Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Choice of data-collection parameters based on statistic modelling

Alexander N. Popov[a,c]* and Gleb P. Bourenkov[b,c]

[a]European Molecular Biology Laboratory (EMBL) Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany, [b]Max-Plank-Arbeitsgruppen fur Strukturelle Molekularbiologie, Arbeitsgruppe Proteindynamik, Notkestrasse 85, 22603 Hamburg, Germany, and [c]A. V. Shubnikov Institute of Crystallography, Russian Academy of Sciences, Leninski Prospekt 59, 117333 Moscow, Russia

Correspondence e-mail:
sasha@embl-hamburg.de

Received 19 February 2003
Accepted 10 April 2003

A new method and the software program *BEST* for optimal planning of X-ray data collection from protein crystals using the rotation method are presented. From one or a few initial diffraction images, *BEST* estimates the statistical characteristics of the data set for different combinations of data-collection parameters and suggests the most optimal ones. The anisotropy in diffraction and the permitted width of oscillation without spatially overlapping reflections are taken into account. According to the option chosen, the optimal set of parameters provides a given average signal-to-noise ratio at a given resolution either in the shortest time or with the minimum total radiation dose. *BEST* has been successfully used at the protein crystallography beamlines at DORIS (DESY). The software proved to be extremely useful in using the available data-collection time in the most efficient way.

1. Introduction

The quality of X-ray diffraction data depends drastically on correct choices of experimental conditions and data-collection strategy. Most of the parameters involved in data collection should be considered for each individual application. Owing to advances in synchrotron radiation, area detectors, cryotechniques and software in recent years, collection of X-ray diffraction data from macromolecular crystals has become easier and faster. However, the choice of crystal-to-detector distance, oscillation width, exposure time per frame and detector diameter still requires decisions to be made by the experimenter and cannot be treated in a fully automatic manner.

Usually, the data-collection strategy is chosen on the basis of an examination of several initial diffraction images that are measured with different exposure times, crystal-to-detector distances and rotation angles. The choice is guided by general rules and recommendations (*e.g.* Arndt & Wonacott, 1977; Dauter, 1999; Pflugrath, 1999; Evans, 1999). The shortest total rotation range (continuous or discontinuous) that provides a complete data set and the maximum rotation range per frame that excludes reflection overlaps are estimated on the basis of known crystal class, unit-cell parameters and mosaicity using available programs (*e.g.* Ravelli *et al.*, 1997; Leslie, 1992; Otwinowski & Minor, 2001). In order to optimize further parameters, *e.g.* the resolution of data collection and scan speed, one should take into account characteristics of the particular crystal and X-ray equipment, the available time for measurement and possible radiation damage to the crystal. The relationships between the quality of the data set and the data-collection parameters are complex and experimentalists can predict the results of data collection only approximately.

Here, we present an algorithm for calculation of the statistical results of data collection based on a limited number of initial experimental measurements. Taking into account the anisotropy in the diffraction and the geometrical parameters of data collection, a set of parameters that provides the requested statistics with minimal data-collection time or radiation dose can be estimated.

2. Overview of the method

The quality of measured diffraction data is usually judged by the $R_{\rm merge}$ factor, the ratio of the average squared structure factors to their average uncertainties $\hat{J}/\hat{\sigma}_J$ and by the cumulative distributions of the latter. Simple approximate relationships between the different statistical characteristics are applicable (Diederichs & Karplus, 1999; Weiss, 2001). In the following, we derive expressions for the average squared structure factors $\hat{J}$ and their standard uncertainties $\hat{\sigma}_J$ as functions of the parameters describing the coherent and incoherent scattering by the sample, the instrument and the experiment. Using these expressions, experimental parameters can be found that will yield data with desired statistical characteristics, provided the necessary set of parameters describing the sample is estimated from a small number of initial diffraction exposures.

2.1. Estimation of average squared structure factors

For a majority of protein crystals, the probability density functions for diffraction intensities derived by Wilson (1949) are effectively applicable. Crystals exhibiting pseudosymmetry and/or twinning are exceptions and lie outside the scope of the present considerations. In a canonical derivation (*e.g.* Drenth, 2001), the variation of the first moment of the acentric distribution over reciprocal space is a function of the length of the diffraction vector only. It is defined by the variation in the averaged squared atomic scattering factor and by the overall temperature factor. We attempt to find a more accurate approximation that takes into account the general features of the radial distribution of interatomic distances in protein structures and the overall anisotropic temperature factor.

Neglecting the temperature factors, the pattern of spherically averaged squared structure-factor magnitudes (the first moment of the Wilson distribution) for different protein molecules differs significantly to a resolution of d = 7–5 Å, but appears to converge at higher resolution (Svergun *et al.*, 2001). This suggests that a unique pattern of average squared structure-factor magnitudes should exist at higher resolution for all proteins. In order to derive the distribution $\hat{J}_u(h = 2\sin\theta/\lambda)$, where θ is the Bragg angle and λ is the X-ray wavelength, we scaled together and averaged in resolution bins the diffraction data from 72 randomly chosen protein crystals with different folds, molecular masses, space groups and data-collection resolution (Fig. 1*a*). These data sets have been collected at DESY protein crystallography beamlines during recent years and processed using *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997). The resolution range 12.0–0.9 Å was split into 300 bins in reciprocal space. The resulting curve was tabulated as a function of h. An estimate of the discrepancy between the data sets in fine resolution shells (R factor) is shown in Fig. 1(*b*). Except for the first (low) and last (high) resolution shells, the R factor does

