Quantifying instrument errors in macromolecular X-ray data sets

Kay Diederichs

University of Konstanz, Faculty of Biology, M647, D-78457 Konstanz, Germany

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

An indicator which is calculated after the data reduction of a test data set may be used to estimate the (systematic) instrument error at a macromolecular X-ray source. The numerical value of the indicator is the highest signal-to-noise $[I/\sigma(I)]$ value that the experimental setup can produce and its reciprocal is related to the lower limit of the merging R factor. In the context of this study, the stability of the experimental setup is influenced and characterized by the properties of the X-ray beam, shutter, goniometer, cryostream and detector, and also by the exposure time and spindle speed. Typical values of the indicator are given for data sets from the JCSG archive. Some sources of error are explored with the help of test calculations using *SIM_MX* [Diederichs (2009), *Acta Cryst.* D**65**, 535–542]. One conclusion is that the accuracy of data at low resolution is usually limited by the experimental setup rather than by the crystal. It is also shown that the influence of vibrations and fluctuations may be mitigated by a reduction in spindle speed accompanied by stronger attenuation.

1. Introduction

In general, data collection at a synchrotron beamline is expected to deliver the best possible data set for a given crystal. While it is true that a synchrotron data set is usually much more strongly exposed and therefore delivers higher resolution data than data from a rotating-anode generator, it is sometimes found that home sources deliver more accurate data, *i.e.* data with a higher signal-to-noise $[I/\sigma(I)]$ ratio, than synchrotron data, in particular at low resolution. For some purposes, such as molecular-replacement calculations and refinement, the accuracy of the data may not be of the utmost importance (Borek *et al.*, 2003), but for experimental phasing, in particular in the case of sulfur-SAD phasing, it obviously is. The lack of accuracy of an individual measurement may be partly compensated, within limits given by the time available for the experiment and by the radiation damage to the crystal, by averaging of multiple intensity measurements (observations) of the unique reflections.

Recently, it has been noted (Diederichs, 2009) that at most synchrotron sites and even with good crystals the $I/\sigma(I)$ ratio of the strongest (unmerged) observations is rarely above 30 even in the lowest resolution shell. Obviously, counting statistics are not the limiting factor, as individual reflections may well have many more than 10 000 counts, which would allow $I/\sigma(I)$ ratios of more than 100 and low-resolution R factors of better than 1%.

doi:10.1107/S0907444910014836

The reduced $I/\sigma(I)$ values and elevated R factors are the consequence of several sources of error that cannot be corrected by data-reduction software. These include beam instability, shutter jitter, the goniostat sphere of confusion leading to non-uniform irradiation of the crystal, non-uniformity of the angular speed of the spindle, shutter–spindle desynchronization, detector nonlinearity (*e.g.* errors, owing to dead-time effects or other factors, in the conversion of photon numbers to pixel counts), detector non-uniformity (*e.g.* varying sensitivity and gain across the surface) and detector noise, and crystal vibration arising from the cryostream or other mechanical or electronic influences. Few of these sources of error are under the control of the experimenter. Rather, a beamline user largely depends on the given beamline setup and on the scientists who adjust it.

The wavelength and crystal-to-detector distance depend, for many experiments, on the type and the purpose of the experiment. However, other parameters are available to the experimenter and can be used to optimize the quality of the resulting data (Dauter, 2010) and to minimize the overall error. The most important parameters are the exposure time, the oscillation range, the number of 'passes' (oscillations of the spindle within the exposure time of a frame) and the attenuation of the beam. The maximum oscillation range per frame may be estimated from basic geometrical considerations and under certain assumptions an optimal value may be calculated (Popov & Bourenkov, 2003; Bourenkov & Popov, 2006).

The purpose of this paper is to show that in many cases, after the usual data collection, information is available to quantify the amount of instrument error affecting the data. This may in turn be used to improve the experimental setup and to optimize the data-collection parameters. The focus of the paper is on synchrotron data collection.

2. Methods

Scaling programs account for slow variations in X-ray flux and irradiated crystal volume throughout a data-collection run, compensate for differences in absorption owing to the different paths of incoming and diffracted X-rays and partially adjust for radiation damage. These systematic effects as well as random errors arising from counting statistics and other sources are taken into account in an approach based on a model of the experiment and the probability distributions describing its error components (Borek *et al.*, 2003). However, the model of the experiment and its treatment by the software may be incomplete (*e.g.* owing to a lack of information about a particular effect) and partly inadequate (*e.g.* owing to approximations). Thus, systematic effects that are not properly modelled may give rise to systematic errors.

