Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Simulation of X-ray frames from macromolecular crystals using a ray-tracing approach

Kay Diederichs

Fachbereich Biologie, Universität Konstanz, M647, D-78457 Konstanz, Germany

Correspondence e-mail: kay.diederichs@uni-konstanz.de

Received 6 February 2009
Accepted 19 March 2009

An algorithm is described which simulates a data set obtained from a protein crystal using the rotation method. The diffraction pattern of an ideal crystal is specified by the orientation of the crystal's cell axes with respect to a specified laboratory coordinate system, the distance between the crystal and the detector, the wavelength and the rotation range per frame. However, a realistic simulation of an experiment additionally requires at least a plausible physical model for crystal mosaicity and beam properties. To explore the physical basis of reflection shape and rocking-curve variation, the algorithm simulates the diffraction of a real crystal composed of mosaic blocks which is illuminated with a beam of given divergence and dispersion. Ray tracing for each reflection leads to reflection shapes and rocking curves that appear realistic. A program implementing the algorithm may be used to reproducibly generate data sets that model different physical aspects (imperfections) of the crystal and the experiment. Certain types of systematic errors of the experimental apparatus may also be simulated. Further applications include teaching and characterization of the properties of data-reduction algorithms.

1. Introduction

Structural biology, and macromolecular X-ray crystallography in particular, has progressed significantly from advances in protein expression and purification, large-scale crystallization and rapid data collection, but also in computational analysis of the data. Whereas almost all computational steps in crystallography, such as phasing and refinement, can be cross-checked against results obtained using ideal data, this is not currently possible for the first and fundamental step of data analysis (called 'data reduction'), which reduces the raw frames of a data set to a list of observed (h, k, l) triplets and their associated pairs of diffraction intensities and standard deviations. However, it would also be very helpful to know what the ideal program output should be for this class of programs. Synthetic data have known intensities and standard deviations and offer the advantage that extreme experimental situations can be simulated as easily as common ones. As in other fields of computational analysis of physical phenomena, synthetic data allow the proper operation of a data-reduction program to be verified for specific examples.

The geometry of an ideal diffraction experiment using the rotation method (Arndt & Wonacott, 1977) may be formu-

© 2009 International Union of Crystallography
Printed in Singapore – all rights reserved

Table 1
Keywords and parameters describing synthetic data sets.

The table lists those keywords and parameters that are specific to *SIM_MX*. A number of other keywords (NAME_TEMPLATE_OF_DATA_FRAMES, DATA_RANGE, OSCILLATION_RANGE, X-RAY_WAVELENGTH, DETECTOR, QX, QY, NX, NY, INCLUDE_RESOLUTION_RANGE, DETECTOR_DISTANCE, DIRECTION_OF_DETECTOR_X-AXIS, DIRECTION_OF_DETECTOR_Y-AXIS, ROTATION_AXIS, INCIDENT_BEAM_DIRECTION, ORGX, ORGY) are required to define the geometrical parameters of the experiment and the desired output; these parameters are defined as those documented for *XDS* at http://www.mpimf-heidelberg.mpg.de/~kabsch/xds/html_doc/xds_parameters.html.

Keyword	Default	Unit	Meaning
UNIT_CELL_A-AXIS	—	Å	Components of unit-cell *a* axis with respect to the laboratory coordinate system for the unrotated crystal. Required input.
UNIT_CELL_B-AXIS	—	Å	Same for unit-cell *b* axis. Required input.
UNIT_CELL_C-AXIS	—	Å	Same for unit-cell *c* axis. Required input.
TWIN_EXPOSURE	0.0	—	This factor determines the resulting twinning fraction if a second crystal (*e.g.* nonmerohedral twin) is in the beam.
TWIN_CELL_A-AXIS	—	Å	As UNIT_CELL_A-AXIS. Required input if TWIN_EXPOSURE > 0.
TWIN_CELL_B-AXIS	—	Å	As UNIT_CELL_B-AXIS. Required input if TWIN_EXPOSURE > 0.
TWIN_CELL_C-AXIS	—	Å	As UNIT_CELL_C-AXIS. Required input if TWIN_EXPOSURE > 0.
GAIN	1.0	—	Factor to convert X-ray photons to pixel contents
EXPOSURE_FACTOR	1.0	—	Factor to multiply input intensities by (*e.g.* to simulate larger crystal or stronger beam)
BACKGROUND	100.0	Counts	Background X-ray photons per pixel
BIG_CRYSTAL	TRUE	—	FALSE for infinitesimally small reflections (affecting a single pixel). TRUE distributes counts to four adjacent pixels proportional to distance.
WAVELENGTH_STDDEV	0.0002	Å	Standard deviation of wavelength
BEAM_STDDEV	0.02, 0.02	deg^2	Standard deviation of rotations of primary beam around two directions perpendicular to its direction
ORIENTATION_STDDEV	0.1, 0.1, 0.1	deg^3	Standard deviation of rotations of mosaic crystals around three orthogonal axes
CELL_STDDEV	0.1, 0.1, 0.1	$Å^3$	Standard deviation of unit-cell axes
FLAT_WAVELENGTH	FALSE	—	Choice of Gaussian (FALSE) or top-hat (TRUE) distribution of wavelength
FLAT_BEAM	FALSE	—	Choice of Gaussian (FALSE) or top-hat (TRUE) distribution of primary beam rotations
FLAT_ORIENTATION	FALSE	—	Choice of Gaussian (FALSE) or top-hat (TRUE) distribution of mosaic crystal rotations
FLAT_CELL	FALSE	—	Choice of Gaussian (FALSE) or top-hat (TRUE) distribution of cell parameters
MODULATION_IN_PHI	0, 0, 0	—	If *a*, *b*, *c* are the parameters, the counts of each ray are multiplied by $1 + a\sin(b\varphi + c)$
SENSITIVITY_IN_PHI_SUBRANGE	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1	—	Sensitivity factors for 20 φ ranges within each frame

