Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Data processing

**Michael G. Rossmann* and Cornelis G. van Beek**

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392, USA

Correspondence e-mail: mgr@indiana.bio.purdue.edu

Received 29 January 1999
Accepted 22 June 1999

X-ray diffraction data processing proceeds through indexing, pre-refinement of camera parameters and crystal orientation, intensity integration, post-refinement and scaling. The *DENZO* program has set new standards for autoindexing, but no publication has appeared which describes the algorithm. In the development of the new *Data Processing Suite* (*DPS*), one of the first aims has been the development of an autoindexing procedure at least as powerful as that used by *DENZO*. The resultant algorithm will be described. Another major problem which has arisen in recent years is scaling and post-refinement of data from different images when there are few, if any, full reflections. This occurs when the mosaic spread approaches or exceeds the angle of oscillation, as is usually the case for frozen crystals. A procedure which is able to obtain satisfactory results for such a situation will be described.

## 1. Introduction

Intensity data estimation has been an integral part of structural biology since Bragg used an ionization chamber technique to determine the energy of diffracted reflections from simple salts. Two alternative types of detector have been used: point detectors, which measure the energy of a single reflection, and area detectors [*e.g.* films, imaging plates, wire detectors, charge-coupled devices (CCDs)], which collect numerous reflections on the same two-dimensional device. The latter gave rise to the early rotation and oscillation photography and, subsequently, to the analogue Weissenberg and precession cameras. However, these cameras required the screening out of most of the diffracted rays in order to concentrate on the recording of only a single reciprocal-lattice plane. Xuong *et al.* (1968) and Arndt *et al.* (1973) pointed out that mostly non-overlapping reflections can be selected by removing the screen but reducing the oscillation or precession angles. Fortunately, two-dimensional film-scanning devices became available at about that time, which allowed for both accurate positional determination as well as intensity determination of reflections.

Subsequent to the publication of *The Rotation Method in Crystallography* (Arndt & Wonacott, 1977), the oscillation technique became the method of choice for intensity estimation of diffraction patterns from crystals of biological macromolecules. During the first decade or so of oscillation photography, it was the practice to carefully 'set' a crystal with its axes oriented in known directions relative to the camera axes. The 'American method' (shoot first, think later) was

introduced by Rossmann & Erickson (1983) to avoid radiation damage during the tedious crystal-setting operation and to enhance the rate of data collection while using precious synchrotron time. However, the American method required that a good indexing system was available for determining the crystal setting.

Various methods of determining X-ray intensities were described in the Arndt and Wonacott book (Arndt & Wonacott, 1977). We developed the Purdue system (Rossmann, 1979; Rossmann *et al.*, 1979) on which the popular *DENZO* or *HKL* system was originally based (Otwinowski & Minor, 1997). As the precise algorithms used by *HKL* are not available, we initiated a project to update our old (1979) procedures. We are developing the *Data Processing Suite* (*DPS*) available at, for instance, the Cornell High Energy Synchrotron Source (CHESS), as well as other synchrotron beamlines (see also http://bilbo.bio.purdue.edu/~viruswww/Rossmann_home/rstest.html). This has been performed in collaboration with MacCHESS (Steve Ealick, Dan Thiel, Marian Szebenyi) and Chris Nielson of Area Detector Systems Corp.

Modern data processing can be divided into a series of steps.

(i) Autoindexing. This requires a peak-picking procedure (*c.f.* Kim, 1989), followed by an analysis of the position of the peaks to determine unit-cell dimensions, Bravais lattice and crystal orientation.

(ii) Pre-refinement of the camera parameters (crystal-to-detector distance, scanning direction relative to oscillation direction, detector tilt away from being normal to the X-ray beam), crystal orientation and effective mosaic spread (actual mosaic spread convoluted with beam divergence).

(iii) Intensity integration by profile fitting, assuming reflection position as calculated from the pre-refined camera and crystal parameters. [Error estimates can be made for each reflection; overlap and overloaded (non-linear response of detector) corrections can be applied; partiality of reflections can be computed.]

(iv) Lorentz and polarization corrections, followed by reduction to a unique asymmetric unit in reciprocal space (this is a Laue-group-dependent step). The reflections then need to be sorted on the basis of their indices reduced to a selected asymmetric unit in reciprocal space. This permits ready comparison of symmetry-related reflections which will be adjacent in the reflection list.

(v) Scaling of images onto a common scale.

(vi) Display of input and output on a graphical user interface.

Although unpublished, *DENZO* has an exceptionally good autoindexing procedure. It was clear that *DPS* would require an algorithm (*cf.* Steller *et al.*, 1997) at least as good as *DENZO* if it were to become useful. As older scaling procedures depended primarily on the matching of whole reflections, we recognized that with the advent of frozen crystals and the correspondingly larger mosaic spreads, new methods of scaling and post-refinement were also required (Bolotovsky *et al.*, 1998). We describe here our *DPS* procedures, which use our autoindexing algorithm, *MOSFLM* (Leslie, 1992) for integration and have the option of using *SCALA* or our *SNP* scaling procedure.

## 2. Autoindexing – introduction

A variety of techniques was suggested to determine the crystal orientation, some of which required initial knowledge of the unit-cell dimensions (Vriend & Rossmann, 1987; Kabsch, 1988), while more advanced techniques (Kim, 1989; Higashi, 1990; Kabsch, 1993) determined both unit-cell dimensions and crystal orientation. All these methods start with the determination of the reciprocal-lattice vectors, assuming that the oscillation photographs are 'stills'. The methods of Higashi and of Kabsch, as well as, in part, Kim, analyze the distribution of the difference vectors generated from the reciprocal-lattice vectors. The most frequent difference vectors are taken as the basis vectors defining the reciprocal-lattice unit cell and its orientation. In addition, Kim's technique requires the input of the orientation of a likely zone-axis direction onto which the reciprocal-lattice vectors are then projected. The projections will have a periodicity distribution consistent with the reciprocal-lattice planes perpendicular to the zone axis. Duisenberg (1992) used a similar approach for single-point detector data, although he did not rely on prior knowledge of the zone-axis direction.

A major advance was made in the program *DENZO*, a part of the *HKL* package (Otwinowski & Minor, 1997), which not only has a robust indexing procedure but also has a useful graphical interface. The indexing technique used in the procedure has not been described, except for a few hints in the manual on the use of a fast Fourier transform (FFT). Indeed, Bricogne (1986) suggested that a three-dimensional Fourier transformation might be a powerful indexing tool. However, for large unit cells, this procedure requires an excessive amount of memory and time (Campbell, 1997).

