225.6	160.4	90.0	90.1	89.9	$1100/\bar{1}100/0010$
32	*oP*	2.4	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
10	*mC*	2.5	225.1	225.6	160.4	90.0	90.1	90.1	$\bar{1}\bar{1}00/\bar{1}100/00\bar{1}0$
11	***tP***	**3.4**	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
25	*mC*	5.9	226.0	226.2	159.3	90.0	90.2	89.7	$0110/0\bar{1}10/1000$
20	*mC*	6.4	226.0	226.2	159.3	90.0	90.2	90.3	$0\bar{1}\bar{1}0/0\bar{1}10/\bar{1}000$
4	*hR*	7.4	225.6	226.2	276.2	90.3	89.9	119.9	$1\bar{1}00/\bar{1}010/\bar{1}\bar{1}\bar{1}0$
23	*oC*	7.8	226.0	226.2	159.3	90.0	90.2	89.7	$0110/0\bar{1}10/1000$
3	*cP*	7.8	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
21	*tP*	8.2	159.4	160.4	159.3	90.1	90.1	90.1	0100/0010/1000
2	*hR*	8.7	225.1	225.8	276.9	90.2	90.0	119.8	$1100/\bar{1}0\bar{1}0/\bar{1}110$
5	*cI*	173.6	225.8	225.1	226.0	60.2	59.9	60.2	1010/1100/0110

$$\frac{\partial R_j}{\partial \varphi} = [\exp(-z_2^2) - \exp(-z_1^2)]|\zeta|/[\sigma_M(2\pi)^{1/2}]$$
$$\frac{\partial R_j}{\partial \sigma_M} = [z_2 \exp(-z_2^2) - z_1 \exp(-z_1^2)]/(\sigma_M \pi^{1/2})$$
$$\frac{\partial R_j}{\partial |\zeta|} = [z_1 \exp(-z_1^2) - z_2 \exp(-z_2^2)]/(|\zeta| \pi^{1/2}).$$

The derivatives $\partial\varphi/\partial s_\mu$, $\partial\sigma_M/\partial s_\mu$ and $\partial|\zeta|/\partial s_\mu$ remain to be worked out (not shown here). As discussed in detail by Greenhough & Helliwell (1982), spectral dispersion and asymmetric beam cross-fire lead to some variation in σ_M, which makes it necessary to include additional parameters in the list s_μ. The effect of these parameters on the partiality is dealt with easily by the derivatives $\partial\sigma_M/\partial s_\mu$.

The refinement scheme described above requires initial scaling factors g_j. With the now improved estimates for the partialities R_j, a new set of scaling factors can be obtained using the method outlined in §4. This alternating procedure of scaling and post-refinement usually converges within three cycles.

The use of error functions for modelling partiality, as implicated by a Gaussian model for describing spot shape, was chosen here for reasons of conceptual simplicity and coherence. This choice is unlikely to significantly alter the results of post-refinement that are based on other functions of similar form (see the discussion by Rossmann, 1985).

6. Space-group assignment

Identification of the correct space group is not always an easy task and should be postponed for as long as possible. Sometimes, the true space group only becomes known when the structure has been successfully solved and refined! However, one can expect to identify a small number of possibilities from the diffraction experiment.

Fortunately, all data processing as implemented in the program *XDS* can be carried out in the absence of any knowledge of the crystal symmetry and unit-cell parameters. In this case, a reduced cell is extracted from the observed diffraction pattern and processing of the data images continues to completion as if the crystal were triclinic. Clearly, the reflection indices then refer to the reduced cell and must be reindexed once the space group is known. For all space groups, the required reindexing transformation is linear and involves only whole numbers, as shown in Part 9 of Vol. *A* of *International Tables for Crystallography* (1989).

Automatic space-group assignment is carried out in two steps once integrated intensities of all reflections are available (see Kabsch, 2010). Firstly, the Bravais lattices are identified that are compatible with the reduced cell derived from the observed diffraction pattern. In the second step, all enantiomorphous space groups compatible with the observed lattice symmetry are rated by a redundancy-independent *R* factor. The group is selected that explains all integrated intensities in the data set at an acceptable *R* factor requiring a minimum number of unique reflections (Occam's principle). This approach deliberately avoids any test for the presence of screw axes as these tests would depend strongly on the completeness of the data. Fortunately, the presence or absence of screw axes is irrelevant for the determination of data correction/scaling factors (see §4).