As a result of the processing, each reflection is assigned an estimate of the variance σ^2_{hkl} of its intensity I_{hkl} (the hkl subscript is omitted in the following). A simple model, which (as required by error-propagation theory; Bevington, 1969) adds the random and systematic variance components from all sources of error, has its historical foundation in the error model applicable to scanners for X-ray film (Leslie, 1999; Evans, 2006). It is given by

$$\sigma^2 = \sigma^2_{\rm counting} + KI^2. \quad (1)$$

Here, for each reflection, the first term $\sigma^2_{\rm counting}$ gives the variance from Poissonian counting statistics and includes the background term. This approach uses a single constant K for each data set. K may be adjusted such that the observed spread of intensities of symmetry-related reflections on average matches their variance, which is equivalent to stating that on average the χ^2 values should be around unity. For strong reflections, the background is negligible and $\sigma^2_{\rm counting}$ is approximately equal to the intensity I. In this case, taking the square root and dividing by I leads to

$$\sigma/I \simeq (K + I^{-1})^{1/2}. \quad (2)$$

Therefore, I/σ cannot be larger than $1/K^{1/2}$. Current data-reduction programs usually employ a more elaborate approach with, for example, two adjustable constants K_1 and K_2 for a data set,

$$\sigma^2 = K_1\sigma^2_{\rm counting} + K_2I^2, \quad (3)$$

such that deviations of the actual detector gain (the conversion factor from photons to counts) from the value used for calculating the Poissonian counting statistics can also be accounted for (Leslie, 2006).

The variable names K_1 and K_2 are used throughout this paper to indicate their general nature. Current data-processing programs employ mathematically equivalent variations of (3) and often name the variables in a descriptive way. For example, *SCALA* (Collaborative Computational Project, Number 4, 1994; Evans, 2006) uses the names SDFAC and SDADD, *SCALEPACK* (Otwinowski & Minor, 1997) uses variables called 'error scale factor' and 'estimated error', and *DSCALEAVERAGE* (from the *d*TREK* package; Pflugrath, 1999) employs the variables $E_{\rm mul}$ and $E_{\rm add}$.

2.1. Upper theoretical limit of I/σ in a data set

XDS (Kabsch, 2010*a*,*b*), which was used to process the simulated and measured data sets in this work, employs the following variation of (3),

$$\sigma^2 = K_1(\sigma^2_{\rm counting} + K_2I^2), \quad (4)$$

where the variables K_1 and K_2 are termed a and b in Kabsch (2010*a*). With this choice of error model,

$$(I/\sigma)^{\rm asymptotic} = \frac{1}{(K_1K_2)^{1/2}} \quad (5)$$

is the limiting value of I/σ for large I and can serve to quantify the contribution of systematic error to the total variance of a reflection.

To use (5), a peculiarity of *XDS* has to be taken into account: the program adjusts the variances in two phases of data processing. Fixed values of $K_1' = 4$ and $K_2' = 0.0001$ are employed in the INTEGRATE step and adjustable values K_1'' and K_2'' are computed in the CORRECT step; the latter are

denoted by a and b in the logfile `CORRECT.LP`. Combining these corrections, the resulting constants are $K_1 = 4K_1''$ and $K_2 = K_2''/4 + 0.0001$ and thus

$$(I/\sigma)^{\text{asymptotic}} = \frac{1}{[K_1''(K_2'' + 0.0004)]^{1/2}}. \quad (6)$$

Similar formulae can be found for other data-processing software. However, as the definitions of the variables differ, the relevant documentation has to be consulted. Furthermore, the least-squares procedures used to obtain a χ^2 value of 1 differ between these programs and some of them also let the user specify the values of the variables. In addition, different programs employ different methods for outlier rejection. Thus, values of $(I/\sigma)^{\text{asymptotic}}$ from different programs are not directly comparable.

For strong reflections, $(I/\sigma)^{\text{asymptotic}}$ is almost reached and can thus be obtained from a plot of I/σ *versus* I (Fig. 1). The latter analysis can be performed routinely for any data-processing software, but it requires data that are strong enough to show the asymptotic behaviour and does not deliver information about K_1 and K_2. Furthermore, the unmerged intensities have to be analyzed, which is the default output format for *XDS* and can be optionally chosen for other packages.