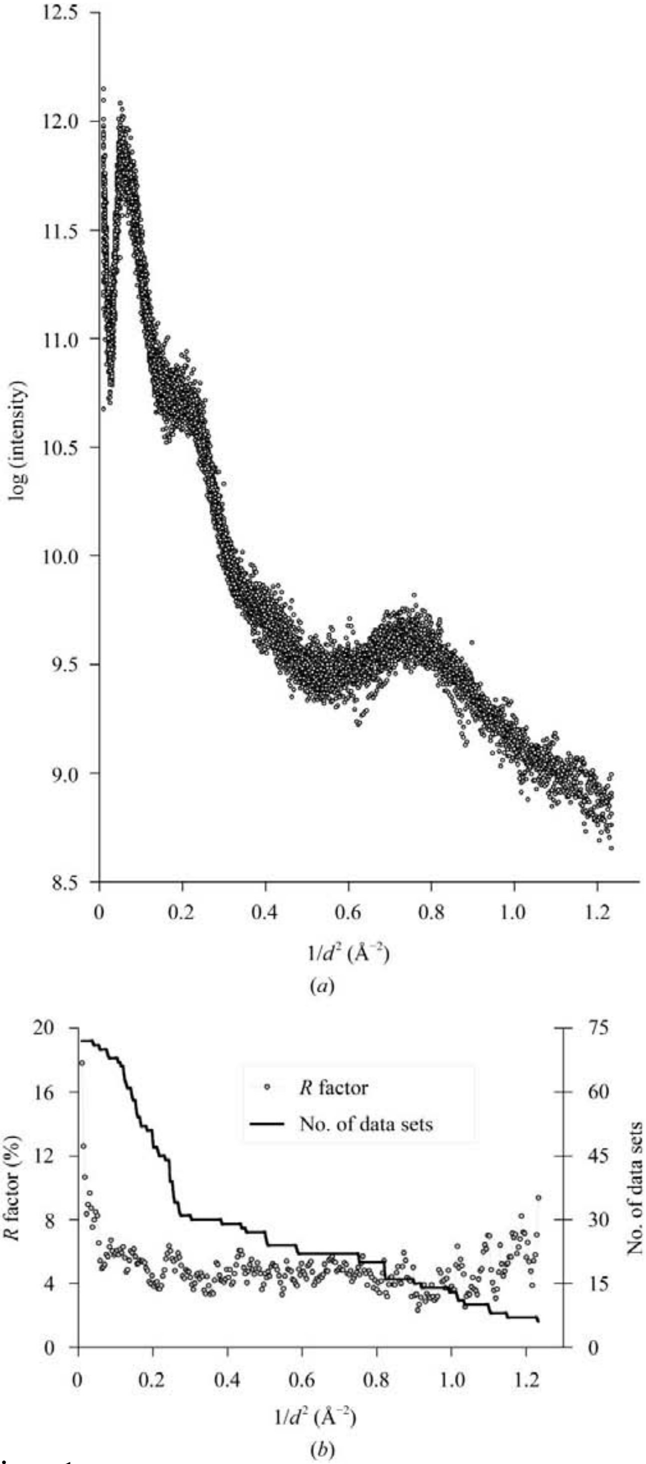

Figure 1
(*a*) The average intensity *versus* resolution (from 12 to 0.9 Å) for 72 protein crystals. (*b*) The number of data sets used and the discrepancy between data sets ($R = \sum_{i=1}^{n} |\hat{I}_c - \hat{I}_i| / \sum_{i=1}^{n} |\hat{I}_i|$), where n is the number of data set in the resolution shell) *versus* resolution.

not exceed 5%. The increase in the discrepancy for the highest resolution is a result of the poorer statistics for these measurements.

The expectation value of a squared structure-factor magnitude for a protein crystal at any point in reciprocal space can be approximated by

$$\hat{J}(\mathbf{h}) = \frac{1}{s}\hat{J}_u(h)\exp(-\mathbf{hBh}^T), \tag{1}$$

where $\mathbf{h}$ is a vector of position in reciprocal space with magnitude $h = |\mathbf{h}|$, $\mathbf{B}$ is an overall anisotropic thermal tensor (obeying the metric point symmetry of the crystal), s is an overall scale factor that describes various factors, such as the intensity of the incident beam, the volume of the crystal sample, the volume of the unit cell $V_{\rm cell}$ *etc*.

For any experiment in which a reciprocal-space volume $V >> 1/V_{\rm cell}$ has been sampled, we can estimate the average value of the squared structure factor,

$$\hat{J}(V) = \frac{1}{V}\int_V\int_0^\infty J(\mathbf{h})p[J|\hat{J}(\mathbf{h})]\,{\rm d}J\,{\rm d}V. \tag{2}$$

Here, p is the well known Wilson's probability density function. For a thin resolution shell with volume v, the average intensity is

$$\hat{J}(h, v) = \frac{1}{v}\int_v \hat{J}(\mathbf{h})\,{\rm d}v. \tag{3}$$

Further, we consider the case when the volume v corresponds to a narrow rotation interval with mean spindle position φ. It is easy to show that v can be partitioned into N equal subvolumes by a set of parallel planes that are perpendicular to the rotation axis and intersect the axis at equidistant points $-\sin 2\theta/\lambda < \zeta_i < \sin 2\theta/\lambda$, where ζ is a reciprocal-lattice coordinate parallel to the rotation axis of the crystal. (3) can then be approximated with any required precision by the discrete sum

$$\hat{J}(h, \varphi) = \frac{1}{2N}\sum_{i=1}^{N}[\hat{J}(\mathbf{h}_{i1}) + \hat{J}(\mathbf{h}_{i2})]. \tag{4}$$

$\mathbf{h}_{i1}$ and $\mathbf{h}_{i2}$ are the coordinates of the intersection points of a resolution sphere h, a plane passing through ζ_i and the Ewald sphere. The expectation values of the squared structure-factor modulus $\hat{J}(h)$ over a large rotation interval measured with constant or variable $\omega(\varphi)$ and $s(\varphi)$ can be found *via* summation over φ.

2.2. Estimation of averaged standard uncertainties

According to Darwin's formula (Darwin, 1914), the integrated reflection intensity is given by

$$I(\mathbf{h}) = \frac{1}{\omega}L(\mathbf{h})P(\mathbf{h})A(\mathbf{h})J(\mathbf{h}), \tag{5}$$

where $L(\mathbf{h}) = L(h, \zeta)$ is the Lorentz factor and $P(\mathbf{h}) = P(h, \zeta)$ is the correction for polarization (Kahn *et al.*, 1982). A is the correction for absorption in the diffracted beam path and ω is the angular velocity of the crystal rotation. Neglecting the absorption in the crystal, $A(\mathbf{h}) = A(h)$ and basically describes absorption in air.

As is known from the basics of rotation-method data processing (see, for example, Leslie, 1999), the variance in the integrated intensity measured by the box summation can be approximated by the second-order polynomial function of I. Similar representation of σ_J as a function of J is valid according to (5). The use of profile fitting usually results in a reduction in the random error associated with weak intensities by a factor of 1.2–1.3 (Leslie, 1999).