6.1. Determination of the Bravais lattice

The determination of possible Bravais lattices is based upon the concept of the reduced cell whose metric parameters characterize 44 lattice types as described in Part 9 of Vol. *A* of *International Tables for Crystallography* (1989). A primitive basis $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ of a given lattice is defined there as a reduced cell if it is right-handed and if the components of its metric tensor

$$A = \mathbf{b}_1 \cdot \mathbf{b}_1, \quad B = \mathbf{b}_2 \cdot \mathbf{b}_2, \quad C = \mathbf{b}_3 \cdot \mathbf{b}_3$$
$$D = \mathbf{b}_2 \cdot \mathbf{b}_3, \quad E = \mathbf{b}_1 \cdot \mathbf{b}_3, \quad F = \mathbf{b}_1 \cdot \mathbf{b}_2$$

Table 2
Identification of possible space groups.

							Conventional unit-cell parameters (Å, °)					
Space group		Lattice type		R_{meas} (%)	UNIQUE	COMPARED	*a*	*b*	*c*	α	β	γ
1	*P*1	44	*aP*	5.8	35341	22207	159.3	159.4	160.4	90.1	90.1	90.1
1	*P*1	31	*aP*	5.8	35341	22207	159.3	159.4	160.4	90.1	89.9	89.9
3	*P*2	33	*mP*	6.5	21904	35644	159.3	159.4	160.4	90.0	90.1	90.0
3	*P*2	34	*mP*	7.0	26743	30805	159.3	160.4	159.4	90.0	90.1	90.0
5	*C*2	10	*mC*	7.7	22207	35341	225.1	225.6	160.4	90.0	90.1	90.0
5	*C*2	14	*mC*	7.7	22207	35341	225.1	225.6	160.4	90.0	90.1	90.0
16	*P*222	32	*oP*	7.9	14461	43087	159.3	159.4	160.4	90.0	90.0	90.0
21	*C*222	13	*oC*	8.0	15094	42454	225.1	225.6	160.4	90.0	90.0	90.0
3	*P*2	35	*mP*	8.2	25786	31762	159.4	159.3	160.4	90.0	90.0	90.0
75	*P*4	11	*tP*	8.5	14944	42604	159.4	159.4	160.4	90.0	90.0	90.0
89	***P*422**	**11**	***tP***	**9.0**	8086	49462	159.4	159.4	160.4	90.0	90.0	90.0
146	*R*3	2	*hR*	45.2	20068	37480	225.5	225.5	276.9	90.0	90.0	120.0
5	*C*2	20	*mC*	46.9	23125	34423	226.0	226.2	159.3	90.0	90.2	90.0
5	*C*2	25	*mC*	46.9	23125	34423	226.0	226.2	159.3	90.0	90.2	90.0
75	*P*4	21	*tP*	49.2	14828	42720	159.9	159.9	159.3	90.0	90.0	90.0
89	*P*422	21	*tP*	50.7	7876	49672	159.9	159.9	159.3	90.0	90.0	90.0
21	*C*222	23	*oC*	51.3	15155	42393	226.0	226.2	159.3	90.0	90.0	90.0
195	*P*23	3	*cP*	57.3	5344	52204	159.7	159.7	159.7	90.0	90.0	90.0
207	*P*432	3	*cP*	58.1	2896	54652	159.7	159.7	159.7	90.0	90.0	90.0
155	*R*32	4	*hR*	59.7	9038	48510	225.9	225.9	276.2	90.0	90.0	120.0
155	*R*32	2	*hR*	60.7	10487	47061	225.5	225.5	276.9	90.0	90.0	120.0
146	*R*3	4	*hR*	61.1	16751	40797	225.9	225.9	276.2	90.0	90.0	120.0

satisfy a number of conditions (inequalities). The main conditions state that the basis vectors are the shortest three linear independent lattice vectors with either all acute or all non-acute angles between them. As specified in *International Tables for Crystallography*, each of the 44 lattice types is characterized by additional equality relations among the six components of the reduced-cell metric tensor. As an example, for lattice character 11 (Bravais type *tP*) the components of the metric tensor of the reduced cell must satisfy

$$A = B, \quad B \leq C, \quad D = 0, \quad E = 0, \quad F = 0.$$

(Note that the other tetragonal primitive lattice character 21 requires $A \leq B = C$ with the fourfold as the shortest axis.)