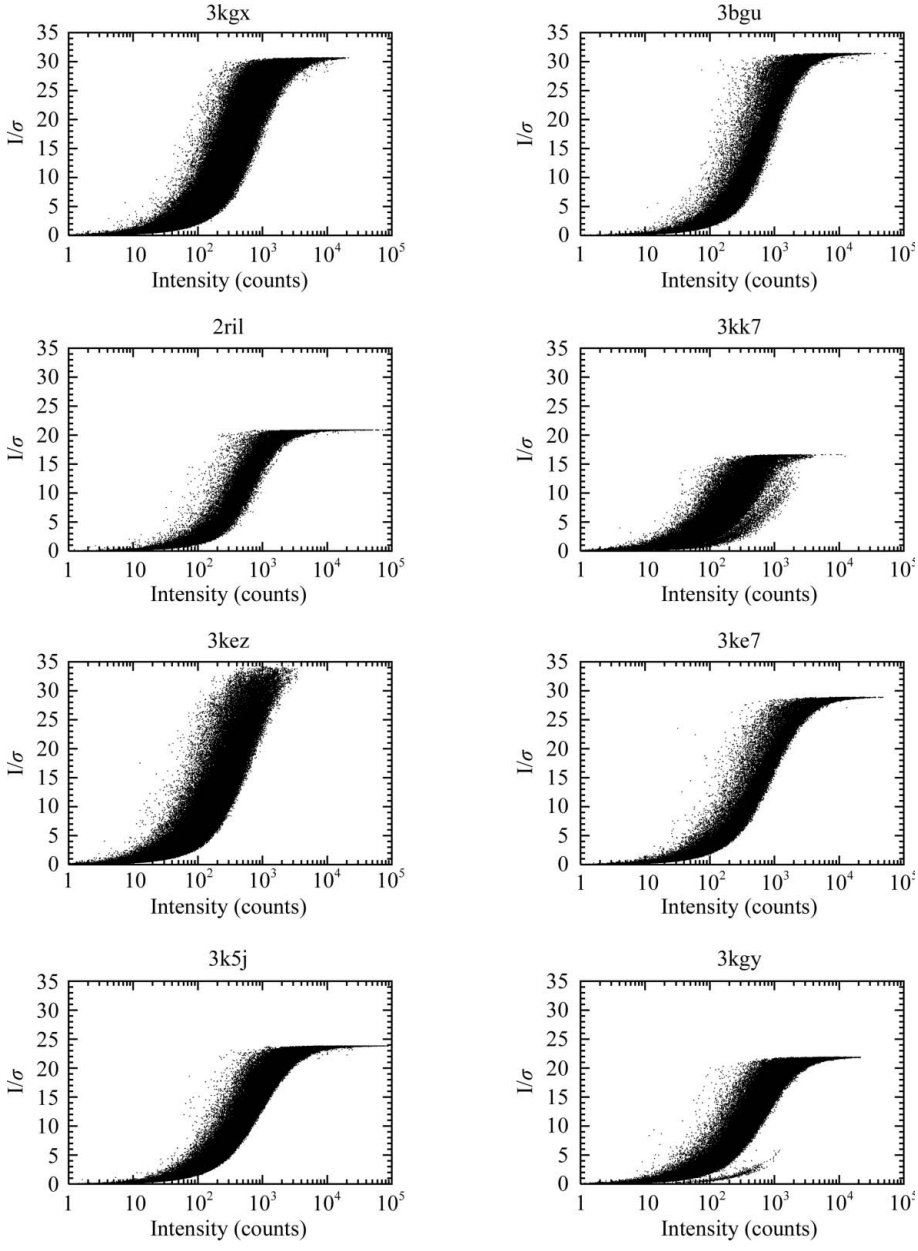

Figure 1
I/σ plots for the first data set of each project listed in Table 1. The data points of 3kk7 and 3kgy which form a weak elongated cloud separated from the strongly populated region belong to reflections near ice rings.

2.2. Processing of experimental data sets

A random selection of eight experimental data sets (Joint Center for Structural Genomics, unpublished work) were downloaded from the JCSG data-set archive (Jaroszewski *et al.*, 2009). These represent successful structure solutions using the molecular-replacement method or experimental phasing by the SAD or MAD method. All data sets were processed with *XDS* (version 28 December 2009) using standard procedures. A single pass of data processing and thus no data-set-dependent optimization of geometric parameters and those describing beam divergence and crystal mosaicity was employed. The resolution of the data sets, as judged from the statistics of data processing, was in good agreement with that reported for the deposited PDB files in all cases.

Some data sets were also processed with *MOSFLM* (Leslie, 1992), with generally similar results (data not shown).

2.3. Simulation of instrument errors

The program *SIM_MX* (Diederichs, 2009) can be used to generate artificial data sets with and without (systematic) instrument errors. For this work, artificial data sets with 1.6 Å resolution were generated using *SIM_MX* with intensities corresponding to an insulin model (PDB code 2bn3; Nanao *et al.*, 2005) in space group $I2_13$ with a = 77.9 Å. The intensity values were calculated in *phenix.refine* (Adams *et al.*, 2010), using anisotropic atoms, added H atoms and a solvent model, and multiplied by 0.01.