lated with respect to a general right-handed orthonormal laboratory coordinate system (Kabsch, 2006*a*,*b*). However, the shape of real reflections on the detector and their intensity distribution over several frames (the 'rocking curve') is not easily calculated even for a simple isotropic model for the intensity distribution of a reflection in reciprocal space. This is a consequence of the fact that in the rotation method the path of each reflection through the Ewald sphere is different, which translates into different reflection shapes.

'Two-dimensional' data-reduction programs such as *MOSFLM* (Leslie, 1992*a*,*b*) and *DENZO* (Otwinowski & Minor, 1997) take variation in reflection shape into account by adapting the shape of an integration box and by estimating a parameter describing the rocking curve. 'Three-dimensional' data-reduction programs such as *XDS* (Kabsch, 1988, 1993, 2006*a*,*b*) and *d*TREK* (Pflugrath, 1999) build up an internal three-dimensional representation of each reflection's profile. This representation is obtained through a mathematical transformation developed by Kabsch (1988), which results in reflection profiles that are more uniform across the detector surface and directly related to the intensity distribution in reciprocal space.

The physical parameters affecting the observed shape and rocking curve of a reflection are the size, shape and mosaicity (as defined below) of the crystal, the size, wavelength dispersion and divergence of the beam, the pixel size and point-spread function of the detector, the position of the reflection relative to the geometry of the experiment (determining the angle of incidence of X-rays and the Lorentz and polarization factors) and the amount of rotation around the spindle. The specific influence of some of these effects on the reflection shapes and reflection widths may be simulated, as is shown below.

Ray tracing has been used in computer graphics as a technique for generating a realistic image by tracing the path of light through pixels in an image plane. In crystallography, ray tracing has been used, for example, for an analysis of X-ray optics (Yamada *et al.*, 2001; Artemiev *et al.*, 2004) and to

simulate powder diffraction line profiles (Lambert & Guillet, 2008). However, to the best of the author's knowledge, ray tracing has not so far been used to simulate diffraction data from crystals of macromolecules.

Computational visualization of ideal diffraction patterns has been achieved with *XRayView* (Phillips, 1995) and *ROTGEN* (Campbell, 1996) and simulation of data affected by diffuse scattering has been performed in *XCADS* (Kolatkar *et al.*, 1994). However, with the exception of *MLFSOM* (Holton, 2008*a*), which employs a user-supplied model for reflection shape and rocking curve (Holton, 2008*b*), there is no currently available program that calculates diffraction data from intensities, thus reversing the calculations performed in data reduction.

2. Materials and methods

A well known concept for describing a real crystal, as opposed to an ideal crystal, is the notion of a 'mosaic crystal', which considers the real crystal to be composed of a large number of ideal crystals ('mosaic blocks').

The mosaic blocks not only differ slightly in their orientation. As suggested and discussed by Nave (1998) and independently found in this work, the mosaic blocks must also be allowed to have unit-cell parameters that slightly differ from the average in order to arrive at realistic reflection shapes. Together, these two kinds of deviations from the properties of an ideal crystal, for the purposes of this work, constitute the 'mosaicity' of a crystal.

In addition, an algorithm for the realistic simulation of diffraction patterns has to consider the primary beam to be composed of individual beams sampling a distribution of wavelengths and directions.

2.1. Implementation

A program (called *SIM_MX*) has been written in standard Fortran95. It was developed under Linux and requires at least version 0.8.1 of CBFlib (Bernstein & Hammersley, 2005; Bernstein & Ellis, 2005), which is available from http://sourceforge.net/projects/cbflib. As the complete data set is synthesized in the memory of the computer, a large amount of memory (several gigabytes) and, for some data sets, a 64-bit operating system may be required. However, it is possible to divide up a large data set into several frame ranges. In this way, the large data set may be stitched together at the expense of more CPU time.