## 3. The crystal orientation matrix

The position $\mathbf{x}$ $(x,y,z)$ of a reciprocal-lattice point can be given as

$$\mathbf{x} = [\Phi][A]\mathbf{h}. \tag{1}$$

The matrix $[\Phi]$ is a rotation matrix around the camera's spindle axis for a rotation of $\varphi$. The vector $\mathbf{h}$ represents the Miller indices $(hkl)$ and $[A]$ defines the reciprocal unit-cell dimensions and the orientation of the crystal lattice with respect to the camera axes when $\varphi = 0$. Thus,

$$[A] = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}, \tag{2}$$

where $a_x^*$, $a_y^*$ and $a_z^*$ are the components of the crystal $a^*$ axis with respect to the orthogonal camera axes. When an oscillation image is recorded, the position of a reciprocal-lattice point is moved from $\mathbf{x}_1$ to $\mathbf{x}_2$, corresponding to a rotation of the crystal from $\varphi_1$ to $\varphi_2$. The recorded position of the reflection on the detector corresponds to the point $\mathbf{x}$ when it is on the

Ewald sphere somewhere between $\mathbf{x}_1$ and $\mathbf{x}_2$. The actual value of $\varphi$ at which this crossing occurs cannot be retrieved directly from the oscillation image. We shall, therefore, assume here, as is the case in all other procedures, that $[\Phi][A]$ defines the crystal orientation in the center of the oscillation range. Defining the camera axes as in Rossmann (1979), it is easy to show that a reflection recorded at the position $(X,Y)$ on a flat detector normal to the X-ray beam at a distance $D$ from the crystal corresponds to

$$\begin{aligned} x &= X/[\lambda(X^2 + Y^2 + D^2)^{1/2}], \\ y &= Y/[\lambda(X^2 + Y^2 + D^2)^{1/2}], \\ z &= D/[\lambda(X^2 + Y^2 + D^2)^{1/2}], \end{aligned} \tag{3}$$

where $\lambda$ is the X-ray wavelength.

If an approximate $[A]$ matrix is available, the Miller indices of an observed peak at $(X,Y)$ can be roughly determined using (3) and (1), where

$$\mathbf{h} = [A]^{-1}[\Phi]^{-1}\mathbf{x}, \tag{4}$$

with the error being dependent on the width of the oscillation range, the error in the detector parameters and errors in determining the coordinates of the centers of the recorded reflections.

## 4. Fourier analysis of the reciprocal-lattice vector distribution when projected onto a chosen direction

If the members of a set of reciprocal-lattice planes perpendicular to a chosen direction **t** are well separated, then the projections of the reciprocal-lattice vectors onto **t** will have an easily recognizable periodic distribution. Unlike the procedure of Kim (1989), which requires the input of a likely zone-axis direction, the present procedure tests all possible directions and analyzes the frequency distribution $f(j)$ of the projected reciprocal-lattice vectors in each case. Also, unlike the procedure of Kim, the periodicity is determined using a one-dimensional FFT (Fig. 1).

## 5. Exploring all possible directions to find a good set of basis vectors

The polar coordinates $\psi,\varphi$ define the direction **t**, where $\psi$ defines the angle between the X-ray beam and the chosen direction **t**. The Fourier analysis is performed for each direction **t** in the range $0 < \psi \leq \pi/2, 0 < \varphi \leq 2\pi$. A suitable angular increment in $\psi$ was determined empirically to be about 0.03 rad (1.7°). For each value of $\psi$, the increment in $\varphi$ is taken to be the closest integral value to $(2\pi\sin\psi)/0.03$. This procedure results in ~7300 separate roughly equally spaced directions. For each direction **t**, the distribution of the corresponding $F(k)$ coefficients is surveyed to locate the largest local maximum. The $\psi$ and $\varphi$ values associated with the 30 largest maxima are selected for refinement by a local search procedure to obtain an accuracy of $10^{-4}$ rad (~0.006°). Directions are chosen from these vectors to give a linearly independent set of three basis vectors of a primitive real-space unit cell. These are then converted to the basis vectors of the reciprocal cell. The components of the three reciprocal-cell axes along the three camera axes are the nine components of the crystal orientation matrix $[A]$ (2). The resultant unit cell is then reduced and analyzed in terms of the 44 lattice types (Burzlaff *et al.*, 1992).

## 6. The effects of errors on indexing

The components of the basis vectors parallel to the X-ray beam are necessarily rather inaccurate when applying any autoindexing procedure. This is because the usual flat detector records data only in a forward direction and because the normal oscillation angle is small, resulting in a lack of information about the extent of the reciprocal lattice along the X-ray beam. Thus, it would be an advantage to combine images of one crystal taken at different rotation angles or, best, separated by a 90° rotation. In principle, this is not difficult, as the vectors **x** from different orientations of the crystal can be combined with different oscillation angles $\delta\varphi$ using (1). However, in practice, the errors in the values of camera parameters used for calculating the positions **x** and the assumption that the crystal is stationary for any given image introduce errors into the calculation of the position **x** for widely separated images.

An attempt was made to combine the reciprocal-lattice vectors derived from three separate images, taken at $\varphi$ = 0, 14.8 and 37.8°, recorded on a CCD detector using a frozen human rhinovirus 16 (HRV16) crystal at beamline SBC-19ID at the Advanced Photon Source (Argonne National Laboratory, Chicago). Each image was indexed successfully when analyzed by itself. However, on combining the information from the three images, the FFT systematically determined an $[A]$ matrix for one of the images which contained about 30% more useful reflections than the other two images. This

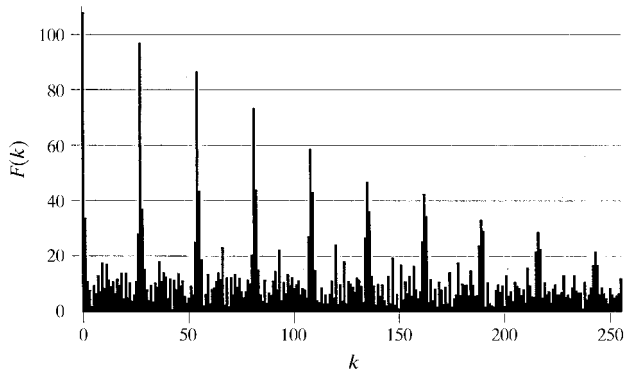


**Figure 1**
Let a line **t** in reciprocal space be divided into discrete intervals $\Delta t$ at a distance $t$ from the origin. Then, let the frequency of reflections at $(x, y, z)$ projected onto this line be given by the function $f(t)$ in an interval of $\Delta t$. The one-dimensional Fourier transform of this line will be given by $F(k) = \sum_{t=0}^{t=m\Delta t} f(t)\exp(2\pi ikt)$, where $m$ is an integer. In the figure, the largest Fourier coefficient other than $F(0)$ corresponds to $k$ = 27 and measures the distance between reciprocal-lattice planes perpendicular to the line of projection. (Reprinted with permission from Steller *et al.*, 1997.)

showed that the FFT found the dominant periodicity and that the positions of the reciprocal-lattice points for the other images did not mesh precisely with those of the dominant image on account of inaccurate camera parameters. Although unsuccessful for the purpose initially proposed, this result is particularly interesting as it shows that split crystals containing a dominant fragment would be readily indexable with the autoindexing procedure described here. Omission of the indexed reflections would then allow indexing of the minor component of the crystal.