Any primitive triclinic cell describing a given lattice can be converted into a reduced cell. It is well known, however, that the reduced cell thus derived is sensitive to experimental error. Hence, the direct approach of first deriving the correct reduced cell and then finding the lattice type is unstable and may in certain cases even prevent identification of the correct Bravais lattice.

A suitable solution of the problem has been found that avoids any decision as to what the 'true' reduced cell is (see Kabsch, 1993). The essential ingredients of this procedure are (i) a database of possible reduced cells and (ii) a backward-search strategy that finds the best-fitting cell in the database for each lattice type.

The database is derived from a seed cell which strictly satisfies the definitions for a reduced cell. All cells of the same volume as the seed cell whose basis vectors can be linearly expressed in terms of the seed vectors by indices -1, 0 or $+1$ are included in the database. Each unit cell in the database is considered as a potential reduced cell, although some of the defining conditions as given in Part 9 of Vol. *A* of *International Tables for Crystallography* (1989) may be violated. These violations are treated as arising from experimental error.

The backward-search strategy starts with the hypothesis that the lattice type is already known and identifies the best-fitting cell in the database of possible reduced cells. In contrast to a forward-directed search, it is now always possible to decide which conditions have to be satisfied by the components of the metric tensor of the reduced cell. The total amount by which all these equality and inequality conditions are violated is used as a quality index. For example, to find out how well a potential reduced cell $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ from the database characterizes lattice character 11 (Bravais lattice *tP*), the quality index

$$p_{11}(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3) = |A - B| + \max(0, B - C) + |D| + |E| + |F|$$

is computed. Positive values of p_{11} indicate that some conditions are not satisfied. All potential reduced cells in the data base are tested and the smallest value for p_{11} is used for rating lattice type 11. A similar test is carried out for all 44 possible lattice types using quality indices based on their defining conditions as listed in Part 9 of Vol. *A* of *International Tables for Crystallography* (1989).

The results obtained using this method are shown in Table 1 for the example of a data set comprising 177 images with each exposure covering 0.5° of spindle rotation. The space group of the protein crystal was $P4_32_12$ (unit-cell parameters a = 159.4, b = 159.4, c = 160.3 Å), but this knowledge was not used in the processing. Instead, the data were processed with respect to a triclinic reduced cell derived from the observed diffraction pattern as described above. The images contained a total of 292 998 reflections within the resolution range 20.0–3.0 Å; 57 548 reflections in the resolution range 10.0–5.0 Å were used for space-group determination. For determination of the lattice symmetry all 44 possibilities were considered and rated

by their quality index. The table shows the possible lattice symmetries, their implied conventional unit-cell parameters and a reindexing transformation. The table entries are sorted by increasing quality index and reveal a nearly cubic lattice symmetry. A lattice symmetry is considered to be acceptable if it has a low quality index and its implied unit-cell parameters do not violate the ideal values by more than 3.0° in angles and 3% in cell axes. Thus, except for the last entry, all of the lattice symmetries in the table are acceptable; the correct lattice type 11 *tP* is highlighted. Lattice symmetries that are not accepted include all body-centred lattices or those that are centred on all faces; they are omitted from the table.

The reindexing transformation REIDX() consists of 12 integers that relate the original indices h, k, l used during the integration to the indices h', k', l' with respect to the new cell.

$$h' = \frac{[\mathrm{REIDX}(1)\cdot h + \mathrm{REIDX}(2)\cdot k + \mathrm{REIDX}(3)\cdot l]}{\mathrm{IDXV}} + \mathrm{REIDX}(4)$$

$$k' = \frac{[\mathrm{REIDX}(5)\cdot h + \mathrm{REIDX}(6)\cdot k + \mathrm{REIDX}(7)\cdot l]}{\mathrm{IDXV}} + \mathrm{REIDX}(8)$$

$$l' = \frac{[\mathrm{REIDX}(9)\cdot h + \mathrm{REIDX}(10)\cdot k + \mathrm{REIDX}(11)\cdot l]}{\mathrm{IDXV}} + \mathrm{REIDX}(12).$$

The value of the integer IDXV depends on the lattice type used for specifying reflection indices in the integration step. IDXV is 1 for a primitive lattice, 2 for a face-centred or body-centred lattice, 3 for a rhombohedral lattice and 4 for a lattice centred on all faces. In the example case we have IDXV = 1 because integration was carried out in space group *P*1.