2.2. Input to the program

The program *SIM_MX* reads a control file `SIMULATE.INP` that consists of lines describing the experiment (Table 1). In addition, a free-format file `intensities.hkl` with (h, k, l) triplets and intensity values is read. These intensities may, for example, be calculated values corresponding to a known protein structure or measured values from a data-reduction program. As *SIM_MX* has no built-in space-group library, no symmetry expansion is performed, nor is Friedel symmetry imposed. All (h, k, l) triplets for which reflections within the synthetic data set should be generated have to be supplied. Thus, effects such as radiation damage (which breaks space-group symmetry) and anomalous scattering may be simulated by suitable user manipulation of the intensity values that the program reads.

As a simple safeguard against input errors, the program prints out a list of reflections missing from a full sphere in reciprocal space.

2.3. Operation of the program

In the context of the program, ray tracing means that a large number of X-rays, each generated by a primary beam with a specific direction and diffracted by a specific mosaic block, together constitute the total diffraction pattern of the crystal. Each mosaic block contributes one ray for each reflection of the diffraction pattern.

Computationally, this is achieved by sampling a nine-dimensional distribution describing variation of mosaic block orientation (three-dimensional), variation of mosaic block cell axes (three-dimensional; angles are fixed), variation of beam direction (two-dimensional) and variation of wavelength (one-dimensional). The exact diffraction geometry for each ray is calculated and its intensity contribution is added to a storage location representing a pixel on a frame which covers the rotation angle determined for this ray. The intensity contribution is given by the intensity of the respective reflection (as read from `intensities.hkl`) divided by the number of rays and divided by the Lorentz factor of this ray.

For each physical parameter considered, a Gaussian distribution function with given standard deviation is employed by default. Alternatively, a flat 'top-hat' distribution may be chosen.

The typical dimension of mosaic blocks of protein crystals may be assumed to be of the order of 1 µm. Owing to this finite size, the reflections cannot be considered as infinitesimally sharp. This effect may be taken into account by distributing the counts of a ray to the four nearest pixels instead of adding them to the contents of a single pixel. The effect of the corresponding keyword (BIG_CRYSTAL; see Table 1) may also be considered to crudely account for the effect of a finite-sized crystal and a finite-sized beam.

The diffraction of a second crystal with a different orientation matrix (twin) may be added by the program with a weight corresponding to the desired twinning fraction.

The program reports the number of pixels contributing to each reflection. The number of rays traced should exceed the number of pixels by about an order of magnitude. If the number of rays traced is lower, it may be increased by using a command-line option at the expense of increasing CPU time. The default number is 1000 rays, which is ample for a realistic simulation of low-mosaicity crystals.

2.4. Simulation of statistical and systematic errors

Using a pseudo-random number generator (see below), counts in each pixel are obtained by sampling a Poisson

distribution (with a mean as accumulated by the ray-tracing computation), thus producing the appropriate counting statistics.

Different kinds of systematic errors may be simulated by modifying the pixel contents while they are computed. This is possible because each reflection results from contributions that are exactly characterized with respect to position on the detector and rotation angle.

For example, the Pilatus detector (Hülsen *et al.*, 2006) may be run in a quasi-continuous data-acquisition mode that uses a short time period during the collection of each frame for readout. As during this time the shutter is not closed and the spindle movement is not stopped, all X-rays reaching the detector during this timeframe are not actually stored on the resulting frames. Such a detector property may be simulated in *SIM_MX* using the SENSITIVITY_IN_PHI_SUBRANGE keyword.

A sinusoidal modulation of the conversion factor from photons to counts, with arbitrary period, may be obtained with the keyword MODULATION_IN_PHI that accepts three parameters (Table 1). This can be used to simulate the effects of, for example, mechanically or electrically induced vibrations or sensitivity changes, primary beam-intensity changes and of changes in the irradiated crystal volume during rotation of the crystal.

2.5. Output of the program

The program generates a data set consisting of the desired number of frames in either (i) a compressed format (Abrahams, 1993) with 1200 or 2000 pixels in x and y (filenames ending in `.pck`), (ii) an SMV format with a 512-byte header and a number of pixels in x and y that is a multiple of 512 (filenames ending in `.img`) that is compatible with software that can read frames from ADSC detectors or (iii) a self-documenting compressed CBF format (filenames ending in `.cbf`) available through the latest version of CBFlib (H. J. Bernstein, personal communication).