## 7. Scaling and post-refinement – introduction

A least-squares procedure frequently used for scaling frames of data which contain a substantial number of 'full' reflections is the Hamilton, Rollett and Sparks (HRS) method (Hamilton *et al.*, 1965). The target for this least-squares minimization is

$$\Psi = \sum_h \sum_i W_{h_i}(I_{h_i} - G_m I_h)^2, \tag{5}$$

where $I_h$ is the best least-squares estimate of the intensity of a reflection with reduced Miller indices $h$, $I_{h_i}$ is the intensity of the $i$th measurement of reflection $h$, $W_{h_i}$ is a weight for reflection $h_i$ and $G_m$ is the inverse linear scale factor for the frame $m$ on which reflection $h_i$ is recorded. The HRS expression (5) assumes that all reflections $h_i$ are full; that is, their reciprocal-lattice points have completely passed through the Ewald sphere.

For all $h$, the values of $I_h$ must correspond to a minimum in $\Psi$. Thus,

$$(\partial\Psi/\partial I_h) = 0. \tag{6}$$

Therefore, $I_h$ is given by

$$I_h = \left(\sum_i W_{h_i} G_m I_{h_i}\right) \Big/ \left(\sum_i W_{h_i} G_m^2\right). \tag{7}$$

Since $\Psi$ is not linear with respect to the scale factors $G_m$, the values of the scale factors have to be determined by an iterative non-linear least-squares procedure. As the scale factors are relative to each other, the HRS procedure requires that one of them is arbitrarily fixed. If there are frames which have too few or no common reflections with any other frames, the normal equations matrix will be singular.

An improved method of solving the HRS normal least-squares equations is described by Fox & Holmes (1966). Their approach is based on the singular-value decomposition of the normal equations matrix. Apart from an accelerated convergence of the least-squares procedure, the advantage of the Fox and Holmes method is that no *ad hoc* decision needs to be made as to which scale factor should be fixed. Furthermore, 'troublesome' frames of data can be identified as causing negligibly small eigenvalues in the normal equations matrix.

**Table 1**
Scaling and post-refinement parameters.

| Parameter | Method 1 | Method 2 |
|---|---|---|
| Scale factors | Yes | Yes |
| Temperature factors | Yes | Yes |
| Crystal orientation | No | Yes |
| Effective mosaicity | No | Yes |

## 8. Generalization of the Hamilton, Rollett and Sparks equations to take into account partial reflections

In general, a Bragg reflection will occur on a number of consecutive frames as a series of partial reflections, and the full intensity can only be estimated from the measured intensities of the partial reflections. Let $I_{h_{im}}$ represent the intensity contribution of reflection $h_i$ recorded on frame $m$. If all the parts of reflection $h_i$ are available in the data set, then

$$I_{h_i} = \sum_m (I_{h_{im}}/G_m). \tag{8}$$

In practice, there will always occur reflections which do not have all their parts available. In such cases, the only way to estimate the full intensity of a reflection is to apply an estimated value of partiality to the measured intensities of available partial reflections.

Various models have been proposed in the literature to calculate the reflection partiality. In this study, we use Rossmann's model (Rossmann, 1979; Rossmann *et al.*, 1979) with Greenhough and Helliwell's correction (Greenhough & Helliwell, 1982). This model treats partiality as a fraction of a spherical volume swept through a nest of Ewald spheres. The coordinates of the spherical volume are defined by the Miller indices of the reflection, crystal orientation matrix and rotation angle. The divergence of the Ewald spheres accounts for the crystal mosaicity. Alternative geometrical descriptions of the reciprocal-lattice point passing through the nest of Ewald spheres have been given by Winkler *et al.*(1979), Greenhough & Helliwell (1982) and Bolotovsky & Coppens (1997).

Provided that the reflection partiality $p_{h_{im}}$ is known, the full intensity is estimated by

$$I_{h_i} = I_{h_{im}}/p_{h_{im}} G_m. \tag{9}$$

(9) can produce as many estimates of $I_{h_i}$ as there are parts of reflection $h_i$, while (8) produces only one estimate of $I_{h_i}$ from all parts of reflection $h_i$. Having defined the relationships between measured intensities of partial reflections and estimated full intensities by expressions (8) and (9), two methods of generalizing the HRS equations can be considered.

### 8.1. Method 1

If a reflection $h_i$ occurs on a number of consecutive frames and all intensity parts $I_{h_{im}}$ are available in the data set, the generalized HRS target equation takes the form

**Table 2**
Hierarchy of criteria for selecting reflections for the scaling and averaging procedures.

In methods 1 and 2, reject all parts of a reflection which has
(i) No successfully integrated parts
(ii) No parts with significant intensity (for scaling procedure only)
(iii) Some parts entering and some parts exiting the Ewald sphere (this condition implies that the reflection is too close to the rotation axis and is partly in the blind zone)

In method 1, reject all parts of a reflection which has
(i) any part which is not successfully integrated,
(ii) any part which has a significant intensity, but is not predicted by the scaling program based on the crystal mosaicity and orientation matrix,
(iii) the sum of calculated partialities different from unity by more than a user-chosen value,
(iv) A redundancy of 1.

In method 2, reject a part of a reflection if
(i) the calculated partiality is less than a user-chosen value,
(ii) the intensity is insignificant,
(iii) the calculated partiality is 1 and the redundancy is 1.

$$\Psi = \sum_h \sum_i \sum_m W_{h_{im}} \left\{ I_{h_{im}} - G_m \left[ I_h - \sum_{m' \neq m} (I_{h_{im'}}/G_{m'}) \right] \right\}^2 . \quad (10)$$

Using (6), the best least-squares estimate of $I_h$ will be

$$I_h = \frac{\sum_i I_{h_i} \sum_m W_{h_{im}} G_m^2}{\sum_i \sum_m W_{h_{im}} G_m^2} . \quad (11)$$

### 8.2. Method 2

If the theoretical partiality $p_{h_{im}}$ of partial reflections $h_{im}$ can be estimated, the generalized HRS target equation takes the form

$$\Psi = \sum_h \sum_i \sum_m W_{h_{im}} (I_{h_{im}} - G_m p_{h_{im}} I_h)^2 \quad (12)$$

and, using (6), the best least-squares estimate of $I_h$ becomes

$$I_h = \frac{\sum_i \sum_m W_{h_{im}} G_m p_{h_{im}} I_{h_{im}}}{\sum_i \sum_m W_{h_{im}} G_m^2 p_{h_{im}}^2} . \quad (13)$$

When all reflections in the data set are full, expressions (10) and (12), and (11) and (13), reduce to the 'classical' HRS expressions (5) and (7). Method 1 allows refinement of the scale factors only while method 2 allows refinement of the scale factors, crystal mosaicity and orientation matrix (Table 1), because the latter two factors contribute to the calculated partiality.