Note also that elements 4, 8 and 12 of the transformation are always 0 in this example. These three extra elements were introduced to provide a simple tool for correcting the indices if all reflections are misindexed by a constant.

6.2. Finding possible space groups

For protein crystals, the absence of parity-changing symmetry operators restricts the number of possible space groups to 65 instead of 230. Moreover, the determination of correction factors for the integrated intensities does not depend on the presence or absence of any screw axes so that data processing can be finished without this knowledge. This reduces the problem to the identification of an enantiomorphous space group without screw axes that is compatible with the observed lattice symmetry (see above).

For solution of the problem, a quality indicator of the mean variation in the intensities of symmetry-equivalent reflections ($R_{\rm meas}$) is calculated for each possible group. The decision for a particular group is then based on Occam's principle: the selected group must explain all integrated intensities in the data set at acceptable quality, thereby requiring a minimum number of unique reflections.

A suitable redundancy-independent data quality indicator has been suggested by Diederichs & Karplus (1997) and Weiss (2001),

$$R_{\rm meas} \equiv R_{\rm r.i.m.} = \frac{\sum_{hl}\left(\frac{n_h}{n_h-1}\right)^{1/2}|I_{hl}-I_h|}{\sum_{hl} I_{hl}}.$$

The subscript h represents the unique reflection indices and hl denotes any of the n_h symmetry-related reflections to h. The absolute differences between the observed intensities I_{hl} and their mean intensity I_h are weighted to remove any dependency on n_h and compared with the intensities. Small values of $R_{\rm meas}$ indicate accurate single observations I_{hl} and the use of symmetry operators compatible with the intensity data set.

For the above example data set, Table 2 lists all enantiomorphous groups which are in harmony with the observed lattice symmetry shown in Table 1. For each listed space group, UNIQUE is the number of unique reflections and COMPARED is the number of reflections used to calculate the redundancy-independent R factor $R_{\rm meas}$. Two sets of groups can be distinguished clearly: those implying an acceptable $R_{\rm meas}$ and a second set with $R_{\rm meas} > 45\%$, which is totally unacceptable. Among the acceptable solutions a minimum number of unique reflections is needed if the crystal has the tetragonal space-group symmetry *P*422.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover Publications.

Bricogne, G. (1986). *Proceedings of the EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase III)*, p. 28. Paris: LURE.

Diamond, R. (1966). *Acta Cryst.* **21**, 253–266.

Diamond, R. (1969). *Acta Cryst.* A**25**, 43–55.

Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.

Dijkstra, E. W. (1976). *A Discipline of Programming*, pp. 154–167. New Jersey: Prentice–Hall.

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Ford, G. C. (1974). *J. Appl. Cryst.* **7**, 555–564.

Fox, G. C. & Holmes, K. C. (1966). *Acta Cryst.* **20**, 886–891.

Greenhough, T. J. & Helliwell, J. R. (1982). *J. Appl. Cryst.* **15**, 338–351.

Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *Acta Cryst.* **18**, 129–130.

Harrison, S. C., Winkler, F. K., Schutt, C. E. & Durbin, R. M. (1985). *Methods Enzymol.* **114**, 211–237.

Howard, A. (1986). *Proceedings of the EEC Cooperative Workshop on Position-Sensitive Detector Software (Phases I and II)*, pp. 89–94 Paris: LURE.

International Tables for Crystallography (1989). Vol. *A*, pp. 738–749. Dordrecht: Kluwer Academic Publishers.

International Tables for Crystallography (2001). Vol. *F*, ch. 25.2, pp. 695–743. Dordrecht: Kluwer Academic Publishers.

Kabsch, W. (1988*a*). *J. Appl. Cryst.* **21**, 67–71.

Kabsch, W. (1988*b*). *J. Appl. Cryst.* **21**, 916–924.

Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.

Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Acta Cryst.* A**59**, 228–234.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pflugrath, J. W. (1997). *Methods Enzymol.* **276**, 286–306.
Rossmann, M. G. (1985). *Methods Enzymol.* **114**, 237–280.
Schutt, C. & Winkler, F. K. (1977). *The Rotation Method in Crystallography*, edited by U. W. Arndt & A. J. Wonacott, pp. 173–186. Amsterdam: North-Holland.
Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Wirth, N. (1976). *Algorithms + Data Structures = Programs*, pp. 264–274. New York: Prentice–Hall.