In the following, we express the polynomial coefficients $k_{0\text{–}2}$; the index corresponds to the power of J, as functions of the reciprocal-space coordinates (h, ζ, φ) in the reciprocal-space volume v that was defined earlier.

The coefficient k_0 describes the contribution of background counting statistics to the total variance. The total background can be considered as a sum of three contributions: (i) the incoherent scattering arising from several sources, including scattering from the crystal, liquid around the crystal, the sample holder, the beam stop, air and slits; (ii) the detector dark current, which depends on the exposure time only; (iii) the detector readout noise, which is added once per readout. Let us designate $\rho_c(h)$, $\rho_{\rm darc}$ and $\rho_{\rm readout}$ as the densities of the background components (i), (ii) and (iii), respectively; $k_b = [n_p(n_p + n_b)]/n_b$, where n_p and n_b are the number of pixels in the peak and background region of the measurement box, respectively; $\Delta S_p = S_p/R\cos^2 2\theta$ is the angular size of the pixel with detector pixel size S_p and crystal-to-detector distance R. Elastic scattering originating from the area close to the sample position is assumed. Variation in k_b with R can be neglected when the contribution of the sample size and detector resolution to the spot size dominate over the contribution of mosaicity and beam divergence. Otherwise, the dependence $k_b(R)$ can be introduced when separate contributions are known. Further, we assume that the volume v is sampled by several exposures with an oscillation width $\Delta\varphi(\varphi)$. The reflections are then integrated over $(\Delta\Phi/\Delta\varphi) = [\mu L(h, \zeta)\sin 2\theta/\Delta\varphi] + 1$ frames on average; the corresponding probability distribution is uniform. μ is a convolution of the contributions of the mosaic spread, beam divergence and wavelength bandwidth to the angular width of the reflection (see Helliwell, 1995). Thus, taking into account the influence of polarization and absorption onto the scattering background component and the detector gain G,

$$k_0(h, \zeta, \varphi) = \frac{Gk_b\Delta\Phi(\Delta S^2 PA\rho_s\omega + \rho_{\rm dark}\omega + \rho_{\rm readout}\omega^2/\Delta\varphi)}{(LPA)^2}. \tag{6}$$

The coefficient k_1 that defines the contribution of peak-counting statistics to the total variance is expressed as

$$k_1(h, \zeta, \varphi) = \frac{G\omega}{LPA} \tag{7}$$

and the coefficient k_2, describing the contribution of the instrumental error, is expressed as a constant

$$k_2 = k_{\rm ins}. \tag{8}$$

The value of $k_{\rm ins}$ is defined by imperfections in the detector, goniostat and the X-ray source. $k_{\rm ins} \simeq 0.001$ for currently available diffractometers.

The average standard uncertainty of the observed squared structure factors for a sampled volume v is approximated by

$$\hat{\sigma}_J(h,\varphi) = \frac{1}{2N}\sum_{i=1}^{N}\int\limits_{0}^{\infty}(k_{0i}+k_{1i}J+k_2J^2)^{1/2} \times \{p[J|\hat{J}(\mathbf{h}_{i1})]+p[J|\hat{J}(\mathbf{h}_{i2})]\}\,{\rm d}J, \qquad (9)$$

where $k_{(0,1)i} = k_{(0,1)}(h, \zeta_{\rm i}, \varphi)$. After merging symmetry-equivalent/redundant observations over a large rotation range subdivided into m equally small intervals centred at spindle positions φ_j and assuming that redundant measurements are uniformly distributed over the rotation range, the expectation values of standard uncertainties can be estimated as

$$\hat{\sigma}_J(h) \cong M^{-1/2}m^{-1}\sum_{j=1}^{m}\hat{\sigma}_J(h,\varphi_j). \qquad (10)$$

Here, M is the mean redundancy of the measurements in the total rotation interval.

2.3. Optimization of data-collection parameters

Using (4) and (9), we can estimate the ratio of intensities to their uncertainties $\hat{J}/\hat{\sigma}_J(h)$ for any set of data-collection parameters. We can also perform the reverse task and determine the sets of parameters corresponding to a given $\hat{J}/\hat{\sigma}_J(h) = C$, assuming that the total rotation range ($\varphi_{\rm start}$–$\varphi_{\rm end}$) is defined (*e.g.* as the shortest range providing a complete data set) and the average redundancy M over this range is calculated. For every $\varphi_{\rm start} \le \varphi_j \le \varphi_{\rm end}$, solutions to the equations

$$\frac{\hat{J}}{\hat{\sigma}_J}(h,\varphi_j) = \frac{C}{M^{1/2}} \qquad (11)$$

with respect to ω_j, $\Delta\varphi_j$ and R_j are found that satisfy the condition of total data-collection time,

$$T_{\rm total}(h) = \frac{\varphi_{\rm end}-\varphi_{\rm start}}{m}\sum_{j=\varphi}^{m}\frac{1}{\omega_j}\left(1+\frac{t_{\rm det}}{\Delta\varphi_j}\right), \qquad (12)$$

where $t_{\rm det}$ is the detector readout time, or the total radiation dose

$$D_{\rm total}(h) \propto \frac{\varphi_{\rm end}-\varphi_{\rm start}}{m}\sum_{j=\varphi}^{m}\frac{1}{\omega_j} \qquad (13)$$

being minimal. In addition, the upper limit $\Delta\varphi_{\rm max}(h)$ defined by the geometrical conditions that the reflections in the range (0–h) do not overlap must be taken into account. A non-zero value of $\rho_{\rm readout}$ ensures $\Delta\varphi_j > 0$ in the solution. If several modes of detector operation are available that differ in $t_{\rm det}$ and radius (such as for a MAR imaging plate), optimization is carried out for each mode separately and the optimal mode providing the shortest $T_{\rm total}$ is selected. Other constants and functions involved in (1–13) are considered to be invariant for a particular experiment.