## 9. Selection of reflections useful for scaling

Method 1 requires that all parts of a reflection are available in order to incorporate that reflection into expression (10). Thus, reflections which occur at the beginning or end of the crystal rotation or at gaps within the rotation range must be rejected. Even when all the necessary parts of a reflection are recorded, at least one of these parts could have a problem during peak integration, thus making the rest of the reflection useless for scaling.

Method 2 allows the use of all reflections for scaling, because every observation of a partial reflection is sufficient to estimate the full reflection intensity by expression (9). However, the smaller the calculated partiality, the greater the error of the estimated full intensity. Therefore, a reasonable lower limit of calculated partiality has to be imposed in selecting partial reflections useful for scaling purposes.

Based on the above, the algorithm for selecting reflections is as follows.

(i) Sort all reflections in the data set according to (*a*) symmetry-reduced Miller indices, (*b*) original Miller indices, (*c*) oscillation range of the frame on which the reflection is recorded.

(ii) Reject some of the reflections according to criteria listed in Table 2.

## 10. Restraints and constraints

Scale factors will depend on intensity variations of the incident X-ray beam, variation of the developing conditions if films are used, crystal absorption and radiation damage. When using frozen crystals, scale factors will be mostly a measure of absorption variation as the crystal is rotated from frame to frame, although abrupt changes will occur when the intensity of the beam is changed, as occurs at the beginning of a new injection of electrons or positrons into the synchrotron ring (a 'fill'). Hence, in general, scale factors can be constrained to follow an analytical function or restrained [adding a term $w(G_n - G_{n+1})^2$ to $\psi$, where $G_n$ and $G_{n+1}$ are scale factors for the $n$th and $(n + 1)$th frame] to minimize variation between successive frames. Such procedures will increase $R_{\rm merge}$ because there are fewer parameters, but will increase the accuracy of the measured intensities as additional reasonable physical conditions have been applied.

The angular mis-setting angles of a single crystal should remain entirely constant. Thus, in principle, the refinement of mis-setting angles should constrain the mis-setting angles to be the same for all frames associated with a single crystal in the data set. However, in practice, independent refinement of these angles can indicate problems in the data sets when there are discontinuities in the plots of setting angle *versus* frame number.

Unit-cell dimensions can be reasonably assumed to be the same for all crystals and might, therefore, be constrained to be such. However, the exact conditions of freezing may cause some crystal-to-crystal variation.

Mosaicity is likely to increase as radiation damage proceeds. Thus, restraint between the independently refined mosaicities of neighboring frames can be useful.

## 11. Generalization of the procedure for averaging reflection intensities

Once the frame scale factors are determined, they need to be applied to reflection intensities and error estimates. The intensities of reflections with the same reduced Miller indices can then be averaged.

Two methods of intensity averaging may be considered based on the two different expressions (8) and (9) for the estimates of full intensities. For method 1, the intensity average is

$$\langle I_h \rangle = \frac{\sum_i I_{h_i} W_{h_i}}{\sum_i W_{h_i}} = \frac{\sum_i \left[ \sum_m (I_{h_{im}}/G_m) \right] W_{h_i}}{\sum_i W_{h_i}}. \qquad (14)$$

When method 2 is used for averaging, the determination of $\langle I_h \rangle$ is more complicated because there are as many estimates of the full intensity $I_{h_i}$ as there are partial reflections $h_{im}$. Therefore, intensity averaging for reflection $h$ has to be performed in two steps. Firstly, for every reflection $h_i$, the intensity estimates from all partial observations are averaged,

$$\langle I_{h_i} \rangle = \frac{\sum_m W_{h_{im}} [I_{h_{im}}/(G_m p_{h_{im}})]}{\sum_m W_{h_{im}}}, \qquad (15)$$

where the reciprocal variance weights are $W_{h_{im}} = G_m^2 p_{h_{im}}^2 / \sigma^2(I_{h_{im}})$. Secondly, the $\langle I_{h_i} \rangle$ values are averaged as

$$\langle I_h \rangle = \left( \sum_i W_{h_i} \langle I_{h_i} \rangle \right) / \sum_i W_{h_i}, \qquad (16)$$

where $W_{h_i} = 1/\sigma^2(\langle I_{h_i} \rangle)$ and $\sigma(\langle I_{h_i} \rangle)$ can be derived from (15).

While averaging estimated intensities of full reflections, special treatment has to be given to outliers and discordant pairs (Blessing, 1997). For samples of three or more equivalent reflections, it is necessary to consider the absolute values of the differences between individual intensities and the median of the sample, $|I_{h_i} - I_{\rm median}|$. The outliers can be detected by several statistical tests and can then be either down-weighted or rejected. When the sample consists of only two reflections, they can be considered as a 'discordant pair' if the difference between their intensities is not warranted by the estimated errors and, hence, both reflections can be rejected.

Averaging intensities by method 2 has an advantage over method 1 because outliers and discordant pairs can be