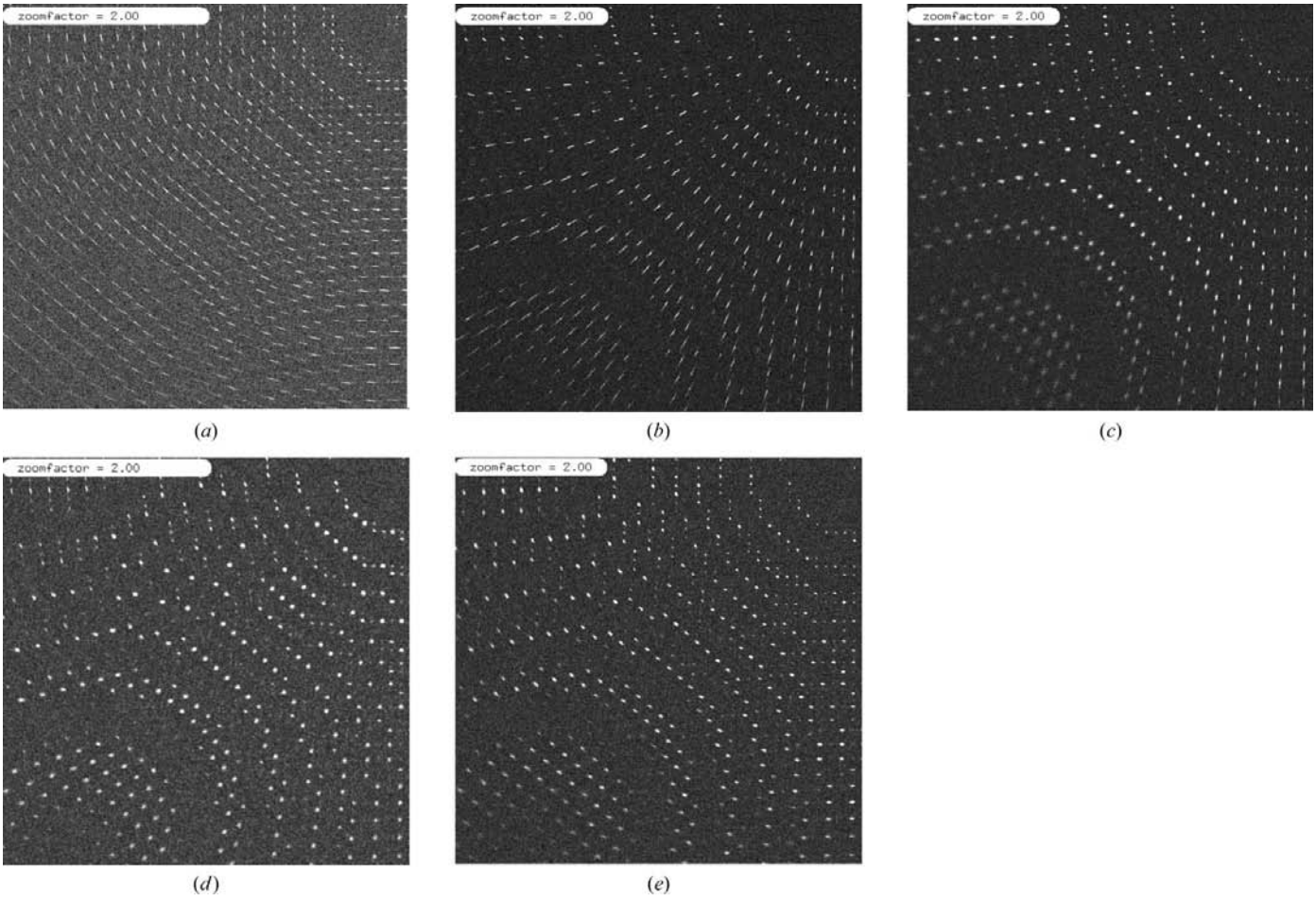

Figure 1
Diffraction patterns showing the influence of the parameters WAVELENGTH_STDDEV, BEAM_STDDEV, ORIENTATION_STDDEV and CELL_STDDEV (see Table 1) on reflection shapes. Further parameters are given in the text. The direct-beam position is in the upper right corner. The rotation axis is parallel to the upper rim of each plot. The figures were produced with the *VIEW* program of the *XDS* package. (*a*) WAVELENGTH_STDDEV = 0, BEAM_STDDEV = 0 0, ORIENTATION_STDDEV = 0.5 0.5 0.5, CELL_STDDEV = 0 0 0. (*b*) WAVELENGTH_STDDEV = 0.01, BEAM_STDDEV = 0 0, ORIENTATION_STDDEV = 0 0 0, CELL_STDDEV = 0 0 0. (*c*) WAVELENGTH_STDDEV = 0, BEAM_STDDEV = 0 0, ORIENTATION_STDDEV = 0 0 0, CELL_STDDEV = 0.4 0.4 0.4. (*d*) WAVELENGTH_STDDEV = 0, BEAM_STDDEV = 0.1 0.1, ORIENTATION_STDDEV = 0 0 0, CELL_STDDEV = 0 0 0. (*e*) WAVELENGTH_STDDEV = 0.001, BEAM_STDDEV = 0.02 0.02, ORIENTATION_STDDEV = 0.2 0.2 0.2, CELL_STDDEV = 0.1 0.1 0.1.