Table 1
Statistics of JCSG data sets.

The data sets were collected in the order given, except that the MAD data sets named E1, E2 and E3 (where they exist) were collected in an interlaced fashion, alternating the wavelengths after every 30 frames. The oscillation range was 1°, except for data sets '1' for 3kgx, 3kez and 3k5j, which employed a 0.5° range. The mosaicity (reported as REFLECTING_RANGE_ESD by *XDS*) was 0.09° for 3kgx, 0.24° for 3bgu, 0.18° for 2ril, 0.25° for 3kk7, 0.08° for 3kez, 0.12° for 3ke7, 0.09° for 3k5j and 0.09° for 3kgy. K_1, K_2 and $(I/\sigma)^{\rm asymptotic}$ were calculated from the output of *XDS* using the formulae in §2.1.

PDB code	Synchrotron	Beamline	Date of data collection	Structure solution	Data set	Resolution (Å)	K_1	1000 × K_2	$(I/\sigma)^{\rm asymptotic}$	Exposure time (s)
3kgx	ALS	5.0.3	18 Oct 2009	Molecular replacement	1	1.80	4.9	0.22	30.7	2
3bgu	ALS	8.2.2	4 Oct 2007	MAD	E1	1.50	4.1	0.25	31.5	1
					E2	1.50	4.4	0.21	32.5	1
					2	1.50	3.9	0.35	27.0	1
2ril	ALS	8.2.2	4 Oct 2007	MAD	E1	1.26	4.4	0.52	20.9	1
					E2	1.26	4.3	0.50	21.6	1
					2	1.26	4.4	0.47	21.9	1
3kk7	SSRL	9-2	31 July 2009	MAD	E1	2.46	4.8	0.76	16.6	4
					E2	2.46	4.6	0.89	15.6	4
					E3	2.46	4.8	0.97	14.6	4
					2	2.46	5.0	0.27	27.2	8
					3	2.46	5.5	0.37	22.2	10
3kez	SSRL	9-2	13 May 2009	SAD	1	1.90	4.0	0.21	34.8	5
3ke7	SSRL	11-1	13 May 2009	MAD	E1	1.45	4.1	0.29	29.0	2
					E2	1.45	4.1	0.22	33.1	2
					2	1.45	4.1	0.28	29.3	2
3k5j	SSRL	11-1	8 Jul 2009	SAD	1	1.40	4.9	0.36	23.9	2
					2	1.77	5.8	0.30	24.0	4
3kgy	SSRL	11-1	7 Jul 2009	MAD	E1	1.50	4.4	0.47	22.0	2
					E2	1.50	4.5	0.45	22.2	2
					2	1.50	4.4	0.52	20.8	2

The data sets correspond to measurements at 1 Å wavelength and a crystal-to-detector distance of 150 mm with a MAR345 detector, an oscillation range of 1° and an average background of 30 counts per pixel. The data sets include the very weak anomalous signal of the S atoms at this wavelength.

A data set devoid of systematic error and data sets modified by various types and amounts of instrument error were generated. By default, the crystal mosaicity was specified using a value of 0.1° for both CELL_STDDEV and ORIENTATION_STDDEV; for some calculations 0.2° or 0.4° was used. Data sets by default consisted of 32 frames; for data sets 9 and 10, 64 frames were calculated. In all other respects, default values of the *SIM_MX* program were used unless noted in Table 2.

SIM_MX has two features for assessing the influence of instrument error on simulated data. The first of these simulates non-ideal shutter–spindle synchronization, such as a shutter always opening too late or closing too early. To this end, the program allows modification of the contents of 20 sub-ranges of the oscillation range of each frame. To simulate a 5% error in this category, one sub-range was set to have zero exposure whereas all other sub-ranges had the correct exposure. The expectation was that this error should result in a lower limit for $(I/\sigma)^{\rm asymptotic}$ of around 20 (the reciprocal of 5%).

The other feature offered by *SIM_MX* is a systematic modification of the pixel contents consisting of multiplication of the number of photons on a frame by a sinusoidal function $[f(\varphi) = 1 + c\sin(d\varphi)]$ which depends on the rotation angle φ at which the photons are diffracted. Such a function may be understood as one (sometimes the only) component of a Fourier analysis of the spectrum of fluctuations. The amplitude of modulation is given by the variable c and its period is associated with the variable d. This modification can be used to assess the influence of beam fluctuations of a given period (which need not coincide with the shutter frequency) or periodic mechanical motions of the diffractometer or crystal.