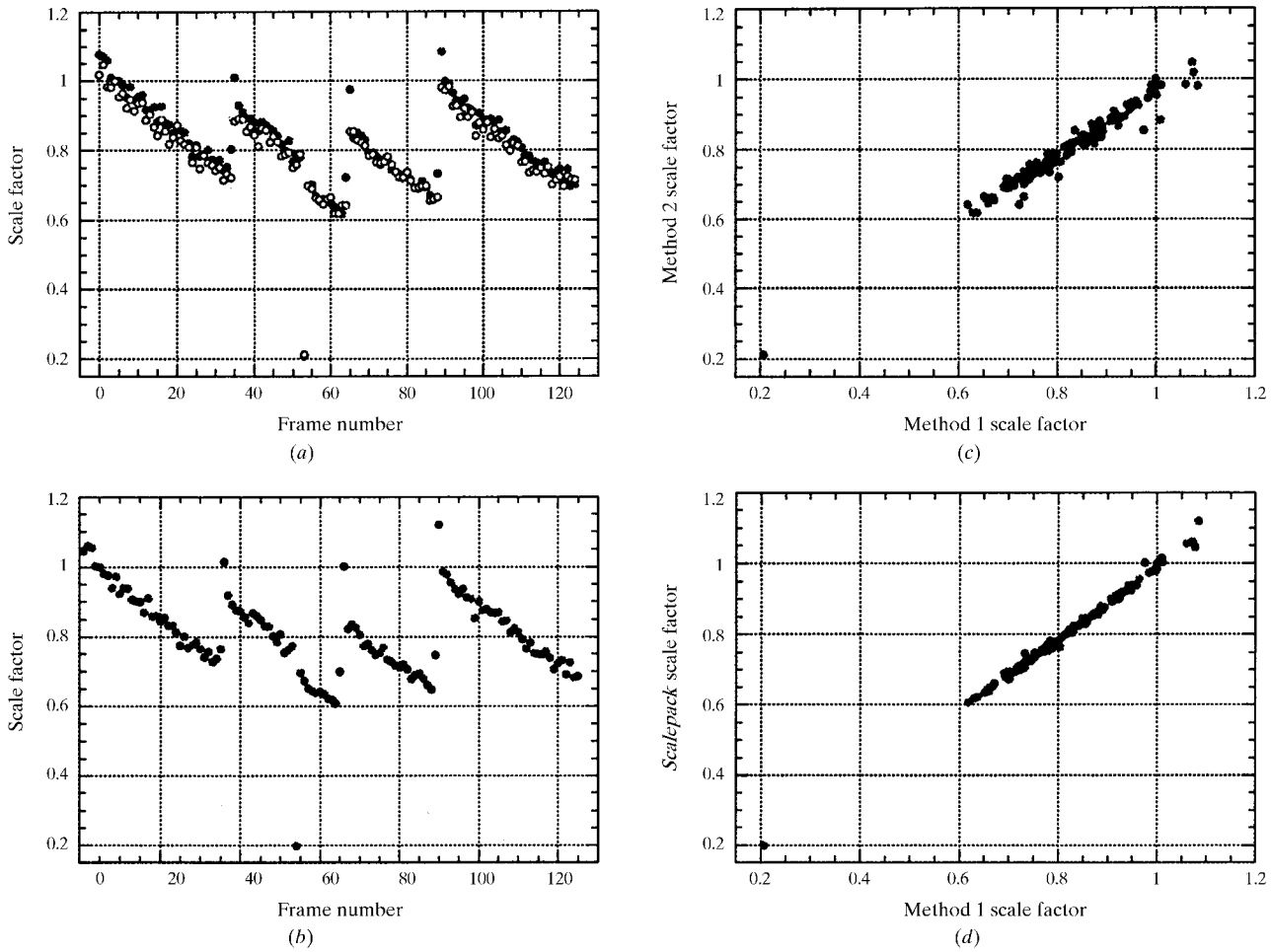


**Figure 2**
Unrestrained linear scale factor as a function of frame number of the φX174 procapsid data set. Results from (*a*) method 1 (filled circles) and method 2 (open circles) and (*b*) *SCALEPACK*. Comparison of (*c*) method 2 *versus* method 1 and (*d*) *SCALEPACK versus* method 1. (Reprinted with permission from Bolotovsky *et al.*, 1998.)

**Table 3**
Experimental information on the data sets processed by the methods described here and by *SCALEPACK*.

The data were integrated using the program *DENZO* (Gewirth, 1996; Otwinowski & Minor, 1997). The mosaicity reported by *DENZO* was used as an initial parameter for the scaling program.

| Data set | Compound name | Ref.† | Space group | Unit-cell parameters | | | | | | Mosaicity (°) | Oscillation range (°) | Total rotation (°) | Data collection information |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *a* (Å) | *b* (Å) | *c* (Å) | *α* (°) | *β* (°) | *γ* (°) | | | | |
| 1 | φX174 procapsid protein | (*a*) | $I2_13$ | 766.9 | 766.9 | 766.9 | 90.0 | 90.0 | 90.0 | 0.35 | 0.30 | 37.50 | CHESS, F1, Fuji IP, temperature = 120 K |
| 2 | Human rhinovirus 14 | (*b*) | $P2_13$ | 437.3 | 437.3 | 437.3 | 90.0 | 90.0 | 90.0 | 0.30 | 0.25 | 28.25 | CHESS, F1, Fuji IP, temperature = 120 K |
| 3 | Sindbis virus capsid protein (114–264) | (*c*) | *P*1 | 35.98 | 59.54 | 71.05 | 109.4 | 101.5 | 90.1 | 0.70 | 1.00 | 201.40 | Rigaku R-AXIS, temperature = 120 K |
| 4 | Alpha3 phage | (*d*) | $P2_1$ | 290.2 | 332.1 | 337.7 | 90.0 | 94.1 | 90.0 | 0.21–0.28 | 0.25 | 180.00 | APS, 14BMC, MAR 345 scanner, temperature = 120 K |

† References: (*a*) Dokland *et al.* (1997); (*b*) Rossmann *et al.* (1985); M. G. Rossmann, C. A. Momany, B. Cheng & S. Chakravarty, unpublished results; (*c*) Choi *et al.* (1991, 1996); (*d*) R. Bernal, B. A. Fane & M. G. Rossmann, unpublished results.

'screened' at two levels: firstly, when the estimates of full intensity $I_{h_i}$, calculated by (9) from different parts of the same reflection, are considered, and secondly, when the mean intensities $\langle I_{h_i} \rangle$, calculated by (15) from different reflections, are compared.

### 11.1. Scale factor *versus* frame number

If scale factors are to make physical sense, their behavior with respect to the frame number has to be in accordance with the known changes in the beam intensity, crystal condition and detector response. Conspicuous deviations from physically reasonable behavior may be attributed to deficiency of the scaling method.

The scaling of the φX174 procapsid data (data set 1 in Table 3) was performed using methods 1 and 2 described here and *SCALEPACK* (Gewirth, 1996; Otwinowski & Minor, 1997; Fig. 2). The graphs (*a*) and (*b*) in Fig. 2 have four segments corresponding to four synchrotron beam fills. All three methods give scale factors within 5% of each other. The only frames for which the results differ by as much as 15% are the first and last frames of each beam fill. Both method 1 and *SCALEPACK* produce physically wrong results in that the scale factors of these frames look like outliers compared with the scale factors of the neighboring frames. By contrast, method 2 provides consistent scale factors for such frames. Although the *SCALEPACK* algorithm for scaling frames with partial reflections has never been disclosed in the literature, the similar behavior of the results obtained by method 1 and *SCALEPACK* suggest that *SCALEPACK* might be using an algorithm similar to method 1.