Situations when (11) has no solution are possible. The trivial case when $\Delta\varphi_{\rm max}(h, \varphi) = 0$ requires either the total interval to be re-chosen or, when this is not feasible, sets up a resolution limit. Others are the cases when the value of $C/N^{1/2}$ exceeds its upper limit defined by the $k_{\rm ins}$ parameter or the $k_0(h, \zeta, \varphi)$ value exceeds the digital limit of the detector (usually 2^{16}) in a particular point in reciprocal space at the solution. In such cases, the total rotation range is expanded and thus M is increased until the solution is feasible.

In each experiment $T_{\rm total}$ is limited. In addition, an estimate of the total dose that the sample can sustain without significant radiation damage may be available, setting up an upper limit to the $D_{\rm total}$. These limitations define the resolution of the data $h_{\rm max}$.

2.4. Other characteristics of data quality

The data quality is traditionally judged by $R_{\rm merge} = \sum_{hkl}\sum_i|J_{hkl,i} - \hat{J}_{hkl}|/\sum_{hkl}\sum_i\hat{J}_{hkl}$, where $\sum_{hkl}$ denotes the sum over all reflections and $\sum_i$ the sum over all equivalent and symmetry-related reflections. We can estimate the $R_{\rm merge}$ factor if we assume Gaussian measurement errors. Then, according to approximation (10), the expectation value of the absolute deviation from the mean can be expressed by

$$\sum_{i}^{M}|J_{hkl,i} - \hat{J}_{hkl}| \simeq [M(M-1)]^{1/2}(2/\pi)^{1/2}\sigma(J_{hkl}) \qquad (14)$$

and $R_{\rm merge}$ is approximated by

$$R_{\rm merge} \simeq \left[\frac{2(M-1)}{\pi}\right]^{1/2}\frac{\hat{\sigma}(J)}{\hat{J}}. \qquad (15)$$

Additional information about the quality of a data set can be derived from the distribution of relative number of reflections $E(q)$ with a $J/\sigma(J)$ ratio less then a given q,

$$E(q) = \frac{1}{2Nm}\sum_{j=1}^{m}\sum_{i=1}^{N}\int\limits_{0}^{z_{ij}}\{p[J|\hat{J}(\mathbf{h}_{ij1})]+p[J|\hat{J}(\mathbf{h}_{ij2})]\}\,{\rm d}J. \qquad (16)$$

z_{ij} is a positive solution of the equation

$$(k_{0i}+k_{1i}z_{ij}+k_2z_{ij}^2)^{1/2} = qz_{ij}. \qquad (17)$$

3. Implementation

The above ideas were implemented in a computer program *BEST* for planning X-ray data collection on the basis of one or a few initial rotation exposures.

The initial image(s) must be evaluated, *i.e.* the unit cell and orientation matrix determined, the mosaicity μ and spot size (k_b) estimated and (partial) reflection intensities integrated. These tasks are readily performed by available data-processing software. In its current implementation, *BEST* runs in combination with *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997).

The procedure consists of several stages, which are described in more detailed below, along with the definitions of requirements for the necessary amount of initial data and conditions for measuring them.

Table 1
Test-data collection.

Protein	Space group	Unit-cell parameters (Å)	Beamline/ detector	λ† (Å)	N‡	$\Delta\varphi$§ (°)	$t_{\rm exp}$¶ (s)	$d_{\rm min}$†† (Å)	μ‡‡ (°)	B§§ ($Å^2$)
Catalase	$P2_13$	$a = b = c = 131.3$	BW7a/MAR CCD	1.00	1	1.0	20	1.50	0.35	10.0
Lysozyme	$P4_32_12$	$a = b = 78.71$, $c = 36.82$	X11/MAR CCD	0.81	2	0.3	20	1.14	0.40	10.9
Isomerase	$C2$	$a = 102.18$, $b = 78.1$, $c = 113.7$, $\beta = 116.1°$	X13/MAR CCD	0.80	4	1.0	20	1.95	0.95	27.6

† X-ray wavelength. ‡ Number of initial images. § Width of oscillation. ¶ Exposure time per initial image. †† Initial image resolution. ‡‡ Crystal mosaicity. §§ Overall B factor.

3.1. Scaling

The scale factor s and the overall anisotropic thermal tensor $\mathbf{B}$ are estimated by fitting the measured intensities to the pattern. Usually, a single diffraction frame contains mostly or only partial reflections. The values of the partiality are estimated and applied to the intensities of available partial reflections to obtain estimates of the J values. The profile of the intensity as a function of the rotation angle is approximated by a triangle with a base that is equal to the angular width of the reflection and is calculated according to the Greenhough & Helliwell (1982) model. The triangular approximation works well for reflections with the centre of the rotation range lying within the scan range of the exposure and when the condition $\Delta\varphi \geq \mu/2$ is satisfied. If the latter condition is not satisfied for a particular sample without severe spatial overlapping of the reflection spots, a series of rotation exposures with smaller $\Delta\varphi$ should be measured and the partial intensities summed.

The estimates of parameters s and $\mathbf{B}$ are obtained by maximizing the likelihood function

$$W(s, \mathbf{B}) = \sum_n \left[\log(s) + \mathbf{h}_n \mathbf{B} \mathbf{h}_n^T - s \exp(\mathbf{h}_n \mathbf{B} \mathbf{h}_n^T) \frac{J_n}{\hat{J}(h_n)} \right], \quad (18)$$

with summation over all measured acentric reflections. Symmetry constraints on $\mathbf{B}$ and the crystal orientation are taken into account. An inverse of the second partial derivatives matrix of W with respect to parameters approximates the covariance matrix. The values of $\hat{J}(\mathbf{h})$ and its relative error $\delta_{\hat{J}}(\mathbf{h})$ are computed using (1) and its first differential with respect to the parameters, respectively. The error in further estimations of $\hat{J}/\hat{\sigma}_J(\mathbf{h})$ is defined by $\delta_{\hat{J}}(\mathbf{h})$ as a dominating term. The average values of $\hat{\delta}_{\hat{J}}(h, \varphi)$ over a volume v are computed as a function of the resolution and spindle position. The φ value corresponding to the maximum average error is selected at a position where further exposure(s) have to be measured in case the initial one(s) do not provide sufficient data for the desired accuracy of prediction. For triclinic or monoclinic crystals, the covariance matrix computed on the basis of measurements made at only a single spindle position is practically always degenerate. In such cases, the direction of a second exposure is found by eigenvector analysis of the second derivative matrix. For higher symmetries, a single orientation is usually sufficient for $\delta_{\hat{J}}$ to be <10% up to the highest resolution shells for well diffracting crystals.