Table 2
Parameters used for the simulation of the insulin data set.

Intensities were the squared structure factors of PDB entry 2bn3.

```
DATA_RANGE = 1 32
OSCILLATION_RANGE = 1
X-RAY_WAVELENGTH = 1
DETECTOR_DISTANCE = 100.0
DETECTOR = MAR345
QX = 0.15 QY = 0.15 NX = 1200 NY = 1200
DIRECTION_OF_DETECTOR_X-AXIS = 1.0 0.0 0.0
DIRECTION_OF_DETECTOR_Y-AXIS = 0.0 1.0 0.0
ROTATION_AXIS = 1.0 0.0 0.0
INCIDENT_BEAM_DIRECTION = 0.0 0.0 1.0
ORGX = 600 ORGY = 600
UNIT_CELL_A-AXIS = −9.538214 45.736636 62.334703
UNIT_CELL_B-AXIS = −20.418634 −62.067525 42.416207
UNIT_CELL_C-AXIS = 74.568985 −11.145223 19.587798
GAIN = 1
EXPOSURE_FACTOR = 0.01
BACKGROUND = 10
BIG_CRYSTAL = TRUE
WAVELENGTH_STDDEV = 0.0002
BEAM_STDDEV = 0.02 0.02
ORIENTATION_STDDEV = 0.1 0.1 0.1
CELL_STDDEV = 0.1 0.1 0.1
```

By default, the intensity contributions belonging to each reflection are summed up by the program. These 'summed intensities' are a byproduct of the book-keeping required to trace all rays diffracted by the crystal and reaching the detector. For this summation, pixels obtaining contributions from more than one reflection are discarded. Depending on the unit-cell parameters, mosaicity and resolution, an overlap situation may occur which results in 'partial' reflections for which only a fraction of the contributing intensity is available for summation. Only those reflections with a partiality of more than 75% are scaled up with the inverse partiality and their intensities are written to an output file `summed_intensities.hkl`.

Based on these optimally summed intensity values, statistics [R factors and $I/\sigma(I)$ values, both as a function of resolution] are calculated that would be obtained from a data-reduction program that accurately sums all pixels contributing to the reflections.

2.6. Example calculations: exploring the physical model

The simulations displayed in Fig. 1 were obtained for a crystal with unit-cell parameters a = 70, b = 80, c = 90 Å, $\alpha = \beta = \gamma = 90^\circ$ (space group $P2_12_12_1$). The orientation of the unrotated crystal ($\varphi = 0^\circ$) was with its a axis along the detector's x axis (horizontal in Fig. 1), its b axis along the detector's y axis and its c axis perpendicular to a and b. The simulation assumed a background of 50 counts, a GAIN and EXPOSURE_FACTOR of 1, a wavelength of 1.4 Å, a crystal-to-detector distance of 100 mm and a rotation interval from 0° to 1°, with the rotation axis parallel to the detector's x axis.

Further details of the individual runs of *SIM_MX* are given in the legend of Fig. 1.

2.7. Example calculation: a complete data set

To compare the statistics from data obtained by book-keeping within *SIM_MX* with those from a data-reduction program, structure factors from an insulin crystal (PDB code 2bn3; Nanao *et al.*, 2005) were used. After downloading these data from the PDB in mmCIF format, they were converted to MTZ format using the *CCP*4 program *CIF*2*MTZ* (Collaborative Computational Project, Number 4, 1994), expanded to *P*1 using the *CCP*4 program *SFTOOLS* and then further expanded to the full sphere (as required by *SIM_MX*) using a custom program. These (h, k, l) triplets with the squared structure factors of entry 2bn3 (in the file `intensities.hkl`) were then used for a *SIM_MX* run (for a complete list of simulation parameters, see Table 2). The frames written by *SIM_MX* were reduced with the *XDS* program (Kabsch, 1988, 1993, 2006*a*,*b*) using default parameters. Internal quality indicators [R factors and $I/\sigma(I)$] were obtained from *XSCALE* (*XDS* package). The intensities in the output files `summed_intensities.hkl` and `XDS_ASCII.HKL` were also compared with the input (in this case, experimental) intensities of entry 2bn3 using a custom program.

3. Results and discussion

3.1. Reproducibility

Care has been taken to ensure that the results of the program, both in terms of pixel contents (which include statistical noise) and summary output, are well reproducible across operating systems and computer hardware. Determinism of sampling of parameter distributions and statistical noise required the use of a thread-safe multiplicative congruential random-number generator (L'Ecuyer, 1999) with known good properties instead of a system-supplied one. The reproducibility of the program output is meant to enable users and authors of data-reduction programs to exchange just the parameters describing a synthetic data set, instead of the data set itself.