Attempts at scaling a data set for a frozen crystal of HRV14 (data set 2 in Table 3) failed with method 1 because of gaps in the rotation range for the first 20 frames, causing singularity of the normal equations matrix. When frames without useful neighbors were excluded, the cubic symmetry of the crystal was sufficient for successful scaling. Method 2, however, did not have any problems with the whole data set, and its results showed greater consistency than those obtained with *SCALEPACK* (Fig. 3). *SCALEPACK* failed to refine the

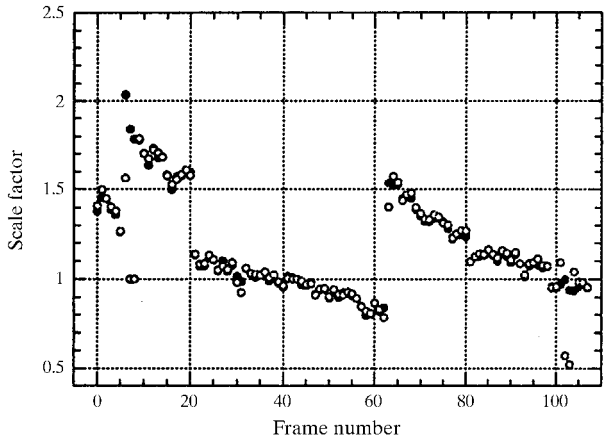


**Figure 3**
Linear scale factor as a function of frame number for the HRV14 data set using *SCALEPACK* (open circles) and method 2 (filled circles). (Reprinted with permission from Bolotovsky *et al.*, 1998).

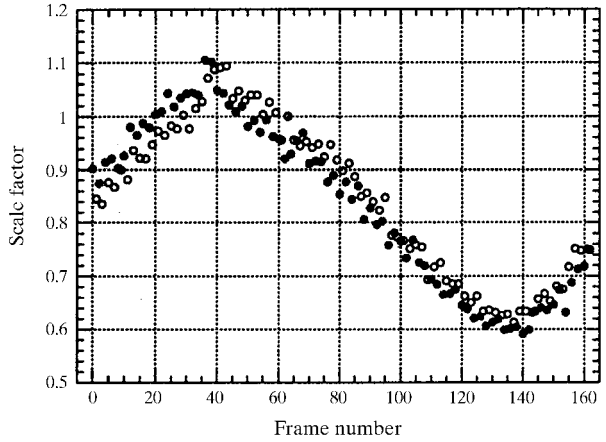


**Figure 4**
Unrestrained linear scale factors, determined by method 2, as a function of even (filled circles) and odd (open circles) frame numbers for the SCP (114–264) data set. The sine-like pattern reflects the anisotropy of a thin plate-shaped crystal. (Reprinted with permission from Bolotovsky *et al.*, 1998.)

scale factors of those frames which did not have a full complement of abutting frames. Their scale factors remained at the initial value of 1. Also, there are other frames for which the scale factors found by *SCALEPACK* look like outliers compared with the scale factors of the neighboring frames.

The accuracy of method 2 is also demonstrated by the scaling results for the Sindbis virus capsid protein (SCP), residues 114–264 (data set 3 in Table 3). The behavior of the scale factor with respect to the frame number reflects the anisotropy of a thin plate-shaped crystal (Fig. 4). For the first 38 frames (numbers 3–40), odd-numbered frames have higher scale factors than even-numbered frames. Data collection was stopped after frame number 40 and restarted. After frame number 41, odd-numbered frames have lower scale factors than even-numbered frames. This effect presumably relates to the use of the two alternative image plates with slightly different sensitivities in the R-AXIS camera.

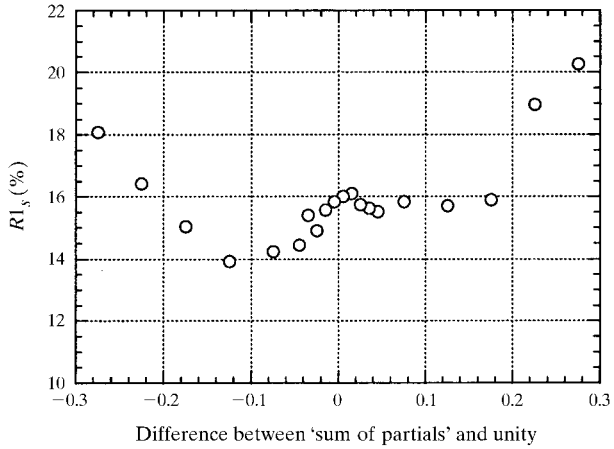


**Figure 5**
*R* factor as a function of the difference of calculated sum of partialities and unity for the estimates of full reflections when method 1 is used for scaling and averaging of the $\varphi$X174 procapsid data set. (Reprinted with permission from Bolotovsky *et al.*, 1998.)

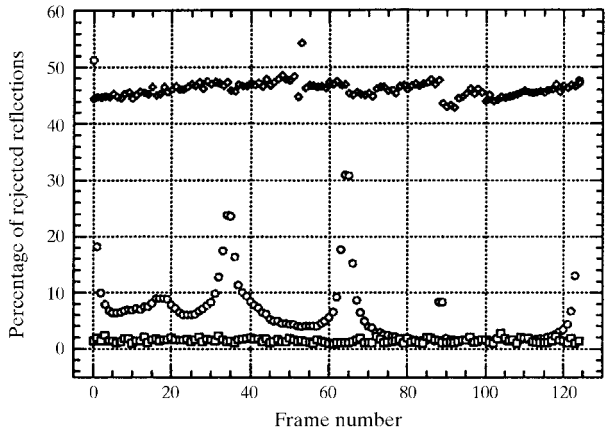


**Figure 6**
Percentage of rejected reflections for method 1 *versus* method 2 for the $\varphi$X174 procapsid data set. The reasons for rejecting reflections are listed in Table 2. (*a*) Open circles represent method 1; (*b*) open squares represent method 2 with mosaicity refinement; (*c*) open diamonds represent method 2 with no mosaicity refinement. (Reprinted with permission from Bolotovsky *et al.*, 1998.)

### 11.2. *R* factor as a function of 'sum of partialities' (method 1)

In order to determine the limits of tolerance which can be permitted when method 1 is used, the *R* factor was examined as a function of the sum of partialities for the $\varphi$X174 procapsid data (Fig. 5). For this evaluation, reflections with sum of partialities $1 \pm 0.3$ were used. The *R* factor changes sharply when the sum of partialities is outside $1 \pm 0.15$. Thus, $\pm 0.15$ were acceptable limits of tolerance for this data set.