The scaling procedure is based on a very limited amount of data. Systematically omitted strong (*e.g.* arising from detector overload) or weak (*e.g.* arising from rejections) reflections may introduce severe bias that is difficult to detect. Test exposures should be measured with a relatively short exposure time and evaluated without applying any rejections based on the estimated standard uncertainties of individual reflections. An appropriate resolution cutoff is permitted. Note that both $\hat{J}(\mathbf{h})$ and $\delta_{\hat{J}}(\mathbf{h})$ can be evaluated up to a resolution limit exceeding the resolution of the data used for scaling.

In the current implementation, s and $\mathbf{B}$ are considered to be fixed during data collection. This corresponds to the assumptions that (i) the exposed crystal volume does not change as a function of spindle position and (ii) there will be no significant radiation damage to the sample during data collection.

3.2. Evaluation of background scattering

The background intensity $\rho_s(h)$ is extracted from the initial diffraction image. The reflection positions and spot sizes are imported from the auxiliary data-reduction software. The pixel intensities are corrected for the detector dark current and the readout noise, the polarization and the angular pixel dimensions. The average values of background density and standard deviations for resolution shells are calculated and outlying pixels (spurious noise or, for example, mis-predicted tails of strong reflections) are rejected in a cyclic way until convergence.

As in the case of the scale factor s, the initial data measured at a one or two spindle positions do not provide a complete description of the three-dimensional background distribution when the exposed volume of the crystal and its environment varies with the spindle position. The approximation obtained from a single orientation will be adequate when assumption (i) in §3.1 holds and sample mounting is optimized to reduce the background scattering.

Since no physical model is available that would permit extrapolation of the background scattering curve to values of h above the limit defined by the geometrical conditions at which the initial exposures has been taken, the background measurements define the absolute resolution limit for further optimization/prediction calculations. This must be taken into account when initial exposure(s) are made.

3.3. Geometrical parameters

Based on the known crystal orientation, unit-cell parameters and crystal class, all reflections that can occur during a 360° rotation of the crystal are simulated and sorted on rotation angle. The rotation range that provides 99% completeness of indices and the reflection multiplicity for this range are then calculated as a function of the starting spindle angle. By default, the shortest total range is used for planning the data collection.

The permitted rotation range $\Delta\varphi_{\max}(h, \varphi)$ has to be determined as a function of both resolution and rotation angle. For every rotation angle, the program simulates the positions of the reciprocal-lattice points and calculates the corresponding volumes of reciprocal space that are used for integrating reflection intensities, taking into account the spot size and the mosaicity. The shortest rotation causing the overlap of neighbouring reciprocal-lattice volumes determines the permitted rotation range. It can be estimated using simple geometrical considerations. The calculations are performed for the area of the diffraction pattern close to the plane containing the incident beam and normal to the rotation axis. In this area, the reciprocal-lattice points pass through the Ewald sphere with the highest velocity.

3.4. Plan

All further calculations are carried out on a two-dimensional grid of the resolution and $\hat{J}/\hat{\sigma}_J$ ratio. By default, the total rotation range is subdivided into sub-ranges with 5° intervals. Given the selected optimization option (time or dose) and the calculated redundancy over the total rotation range, ω and $\Delta\varphi$ are found in each small interval, taking into account the anisotropy of diffraction and the permitted rotation range. The rotation ranges are appropriately adjusted at the edge of the sub-ranges and thus a continuous data-collection protocol is formed. For the sake of convenience, when *BEST* is used in combination with available commercial diffractometer-control programs supporting manual input only, the parameters are merged for the neighbouring intervals if the difference does not account for more then 10% in ω and 0.05° in $\Delta\varphi$. After summation over the sub-ranges, the total data-collection time and total exposure time (proportional to the dose) are tabulated as a function of the resolution and $\hat{J}/\hat{\sigma}_J$ in the last resolution shell. On the basis of the table and according to the specific aim of the experiment, the user can decide which resolution and at which signal-to-noise level can be reached within the total allocated time or within a sustainable dose. For users' choice of $h_{\max}$ and $\hat{J}/\hat{\sigma}_J$, the data-collection plan is output along with the estimates of $\hat{J}$, $\hat{\sigma}_J$, $R_{\rm merge}$ and cumulative distribution $E(q)$ in the resolution shells and for the overall range (0–$h_{\max}$).

4. Testing

Several test data collections and routine use of the program *BEST* at DESY synchrotron beamlines have demonstrated the validity of the above algorithm. In the following, we present three examples of test data collections (Table 1). The