However, strict reproducibility was not achieved as the program uses floating-point arithmetic which implies rounding effects depending on the specifics of the compilers and their options.

3.2. Reflection shapes and rocking curves

Examples of the effects of the major factors affecting reflection shape are shown in Figs. 1(*a*)–1(*e*). Reflection shapes generally appear realistic. It is evident that all factors have a strong influence and need to be taken into account. Furthermore, as is shown below, their effects on reflection shape are different.

One contribution to the observed reflection shape and mosaicity (as defined above) is implemented as a rotation of the crystal around three mutually orthogonal axes ('rotational mosaicity'). Analogous to powder diagrams, crystals with given unit-cell parameters that are rotated with respect to each other generate reflections that are elongated along circles around the origin (Fig. 1*a*) of reciprocal space, with radii

corresponding to their *d*-spacing. Rotational mosaicity cannot therefore explain broad reflections that are often modelled as beam divergence by data-reduction programs. In contrast, the finite dispersion of the wavelength used in a 'monochromatic' experiment results in a radial streaking of reflections (Fig. 1*b*) orthogonal to the type of broadening resulting from rotational mosaicity.

Unit-cell parameter variation (Fig. 1*c*) broadens reflections, as discussed by Nave (1998). Likewise, beam divergence, together with the rotation of the crystal around the φ axis, produces broadening of the reflections (Fig. 1*d*). However, the details of this broadening effect are different: rotational mosaicity produces elongated reflections near the rotation axis, whereas unit-cell parameter variation results in elongated reflections at a right angle to the rotation axis.

It is clear that the physical parameters that affect the simulation likewise influence the reflection shape and the rocking curve. The question arises whether these parameters are suitable for treatment as variable parameters to be fitted to observed reflection shapes and rocking curves. In principle, this would allow a more accurate separation of background and reflection area during data processing, which usually relies on 'profiles', reflection shapes 'learnt' from strong reflections, and a parameterized model fitted to the observed rocking curves.

In practice, beam divergence and dispersion at a synchrotron site may be determined experimentally, for example by measurements from a near-ideal test crystal. The (anisotropic) unit-cell parameter variation and (anisotropic) rotational mosaicity of a given macromolecular non-ideal crystal may then be fitted to the data collected from this crystal, taking the three previously experimentally determined parameters into account. As the effects of the parameters determining mosaicity do not differ greatly, it appears likely that this cannot be performed routinely at the level of single frames, but may be feasible if a complete data set is available. However, this assumes that these six parameters are constant during the time that it takes to collect the data set, a requirement that may not be fulfilled in the presence of radiation damage.

Based on the above findings about the ways that the physical parameters influence the reflection shapes, a simplified approach that can be applied to single frames would involve fitting the parameters of ellipsoids. Contrary to the usual alignment (along *x* and *y*) of the axes of the integration boxes, the axes of these ellipsoids should be oriented radially and tangentially with respect to the origin of reciprocal space.

3.3. Influence of distribution shapes

Computational experiments with the keywords FLAT_WAVELENGTH, FLAT_BEAM, FLAT_ORIENTATION and FLAT_CELL were performed to investigate the role of the distribution of the physical factors that determine reflection shape and rocking curve. Visual inspection of the reflection shapes generated by the program, and statistics obtained from *SIM_MX* and from data-reduction programs, of simulated data sets revealed minor differences between data sets obtained with Gaussian and those obtained with top-hat distributions.

The rocking curves of individual reflections were inspected using the DEBUG keyword and by simulating a range of number of frames with a small rotation increment (0.01°). It was found that if only one physical parameter was different from zero the rocking-curve shape matched that of the distribution of that physical parameter.

However, when several physical parameters were varied (*i.e.* their standard deviations were nonzero), a convolution of distributions occurred, with the result that the rocking curves become closer to a Gaussian. This finding confirms that data integration using a Gaussian model for the rocking curve is a reasonable approach.

3.4. Processing of a simulated data set

The simulated frames based on the 2bn3 intensity data were processed with *XDS* to assess the correspondence of those parameters influencing reflection width and rocking curve and to compare statistical quantities computed in a *SIM_MX* run with those from a data-reduction program.

The simulation parameters determining reflection width and rocking curve are given in Table 2. The values of the parameters EXPOSURE_FACTOR and BACKGROUND were chosen to achieve a very good overall quality of the data set, with average $I/\sigma(I)$ in the 1.59–1.50 Å resolution shell comparable to the value given in the header of the PDB file for the original 2bn3 data reduction (Table 3), without saturating any pixels. The values of WAVELENGTH_STDDEV, BEAM_STDDEV, ORIENTATION_STDDEV and CELL_STDDEV are the default values of the program, which result in sharp reflections corresponding to a well ordered crystal.