The synthetic data sets were processed with *XDS*. The correlation coefficient between signed anomalous differences of intensities referring to random subsets of multiple observations (Schneider & Sheldrick, 2002) was obtained from the *XDS* output file `CORRECT.LP` for the low-resolution shell (50–4.76 Å).

3. Results

3.1. I/σ values of experimental data

The $(I/\sigma)^{\rm asymptotic}$ values of the JCSG data sets, which were collected on ADSC and MAR CCD detectors, are in the range 14–35 (Table 1). If several data sets (often representing different wavelengths) were collected for a given project, their $(I/\sigma)^{\rm asymptotic}$ values usually only show variations of up to about 10%. The exception is 3kk7, which will be discussed below (§4.3).

The K_1 values are in the range 4–5. This is owing to the fact that the point-spread function of CCD detectors extends over several pixels, which means that the gain values for individual pixels (obtained in the INIT step of *XDS*) underestimate the gain value required to obtain accurate variances of the reflection intensities (Leslie, 2006). K_1 compensates for these errors in gain. It may be noted that K_1 values near unity are

Table 2
Statistics of simulated data sets.

The default values are given in the text.

Type of calculation	Amount of instrument error simulated	Synthetic data set No.	Modification with respect to defaults	K_1	$1000 \times K_2$	$(I/\sigma)^{\text{asymptotic}}$
Error-free	None	1	—	0.98	0.04	160.9
Shutter/spindle de-synchronization	No signal during last 5% of oscillation	2	—	0.72	2.89	21.9
	No signal during last 5% of oscillation	3	Double mosaicity	0.93	0.33	57.0
	No signal during last 10% of oscillation	4	Double mosaicity	0.92	1.57	26.3
	No signal during last 5% of oscillation	5	Tenfold flux	1.21	2.04	20.1
	No signal during last 5% of oscillation	6	Double distance	0.11	26.4	18.4
	No signal during last 5% of oscillation	7	Resolution cutoff at 3 Å	0.67	3.97	19.3
	No signal during last 5% of oscillation	8	Crystal rotated by 45° around a general axis	0.79	2.58	22.2
	No signal during last 5% of oscillation	9	Double number of frames	0.79	2.96	20.7
	No signal during last 5% of oscillation	10	Half of oscillation range; double number of frames	0.86	2.93	19.9

usually obtained for the Pilatus (Henrich *et al.*, 2009) silicon pixel detector (data not shown), which has a very narrow point-spread function.

Fig. 1 shows plots of I/σ values as a function of intensity. One data set (3kez) was weakly exposed and the limit of I/σ is not clearly visible. The other seven plots display a sigmoidal curve: the (I/σ) of the low-intensity reflections (up to an intensity of about 100 counts) is most affected by background and therefore does not depend strongly on I, whereas in the range of intensities up to about 500 counts I/σ rises with I. For large I, I/σ approaches a limiting value consistent with the $(I/\sigma)^{\text{asymptotic}}$ value calculated from K_1 and K_2 (6). This means that despite their high quality these data sets were noticeably compromised by systematic errors. Quantitatively, the asymptotic value of I/σ is approached for intensity values above about 500 counts and a large fraction of all reflections is affected in these data sets.

3.2. Systematic errors related to shutter–spindle de-synchronization

A calculation of synthetic data without systematic error (data set 1 in Table 2) was used to establish the highest value of $(I/\sigma)^{\text{asymptotic}}$ that can be reached by simulation from a crystal with a mosaicity of 0.1°. For the data simulated here, the value obtained was 160.9. All other calculations included systematic errors such that the resulting $(I/\sigma)^{\text{asymptotic}}$ values encompass the range relevant to experiments at a synchrotron, which is far below this value, which could be considered ideal. In most calculations, the K_1 values are near unity, as no detector point-spread function is currently simulated by *SIM_MX*. The exception is synthetic data set 6 which, owing to a doubled crystal-to-detector distance, lacks the weak reflections required to properly define K_1 in the least-squares fit; in this case only $(I/\sigma)^{\text{asymptotic}}$ is meaningful.

After establishing a baseline value of 21.9 for $(I/\sigma)^{\text{asymptotic}}$ in a calculation without modification of the default values (synthetic data set 2), eight calculations (synthetic data sets 3–10) were performed to test the influence of several aspects of the data-collection strategy. Doubling the mosaicity of the crystal (synthetic data set 3) increased $(I/\sigma)^{\text{asymptotic}}$ to 57. On the other hand, increasing the flux (tenfold exposure in synthetic data set 5), doubling the distance (synthetic data set 6), cutting the resolution to 3 Å (synthetic data set 7), changing the orientation of the crystal (synthetic data set 8), doubling the number of frames and thereby increasing the multiplicity (synthetic data set 9) or using half of the oscillation range (synthetic data set 10) did not change $(I/\sigma)^{\text{asymptotic}}$ in a significant way. A change of wavelength was not simulated, as none of the systematic errors is explicitly wavelength-dependent.