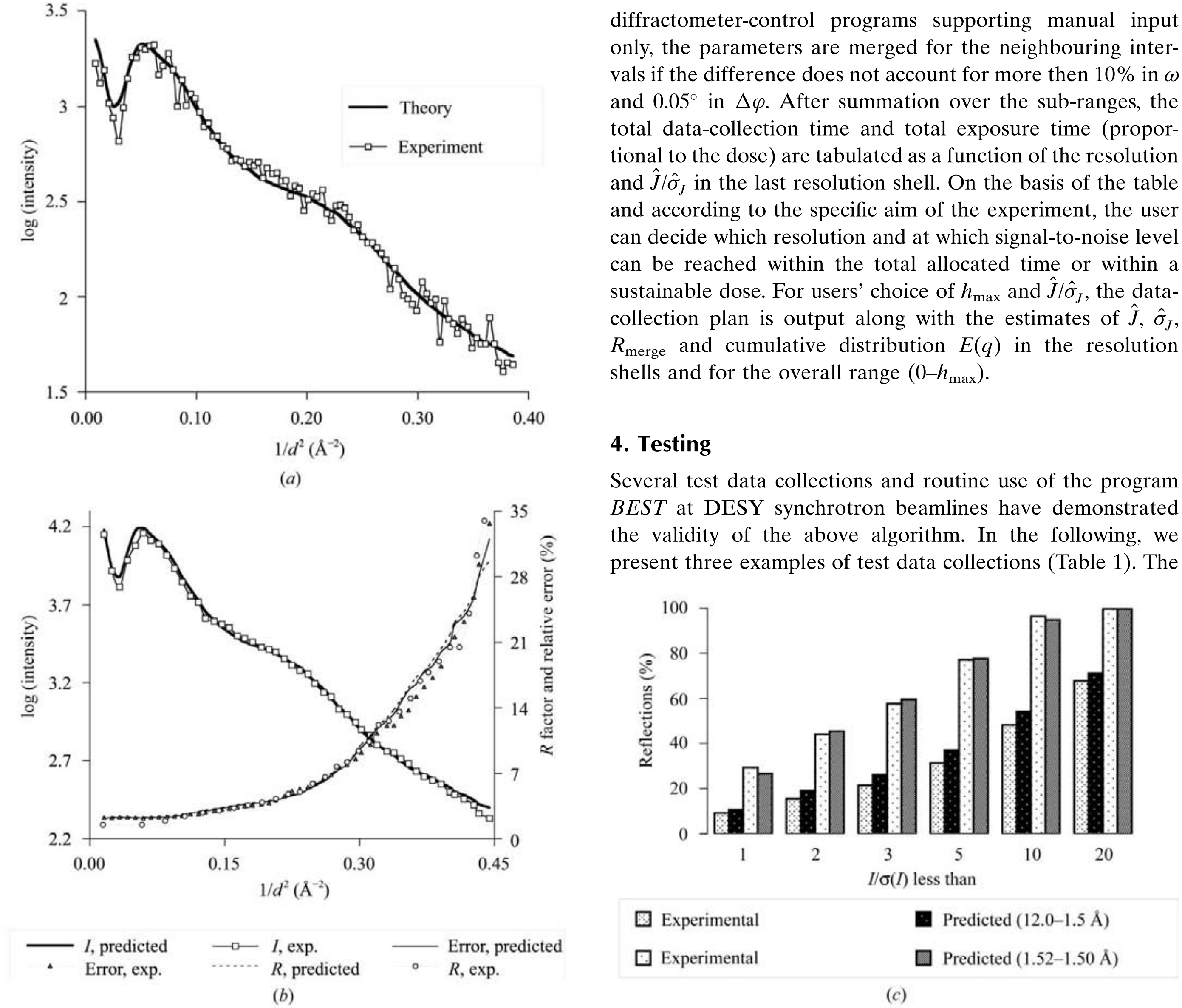

Figure 2

Test data collection of a TTC crystal according to the *BEST* plan (total time of data collection T = 55 min, shortest total angle range Φ = 21°, $\Delta\varphi$ = 0.26° and $t_{\rm exp}$ = 34 s, completeness = 99.6%, redundancy = 2.6). (*a*) Estimation of the average intensity from the resolution using one initial image. Owing to the high symmetry of the TTC crystal and the large number of reflections (6576 full and 6745 partial), one initial image was sufficient for fitting with estimated errors of prediction <1%. (*b*) Comparison of the predicted average intensity, relative error of measurement and $R_{\rm merge}$ with the experimental results. (*c*) Predicted and experimental distributions of the relative number of reflections with given J/σ_J ratio for all data and for the last resolution shell.

programs *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997) were used for both producing the input to *BEST* on the basis of initial exposures and for reduction and scaling of the data used for comparison of the predicted statistics with the experimental results.

4.1. Catalase

Crystals of *Thermus thermophilus* catalase (TTC) belong to the cubic space group and usually exhibit high-quality diffraction. A relatively small crystal of TTC, with size about 0.1 mm, was used for test data collection. Integrated intensities in the resolution range 12–1.85 Å obtained in a single short exposure using the detector distance R = 100 mm were used in scaling (Fig. 2*a*). Following *BEST* recommendations, the data set of TTC was collected at a resolution of 1.5 Å and a predicted $\hat{J}/\hat{\sigma}_J$ of 3 in the last resolution shell. The data statistics predicted by *BEST* are in very good agreement with the experimental results (Figs. 2*b* and 2*c*).

4.2. Lysozyme

A crystal of tetragonal hen egg-white lysozyme was used for test data collection at atomic resolution. Two sequential initial images were sufficient for scaling (Fig. 3*a*) and choice of the optimal plan for data collection. Firstly, data at 1.15 Å resolution and $\hat{J}/\hat{\sigma}_J = 2$ in the last shell of resolution were collected according to the plan. About 5% of reflections at low resolution were overloaded. An additional data set was measure to 1.68 Å resolution and $\hat{J}/\hat{\sigma}_J$ = 8. The experimental and predicted statistics are shown in Figs. 3(*b*) and 3(*c*).

4.3. Isomerase

A monoclinic crystal of human mitochondrial Δ^3-Δ^2-enoyl-CoA isomerase exhibited relatively large mosaicity and moderate diffraction anisotropy. Inspection of the spot profiles indicated minor splitting of the crystal. Four initial images (two sequential images with $\varphi = 0°$ and two images with $\varphi = 90°$) were measured at a geometrical resolution limit of 1.93 Å. Estimations of the permitted rotation ranges showed that the data at 1.93 Å resolution may not be collected without spot overlaps. Thus, the data set was collected at 2.06 Å resolution and $\hat{J}/\hat{\sigma}_J$ = 3 in the last resolution shell. The data-collection plan consisted of four entries with stepwise decreasing rotation speed (Fig. 4*b*). The predicted and experimentally obtained statistics were in good agreement for the highest resolution shells (Fig. 4*c*), where counting statistics predominate over systematic errors. The poor quality of the crystal caused additional errors in measurements that were