According to Table 3, the statistical parameters derived for `summed_intensities.hkl` (written by *SIM_MX*) and for `XDS_ASCII.HKL` (written by *XDS*) are in good agreement. It may be noted that the internal quality indicators [$R_{\rm meas}$ and $I/\sigma(I)$] are better in case of the *XDS* data reduction, possibly because *XDS* employs profile fitting instead of straight summation. On the other hand, the data from *SIM_MX* agree better with the input data, as would be expected.

For the simulated data, *XDS* determined a value of 0.059° for its integration parameter BEAM_DIVERGENCE_E.S.D. (which combines the effects of beam and crystal) and a value of 0.156° for its integration parameter REFLECTING_RANGE_E.S.D. (mosaicity). Compared with the parameters used for the simulation (Table 2), the integration parameters found by *XDS* were in the same range. It was observed (data not shown) that these values were almost twice as high if WAVELENGTH_STDDEV, BEAM_STDDEV, ORIENTATION_STDDEV and CELL_STDDEV were doubled. Further comparison is difficult, as the physical model simulated by *SIM_MX* differs from the computational model used by *XDS*.

During this and other tests it was found that even small errors (*e.g.* omitting a slice of reciprocal space) in the

Table 3
Comparison of the crystallographic statistics of experimental data (PDB entry 2bn3) and those obtained for the files `summed_intensities.hkl` (written by *SIM_MX*) and `XDS_ASCII.HKL` (*XDS* data reduction).

The experimental data were reduced with *XDS* and *XSCALE* (*XDS* package); the values given are from the header of the PDB file. The values for *SIM_MX* and *XDS* were obtained by *XSCALE* after symmetry merging in $I2_13$ and the *R* factors against the intensities of 2bn3 were obtained by a custom program.

	2bn3	*SIM_MX*	*XDS*
Unit-cell parameters (Å)	$a = b = c = 77.90$	$a = b = c = 77.90$ (supplied)	$a = b = c = 77.89$
Resolution range† (Å)	50–1.5/NR/1.59–1.50	50–1.5/50–4.5/1.59–1.50	50–1.5/50–4.5/1.59–1.50
$R_{\rm meas}$†‡ (%)	4.0/NR/27	2.6/0.6/38.4	2.2/0.6/26.4
$\langle I/\sigma(I)\rangle$†	19.3/NR/4.98	46.0/182.0/4.68	47.7/186.0/5.82
Wilson *B* factor (Å^2)	NR	22.2	22.1
Multiplicity†	3.8/NR/3.9	3.9/3.9/3.7	3.8/3.6/3.8
R factor against intensities of 2bn3†§	—	1.3/0.3/18.4	2.3/1.6/15.3

† Values are given for overall/lowest shell/highest shell. NR, not reported. ‡ As defined in Diederichs & Karplus (1997). § $R = 2\sum|I_1 - I_2|/\sum(I_1 + I_2)$.

preparation of the input file `intensities.hkl` led to considerable effects in *R* factors and other statistical quantities of the data reduction. On one hand, this is to be expected as these types of errors violate space-group symmetry. On the other hand, this also confirms that the statistical indicators are indeed sensitive to systematic errors. Future work will be necessary to pinpoint the specific sources of systematic error from their influence on the statistical indicators of data quality and other quantities.

3.5. Comparison with experimental data

The $R_{\rm meas}$ values (Diederichs & Karplus, 1997) agree reasonably well between the simulated data and experimental data (Table 3). However, the overall $I/\sigma(I)$ of the experimental data is reported as only 19.3, whereas the $I/\sigma(I)$ of the simulated data is significantly higher (46.0 and 47.7 for *SIM_MX* and *XDS*, respectively). This finding has its most likely explanation in the fact that the $I/\sigma(I)$ of (unmerged) synchrotron data (at most synchrotron beamlines) is rarely above 30 even in the lowest resolution shell, presumably owing to systematic errors that are difficult to identify and control. No such errors were present in the simulation; as a consequence, the merged $I/\sigma(I)$ values in the lowest resolution shell of the simulated data are very high (182 and 186 for *SIM_MX* and *XDS*, respectively), which has a strong influence on the overall values.

3.6. Limitations of the current implementation

In the current version of the program, the following limitations exist.

(i) The polarization of a synchrotron beam and absorption of the diffracted beam in air are not considered. These effects, which are of minor interest, would be easy to implement, but their implementation would be somewhat costly in terms of CPU requirements.

(ii) Crystal size and shape and absorption in the crystal are not considered explicitly. As for small crystals the integrated intensity of a reflection mainly depends on the crystal volume in the beam. This is a good approximation if the diffraction from a crystal is simulated whose projection in the rotation range considered is smaller than the size of a detector pixel. Such small crystals are indeed often used at synchrotrons.