### 11.3. Statistics for rejecting reflections and data quality as a function of frame number

The percentage of rejected reflections with respect to the frame number in method 2 is more monotonic than in method 1 (Fig. 6). In the latter method, the frames at the beginning and end of the crystal rotation and beam fills have an especially high rejection rate because there are insufficient data available to add up to full reflections (the reasons for rejecting reflections are listed in Table 2).

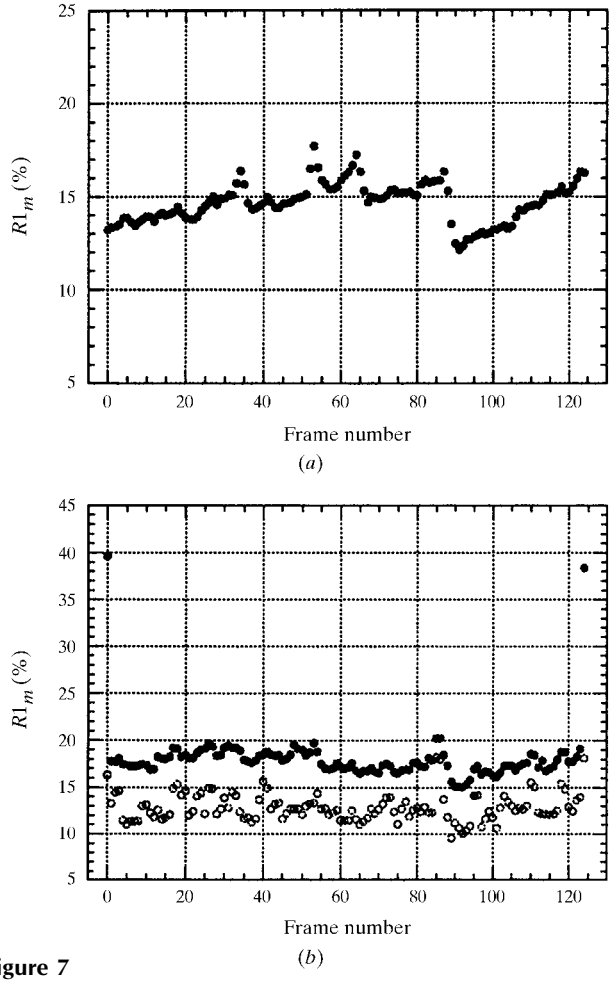


**Figure 7**
*R* factor as a function of the frame number for the $\varphi$X174 procapsid data set using (*a*) method 1 and (*b*) method 2 with no mosaicity refinement (filled circles) and method 2 with mosaicity refinement (open circles). (Reprinted with permission from Bolotovsky *et al.*, 1998.)

The behavior of the *R* factor *versus* frame number (Fig. 7) is more monotonic when method 1 is used compared with method 2. In method 1, the data quality estimates for neighboring frames are strongly correlated because the full reflections used in the statistics are obtained by summing up partials from consecutive frames. In contrast, in method 2 every frame produces estimates of full reflection intensities independently of the neighboring frames. Therefore, the frame *R* factors calculated after scaling with method 2 truly represent the data quality for individual frames.

### 11.4. Observed *versus* calculated partiality

The relationship between observed and calculated partialities (Fig. 8) deviates from the ideal line $p_{obs} = p_{calc}$, especially for the smaller calculated partialities where $p_{obs} > p_{calc}$. This suggests errors in measuring $p_{obs}$ or calculating $p_{calc}$. The latter may be improved by a post-refinement of the orientation matrix and crystal mosaicity (Rossmann *et al.*, 1979).

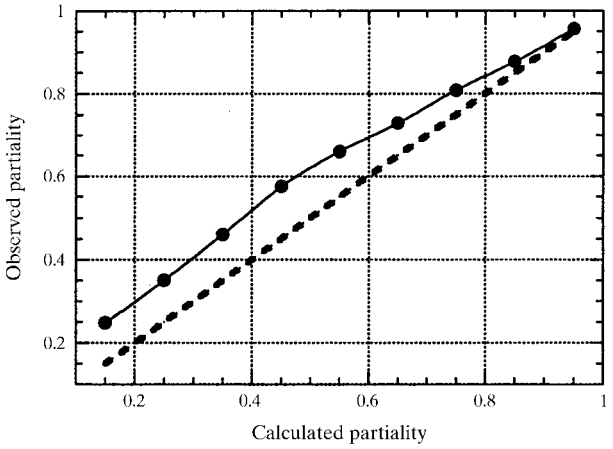


**Figure 8**
The observed partialities plotted against calculated partialities for the $\varphi$X174 procapsid data processed by method 2 with mosaicity refinement. The observed partialities for individual partial reflections were averaged in bins of calculated partialities. The broken line represents the ideal relationship $p_{obs} = p_{calc}$. (Reprinted with permission from Bolotovsky *et al.*, 1998.)

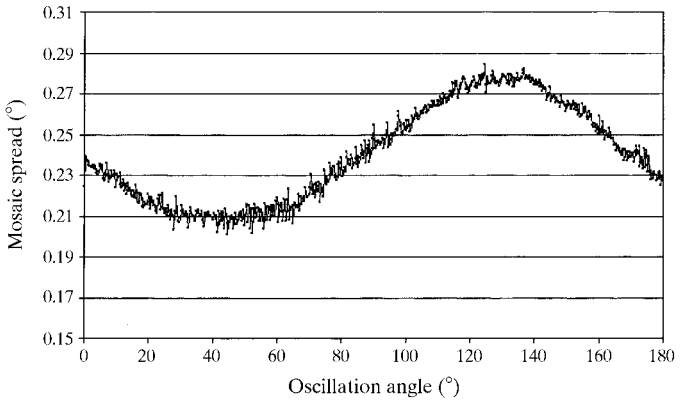


**Figure 9**
Variation of (unrestrained) mosaicity for a monoclinic crystal of the bacterial virus alpha3 showing the crystal anisotropy (data set 4 in Table 3) (Ricardo Bernal, April Burch, Bentley Fane and Michael Rossmann, unpublished data).

### 11.5. Anisotropic mosaicity

Restraint-independent refinement of mosaicity can show both the anisotropic nature of the crystal (Fig. 9) as well as the impact of radiation damage.