3.3. Systematic errors related to modulations

Fig. 2 shows plots of the data simulating modulations. For both low and high periods, the $(I/\sigma)^{\text{asymptotic}}$ values reach the level of the 'ideal' data. In the range 0.01–1 oscillations per second, however, the $(I/\sigma)^{\text{asymptotic}}$ values drop to values below 10, with a minimum of 7.5. The minimum value of $(I/\sigma)^{\text{asymptotic}}$ that is expected in such a calculation is the reciprocal of the average deviation value of $1 + c\sin(x)$ from unity. With $c = 0.2$, a value of $1/(0.2 \times 2/3.14159) = 7.85$ results, which is consistent with the results obtained.

The explanation of these results is the following: for sufficiently long periods of the modulation it can be compensated in the data-reduction software by a scale factor for the whole frame. Similarly, for periods that are sufficiently short compared with the mosaicity of the crystal, the fluctuation at the level of integrated intensities averages out. However, the range 0.01–1 oscillations per second is of practical importance for data collection at a synchrotron. Within this range, the period of the fluctuation (measured in degrees of spindle rotation) is comparable to the mosaicity of the crystal. Consequently, the intensity estimates are strongly affected by the modulation.

3.4. Influence of instrument error on experimental phasing prospects

All JCSG data sets with $(I/\sigma)^{\text{asymptotic}}$ values of 25 and higher allowed straightforward structure solution by experimental phasing (data not shown). Unfortunately, to the best of the author's knowledge experimental data sets which did not allow structure solution are not available from the JCSG data archives.

Synthetic data allow the evaluation of the impact of deteriorated data quality. Synthetic data sets 1 and 3, with their high $(I/\sigma)^{\rm asymptotic}$, could indeed be used for SAD structure solution using *SHELXD*/*E* (Sheldrick, 2008), whereas all of the other data sets did not lead to substructure solution and phasing. However, being able to solve a structure is a qualitative feature of a data set. To obtain a quantitative relation between $(I/\sigma)^{\rm asymptotic}$ and a parameter that has been shown to be strongly related to the prospect of solving a structure with SAD, SIR or MAD, the correlation coefficient between signed anomalous differences within a data set (Schneider & Sheldrick, 2002) was calculated for these synthetic data sets. Fig. 2 shows that $(I/\sigma)^{\rm asymptotic}$ and this anomalous correlation coefficient, which was computed for reflections to 4.76 Å resolution, both depend in the same way on the amount of systematic error introduced.