(iii) The finite thickness and absorption of the detector's active surface is not considered. In the usual geometry, with the direct beam impinging perpendicularly on the centre of a flat detector, the finite thickness leads to an offset in reflection positions owing to oblique incidence of the diffracted X-rays and to brightening of high-angle reflections as these are absorbed along a longer path.

(iv) Detector properties such as point-spread function, nonlinearity of response and read-out noise are not implemented.

(v) The background is assumed to be uniform. Water scattering, ice rings and diffuse scattering are not (yet) implemented.

(vi) The cell variation is treated as if it would only affect positions in reciprocal space. It does not take into account the fact that with a different cell the reciprocal lattice is also sampled at different positions, leading to different intensities. This therefore limits the accuracy of the modelling of the underlying physical phenomenon, but has no negative consequences for testing data-reduction programs as these return a single number for the intensity.

Future versions of the program may remove some of these limitations. In particular, it is planned to make source code available that enables modification of the contents of each pixel before noise is added and frames are written. This would make it possible to address limitations (i)–(v).

4. Summary

A useful property of the program described here is that it was developed independently and using a different theory about reflection shape and rocking curve compared with the data-reduction programs that may be tested with its help. This theory may not directly lead to better data-reduction software, but it does help to better understand the physical factors influencing reflection shapes and rocking curves.

The synthetic data obtained from *SIM_MX* allow a comparison between known (ideal) intensities and intensities from a data-reduction program. On one hand, *SIM_MX* may be used to plan an experiment if the physical parameters governing reflection shape and rocking curve are known (*e.g.* from a global analysis of a data set). On the other hand, optimization of data-reduction options may be achieved with the help of model data produced with *SIM_MX* after the experiment. Systematic errors introduced during data reduction may be avoided in a given experimental situation by a

suitable choice of program options or they may be estimated and even partly corrected.

The program allows the simulation, visualization and quantification of the influence of the major imperfections that occur in a non-ideal experiment. Beam divergence and wavelength dispersion can be experimentally controlled and optimized at synchrotron beamlines and thus play a minor role in the reflection shape and rocking curve of most data sets. Broad reflections, which lead to poor signal-to-noise ratio, mainly arise from physical properties of the protein crystal. The experience of the author is that for most protein crystals reflections are not markedly elongated along circles corresponding to their *d*-spacing; therefore, 'rotational mosaicity' appears to play a minor role. In contrast, and in agreement with the findings of Nave (1998), the model calculations suggest that, apart from inhomogeneity and disorder in unit cells, unit-cell parameter variations are responsible for most of the imperfections that result in poor diffraction properties of crystals.

The program, in the form of an executable for a Linux operating system, may be obtained from the author upon request.

The author wishes to thank W. Kabsch for discussion and H. J. Bernstein for implementation of a Fortran90 interface to CBFlib.

References

Abrahams, J. P. (1993). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **28**, 3–4.

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

Artemiev, N., Hrdý, J., Peredkov, S. & Artemev, A. (2004). *J. Synchrotron Rad.* **11**, 157–162.

Bernstein, H. J. & Ellis, P. J. (2005). *International Tables for Crystallography*, Vol. *G*, edited by S. R. Hall & B. McMahon, ch. 5.6. Heidelberg: Springer.

Bernstein, H. J. & Hammersley, A. P. (2005). *International Tables for Crystallography*, Vol. *G*, edited by S. R. Hall & B. McMahon, ch. 2.3. Heidelberg: Springer.

Campbell, J. W. (1996). *Jnt CCP4/ESF-EACBM Newsl. Protein Crystallogr.* **32**, 15–30.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

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

Holton, J. (2008*a*). *Acta Cryst.* A**64**, C77.

Holton, J. (2008*b*). Personal communication.

Hülsen, G., Broennimann, C., Eikenberry, E. F. & Wagner, A. (2006). *J. Appl. Cryst.* **39**, 550–557.

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

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

Kabsch, W. (2006). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, ch. 11.3. Heidelberg: Springer.

Kabsch, W. (2006). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, ch. 25.2.9. Heidelberg: Springer.

Kolatkar, A. R., Clarage, J. B. & Phillips, G. N. (1994). *Acta Cryst.* D**50**, 210–218.

Lambert, S. & Guillet, F. (2008). *J. Appl. Cryst.* **41**, 153–160.

L'Ecuyer, P. (1999). *Math. Comput.* **68**, 249–260.

Leslie, A. G. W. (1992*a*). *Crystallographic Computing 5: From Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 50–61. Oxford University Press.

Leslie, A. G. W. (1992*b*). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.

Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* D**61**, 1227–1237.

Nave, C. (1998). *Acta Cryst.* D**54**, 848–853.

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

Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.

Phillips, G. N. (1995). *Biophys. J.* **69**, 1281–1283.

Yamada, T., Kawahara, N., Doi, M., Shoji, T., Tsuruoka, N. & Iwasaki, H. (2001). *J. Synchrotron Rad.* **8**, 1047–1050.