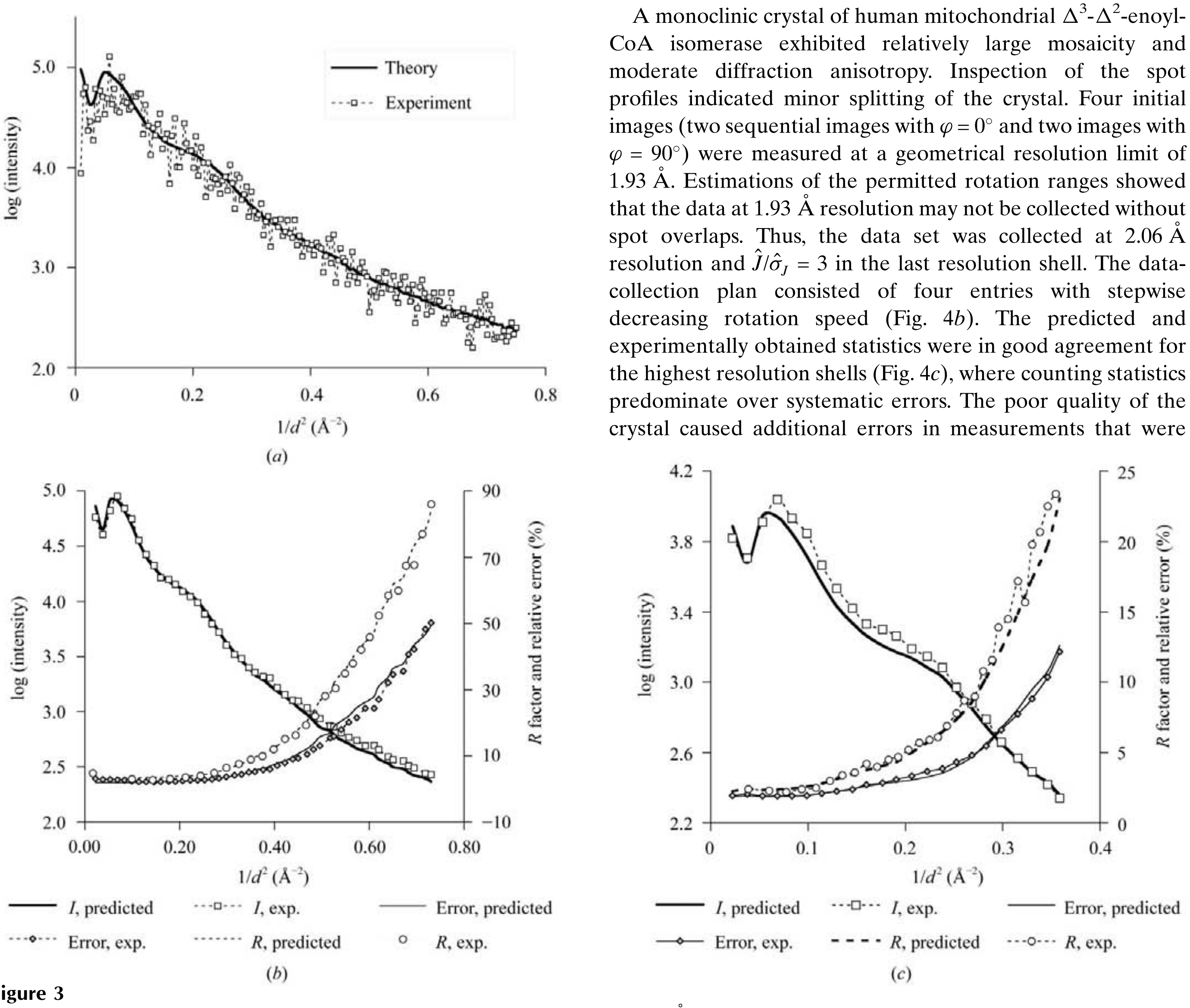

Figure 3
Test data collection of a lysozyme crystal. An atomic resolution data set ($d_{\rm min}$ = 1.15 Å, T = 186 min, Φ = 67°, $\Delta\varphi$ = 0.22°, $t_{\rm exp}$ = 35 s, completeness = 99.1%, redundancy = 5.4) and a low-resolution data set ($d_{\rm min}$ = 1.68 Å, T = 28 min, Φ = 67°, $\Delta\varphi$ = 0.75°, $t_{\rm exp}$ = 10 s) were collected according to the *BEST* plan. (*a*) Estimation of average intensity from resolution using 1820 reflections from two sequential initial images taken with $\Delta\varphi$ = 0.3° and $t_{\rm exp}$ = 20 s. All reflections used were partial. (*b*) Comparison of the predicted and experimental statistics for the atomic resolution data set. (*c*) Comparison of the predicted and experimental statistics for the data set at 1.68 Å resolution.