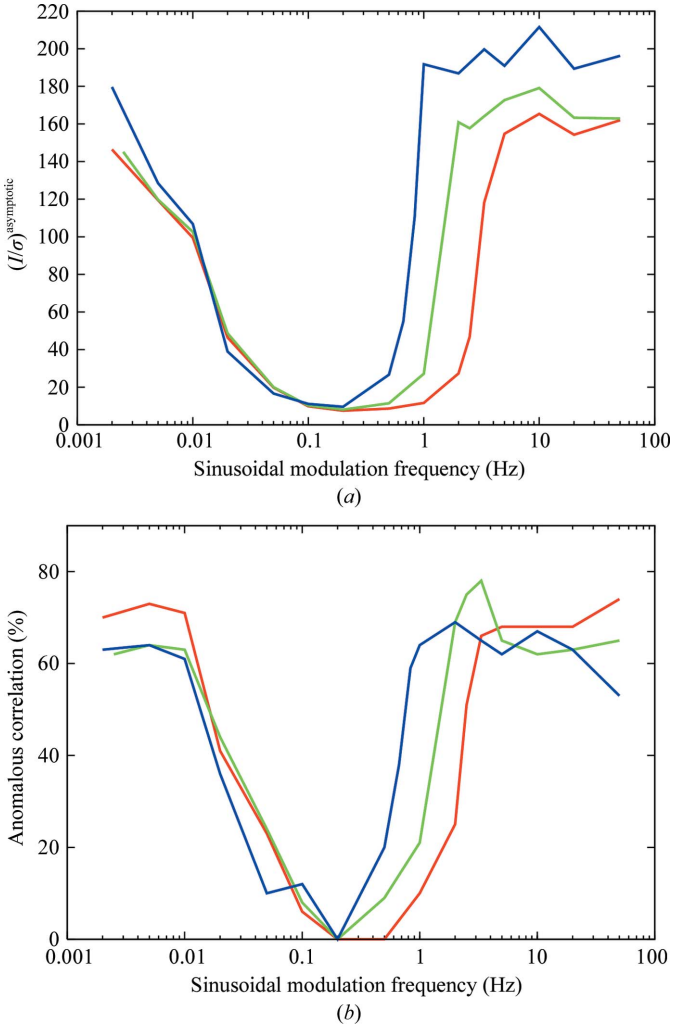

Figure 2
(*a*) Plot of $(I/\sigma)^{\rm asymptotic}$ as a function of the frequency of the harmonic modulation of intensities (*SIM_MX* parameter MODULATION_IN_PHI). The amplitude of the modulation was set to 20% (c = 0.2). The frequency given is relative to 1 s exposure of frames with an oscillation range of 1°; halving the speed of the spindle would have the same effect as doubling the modulation frequency. The variation of the (data set-internal) anomalous correlation coefficient in the 50.0–4.77 Å resolution shell is shown in (*b*). Red, mosaicity of 0.1°; green, 0.2°; blue, 0.4°.

Fig. 3 shows that the anomalous correlation coefficient for the synthetic data with its very weak anomalous signal is higher than 20% whenever $(I/\sigma)^{\rm asymptotic}$ is higher than 20. For values of $(I/\sigma)^{\rm asymptotic}$ less than 10 the anomalous correlation coefficient is negligible and for $(I/\sigma)^{\rm asymptotic}$ values larger than 50 it approaches a value of 74%, the average of the resolution shell.

These values refer to a particularly difficult case, which simulates the anomalous signal of sulfur at a wavelength of 1 Å. For more typical SAD or MAD phasing projects the anomalous signal is usually stronger. Over the years, the author has solved several SeMet MAD structures with $(I/\sigma)^{\rm asymptotic}$ values of 15 or higher; however, whenever $(I/\sigma)^{\rm asymptotic}$ was lower than 20 phasing and model building were usually difficult. These experiences refer to a multiplicity of about 3–4 (for each member of a Friedel pair). The author has also seen cases of good crystals giving $(I/\sigma)^{\rm asymptotic}$ values of 10 or lower. In these cases, radiation damage usually destroyed the crystals before a useful anomalous signal could be obtained from averaging of multiple observations.

4. Discussion

Possible sources of error in the intensity values in an X-ray experiment are the crystal itself, the experimental setup and the software that processes the data. The total error in the scaled intensity of a specific reflection may be split up into a random part (arising from counting statistics and amenable to reduction by averaging of multiple observations) and a systematic part (which has either a lowering or an augmenting effect on a particular observation). Whereas the random component is associated with the scattering of the crystal, the X-ray flux of the beamline and noise from the detector and electronics, the systematic component may be subdivided into the software part, the crystal part and the instrument part

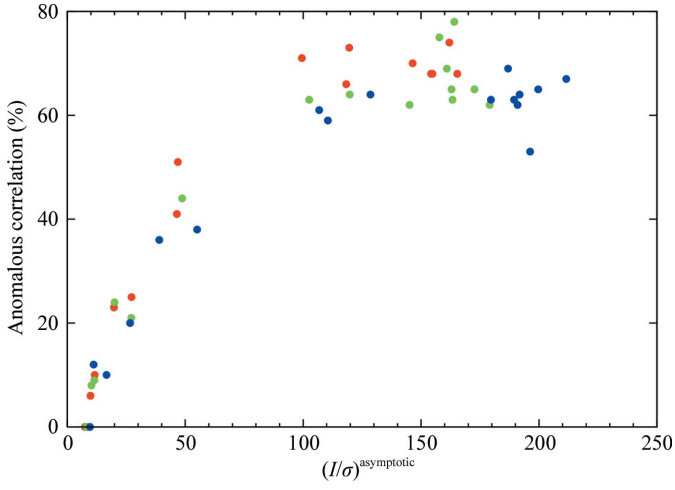

Figure 3
Plot of the relation between anomalous correlation coefficient and $(I/\sigma)^{\rm asymptotic}$ for the synthetic data sets as a function of the frequency of the harmonic modulation of intensities. Colours correspond to those in Fig. 2.

(beamline, diffractometer, detector, cryocooling), which will be discussed in this order.

The goal of the discussion is to judge the relative amounts of these contributions and to dissect their roles, in particular with respect to $(I/\sigma)^{\rm asymptotic}$. It will be seen that for a given mosaicity the value of $(I/\sigma)^{\rm asymptotic}$ mainly depends on the instrument error.