### 11.6. Anomalous scattering

The quality of anomalous dispersion data can be assessed by measuring the scatter $\sigma_{I_h}$ of measurements of non-centric reflections $I_h$ and comparing it with the scatter, $\sigma^+_{I_h}$ or $\sigma^-_{I_h}$, of reflections differing only in absorption while excluding Bijvoet opposites. Thus,

$$\langle\sigma_{I_h}\rangle = (1/h)\sum_h\{[1/(n-1)]\sum_n(I_h - I_{hn})^2\}^{1/2},$$

with corresponding definitions of $\sigma^+_{I_h}$ and $\sigma^-_{I_h}$. The ratios $\langle\sigma_{I_h}\rangle/\langle\sigma^+_{I_h}\rangle$ and $\langle\sigma_{I_h}\rangle/\langle\sigma^-_{I_h}\rangle$ should, therefore, be larger than unity for significant anomalous dispersion data (Fig. 10).

## 12. Availability of source code

The autoindexing program source code has been written in C, implemented on an SGI O2 workstation and is available *via*

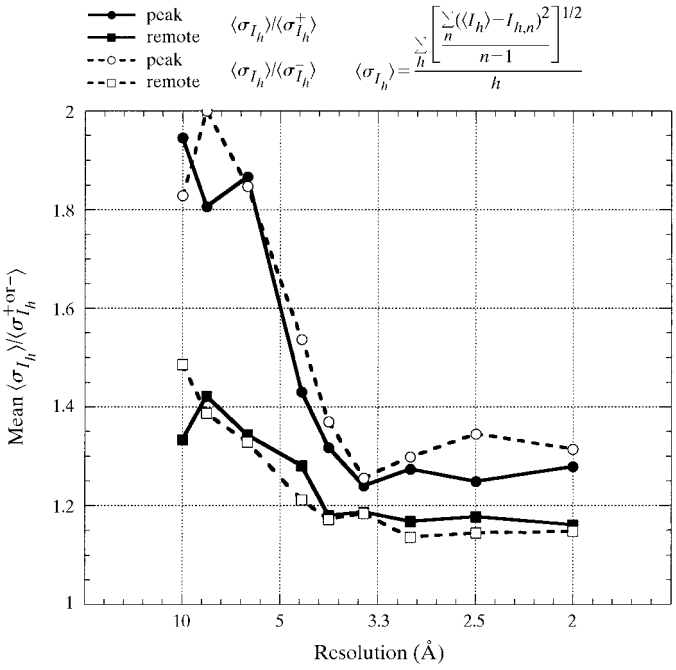


**Figure 10**
Quality of anomalous dispersion data for an SeMet derivative of a dioxygenase Rieske ferredoxin (Christopher Colbert and Jeffrey Bolin, unpublished data). Note the much larger scatter among measurements of $I_h$ for data measured at the absorption edge of Se (filled circles and empty circles) as opposed to measurements remote from the edge (filled squares and empty squares). The decreasing values of $\langle\sigma_{I_h}\rangle/\langle\sigma^+_{I_h}\rangle$ and of $\langle\sigma_{I_h}\rangle/\langle\sigma^-_{I_h}\rangle$ with resolution is a consequence of the decrease of $I_h$ values, thus causing the error in measurements of $I_h$ to approach the difference of intensity of Bijvoet opposites (measured by the inverse-beam procedure to eliminate absorption error).

the WWW at http://bilbo.bio.purdue.edu/~viruswww/Rossmann_home/rstest.html. The run time is sufficiently short for the autoindexing procedure to be run interactively.

The generalized procedure for scaling and averaging crystallographic data with partial reflections has been implemented as a C-language program *SNP* and tested on various data sets collected from crystals of biological macromolecules (Table 3). The source code is available *via* the WWW (http://bilbo.bio.purdue.edu/~viruswww/Rossmann_home/rstest.html).

This paper is largely based on two previous publications (Steller *et al.*, 1997; Bolotovsky *et al.*, 1998) concerning autoindexing and scaling and representing the work of Ingo Steller and Robert Bolotovsky, respectively, while postdoctoral fellows at Purdue University. We are very grateful for the support given to the development of *DPS* by Chris Nielson of ADSC and the staff of MacCHESS (including Steve Ealick, Dan Thiel and Marian Szebenyi) at Cornell University. Also, we would like to thank our colleagues at Purdue University and elsewhere who have provided many helpful suggestions. We are also anxious to acknowledge the outstanding help of Sharon Wilder in many parts of the work, including the preparation of this manuscript. This work was supported by a National Science Foundation grant (MCB-9527131) to MGR.

## References

Arndt, U. W., Champness, J. N., Phizackerley, R. P. & Wonacott, A. J. (1973). *J. Appl. Cryst.* **6**, 457–463.

Arndt, U. W. & Wonacott, A. J. (1977). *The Rotation Method in Crystallography.* Amsterdam: North-Holland.

Blessing, R. H. (1997). *J. Appl. Cryst.* **30**, 421–426.

Bolotovsky, R. & Coppens, P. (1997). *J. Appl. Cryst.* **30**, 65–70.

Bolotovsky, R., Steller, I. & Rossmann, M. G. (1998). *J. Appl. Cryst.* **31**, 708–717.

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

Burzlaff, H., Zimmermann, H. & de Wolff, P. M. (1992). *International Tables for Crystallography*, Vol. *A*, edited by T. Hahn, pp. 738–749. Dordrecht: Kluwer Academic Publishers.

Campbell, J. W. (1997). *CCP4 Newslett.* **33**, 5–16.

Choi, H. K., Lee, S., Zhang, Y. P., McKinney, B. R., Wengler, G., Rossmann, M. G. & Kuhn, R. J. (1996). *J. Mol. Biol.* **262**, 151–167.

Choi, H. K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M. G. & Wengler, G. (1991). *Nature (London)*, **354**, 37–43.

Dokland, T., McKenna, R., Ilag, L. L., Bowman, B. R., Incardona, N. L., Fane, B. A. & Rossmann, M. G. (1997). *Nature (London)*, **389**, 308–313.

Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92–96.

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

Gewirth, D. (1996). *The HKL Manual. A Description of the Programs DENZO, XDISPLAYF and SCALEPACK*, 5th ed., pp. 87–90. New Haven: Yale University.

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.

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

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

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

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

Leslie, A. G. W. (1992). *Crystallographic Computing 5. From Chemistry to Biology*, edited by D. Moras, A. D. Pojarny & J. C. Thierry. Oxford University Press.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

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

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Nature (London)*, **317**, 145–153.

Rossmann, M. G. & Erickson, J. W. (1983). *J. Appl. Cryst.* **16**, 629–636.

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

Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.

Vriend, G. & Rossmann, M. G. (1987). *J. Appl. Cryst.* **20**, 338–343.

Winkler, F. K., Schutt, C. E. & Harrison, S. C. (1979). *Acta Cryst.* A**35**, 901–911.

Xuong, N., Kraut, J., Seely, O., Freer, S. T. & Wright, C. S. (1968). *Acta Cryst.* B**24**, 289–290.