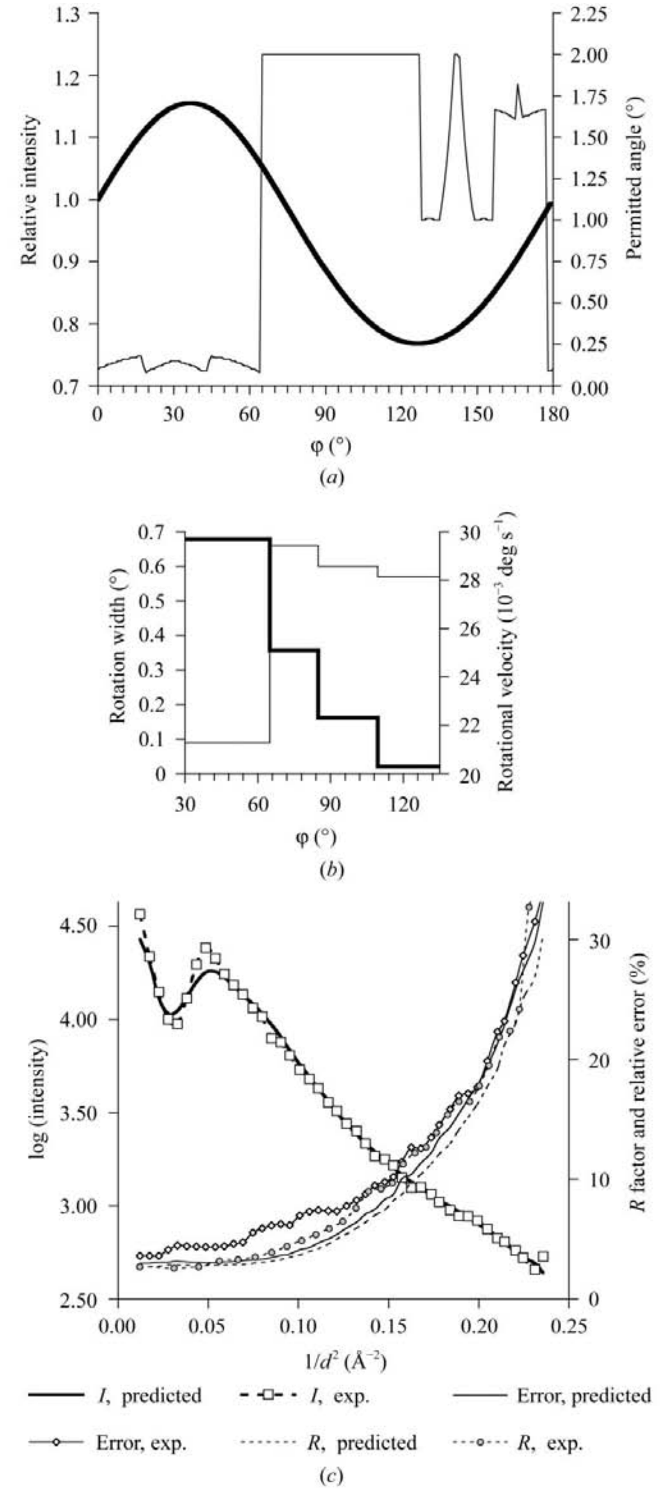

Figure 4
Test data collection of an isomerase crystal. About 7600 partial and 60 full reflections from four initial images were used for determination of the average intensity. (*a*) Relative average intensity (thick line) and permitted rotation range $\Delta\varphi_{\max}$ (thin line) for last resolution shell 2.06 Å *versus* rotation angle. (*b*) Graph presentation of the data-collection plan (the thick line is the rotation velocity and the thin line is the width of rotation). The total rotation range (completeness = 99.3%, redundancy = 2.2) from 30 to 135° is split into four entries according to the diffraction anisotropy and geometrical restrictions. (*c*) Comparison of the predicted and experimental statistics.

not accounted for by the predictions. As a result, the experimental values of $\hat{J}/\hat{\sigma}_J$ at low and medium resolution are about 2% larger then the predicted values.

5. Discussion

The equations and techniques described here provide a foundation for modelling the statistical results of data collection for any combination of set-up parameters and therefore for automatic and optimal planning of the diffraction measurements. The approach can also be used for rational selection of the best diffraction-quality crystal in the crystal screening procedure.

Traditionally, data collection is performed using a constant exposure time (or dose) and oscillation angle per image. Obviously, this is not the optimal method of measurement. In the case of strong diffraction anisotropy, this method of data collection results in systematically worse measurement statistics for the more disordered crystal directions in reciprocal space. To make the distribution of measurement statistics over reciprocal space as uniform as possible, the data can be collected with the scan speed varying with the rotation angle according to the anisotropic behaviour of the diffraction intensity.

Our modelling is based on the assumption that the main uncertainties in the observed intensities are determined by counting statistics. The contribution of the instrumental error is estimated as a constant proportional to the measured intensity. Two constants, G and $k_{\rm ins}$, used in (6–8) must be determined for the given X-ray equipment by the adjustment of the predictions to the experimental results of data collection. In this way, the constant G describes not only the detector gain but also the discrepancy between the box summation and the profile fitting that is usually implemented in the data-processing software. The estimate of the instrumental error has to be about 3–5% and in the absence of strong systematic errors (such as non-uniform crystal rotation speed, instability of the X-ray beam, very poor crystal quality *etc.*) will properly describe the errors in the lower resolution shells. As demonstrated by the example of the isomerase, adequate prediction of data statistics close to the accessible resolution limit can be achieved even in the presence of severe systematic errors. In practice, applications of *BEST* for planning high-resolution data collections aiming at $\hat{J}/\hat{\sigma}_J < 10$ are less sensitive with respect to the calibration of the instrument and quality of a particular crystal. The situation is different when highly accurate data are required; for example, in experiments aiming towards measuring very small anomalous differences. In such cases, the $k_{\rm ins}$ value defines the prediction of the statistics and must be known precisely.

In the current implementation of *BEST*, the assumption is made of invariant coherent and incoherent scattering volumes during experiments. For low-emittance sources, the focal spot sizes are typically smaller then the sample and loop size and thus variation in the exposed crystal volume and its environment with the rotation angle must be taken into account. Extracting the variation from the X-ray measurements alone will require many additional images to be measured in the initial stage of the experiment. Instead, the functions $s(\varphi)$ and $\rho_s(\varphi)$ could be extracted from the optical measurements of the crystal sample, mounting loop and drop. Such optical techniques are currently being developed in context of automated crystal mounting and centring (Wilson, 2002). Provided corresponding data are available, our formulation can take this variation into account without any modification.

Further developments would be required for cases in which radiation damage may not be neglected. Firstly, the variation in s and $\mathbf{B}$ as a function of the absorbed dose must be known. Provided the data are available, this variation can already be accounted for in the present version of *BEST*. Secondly, an additional error arising from merging the data suffering from radiation damage must be modelled. A corresponding correction to σ_J can be approximated within the three-term expansion approach presented here with modified expressions for the dose-dependent k_{0-2}.

The authors would like to thank Drs S. Antonyuk, V. Lamzin and P. Tucker for useful discussions and help.

References

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

Darwin, C. G. (1914). *Philos. Mag.* **27**, 315.

Dauter, Z. (1999). *Acta Cryst.* D**55**, 1703–1717.

Diederichs, K. & Karplus, P. A. (1999). *Nature Struct. Biol.* **4**, 269–275.

Drenth, J. (2001). *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann & E. Arnold, pp. 56–57. Dordrecht: Kluwer Academic Publishers.

Greenhough, T. J. & Helliwell, J. R. (1982). *J. Appl. Cryst.* **15**, 493–508.

Evans, P. R. (1999). *Acta Cryst.* D**55**, 1771–1772.

Helliwell, J. R. (1995). *International Tables for Crystallography*, Vol. C, edited by A. J. C. Wilson & E. Prince, ch. 2.2.7. Dordrecht: Kluwer Academic Publishers.

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

Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696–1702.

Kahn, R., Fourme, R., Gadet, A., Janin, J. & Andre, D. (1982). *J. Appl. Cryst.* **15**, 330–337.

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

Otwinowski, Z. & Minor, W. (2001). *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann & E. Arnold, pp. 226–235. Dordrecht: Kluwer Academic Publishers.

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

Ravelli, R. B. G., Sweet, R. M., Skinner, J. M., Duisenberg, A. J. M. & Kroon, J. (1997). *J. Appl. Cryst.* **30**, 551–554.

Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.

Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–320.

Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.