$(I/\sigma)^{\rm asymptotic}$ was chosen as the focus of the discussion owing to its property as a limiting value of the signal-to-noise ratio of the strongest reflections. Alternatively, its reciprocal value may be considered as the systematic component of the error contributing to the internal R factor of the data, as $R_{\rm merge} \simeq 0.7979\langle\sigma/I\rangle$ (Stout & Jensen, 1968). Thus, if χ^2 is near 1, the lower limit of the low-resolution $R_{\rm merge}$ is close to about $0.8/(I/\sigma)^{\rm asymptotic}$, and its actual value is increased beyond that by random error.

4.1. Systematic error contributions from software

Any investigation into $(I/\sigma)^{\rm asymptotic}$ values of experimental and synthetic data can only be meaningful if the data processing itself produces little error compared with that resulting from the experimental setup. In theory, the value of $(I/\sigma)^{\rm asymptotic}$ for 'perfect' synthetic data, *i.e.* those derived from a model that agrees with that of the data-reduction software, should approach infinity. In practice, perfect data are impossible to generate and likewise data reduction cannot have perfect results, *e.g.* owing to the finite accuracy of computer calculations. Furthermore, approximations and compromises in both the data-producing and data-processing software preclude perfect data being obtained in a simulated experiment.

In the case of *SIM_MX*, the number of X-rays traced is finite, all mosaic blocks are assumed to have the same size and the unit-cell parameters, rotation angles and wavelengths are assumed to be normally distributed around their means. Data-reduction software has to decide which pixels contribute to a reflection and which pixels contribute to the background, and the choice may not be completely consistent with the nature of the simulated data. In addition, bugs may exist in any piece of software and calculations with synthetic data can reveal such bugs. However, there are no known bugs in *SIM_MX* or *XDS* at the time of writing. The large value of $(I/\sigma)^{\rm asymptotic}$ obtained with data devoid of systematic error confirms that the data-reduction program itself introduces little systematic error.

4.2. Systematic error contributions from the crystal

Radiation damage may become the biggest contribution to systematic error if it is allowed to destroy the diffraction pattern of the crystal. Analysis of $(I/\sigma)^{\rm asymptotic}$ for partial data sets offers a simple way to detect radiation damage and may suggest a suitable cutoff on the number of data frames to be processed. For the experimental data sets investigated in §3.1 radiation damage plays a subordinate role: if parts of the data sets are processed (first quarter, first half or first three quarters of all frames) the resulting $(I/\sigma)^{\rm asymptotic}$ values are similar to (but up to 10% higher than) those obtained with the full data sets (data not shown).

In principle, absorption in the crystal and its mounting device (loop) may also contribute to the overall systematic error. However, the currently available data-reduction software has undergone decades of development to properly model these effects; therefore, it may be assumed that the remaining systematic error is small.

Some crystals have defects that are apparent in the shape of their reflections, which may appear to be split or smeared and are difficult to model and profile-fit. Similarly, ice rings result in lowered $(I/\sigma)^{\rm asymptotic}$ values because the integrated intensity of symmetry-related reflections may be affected slightly differently. The scaling procedure takes these intensity variations into account by adjustment of K_1 and K_2, which in turn raises the σ values of all reflections.

Crystals with strong radiation damage, defects in their reflection profiles or ice rings will therefore exhibit a low $(I/\sigma)^{\rm asymptotic}$ and cannot be used to quantify instrument error. However, as long as the profiles of the reflections (even of high-mosaicity crystals) can be adequately modelled by the software and radiation damage is low, the contribution of the crystal to the total systematic error should not be significant. Therefore, in practice any fresh crystal with a clean diffraction pattern that covers the detector is suitable for assessing instrument error, because space-group symmetry provides an ideal internal calibration of the diffraction experiment.

4.3. Systematic error contributions from the instrument

For a given amount of instrument error, it is shown with synthetic data in §3.2 that the $(I/\sigma)^{\rm asymptotic}$ value does not depend on flux, oscillation range, distance, crystal orientation, number of frames or resolution. The experimental data in §3.1 and Table 1 show that it does not depend on wavelength (for MAD data sets) or oscillation range (for 3k5j) either. The only remaining variable is therefore spindle speed. As Figs. 2 and 3 suggest, a reduction in systematic error may be achieved by increasing the exposure time per degree of spindle rotation, as this will smooth out some of the high-frequency systematic errors introduced by the experimental setup.

One of the JCSG projects demonstrates this effect. In the case of 3kk7, three data sets (E1 = peak, E2 = high-energy remote, E3 = inflection) were collected with 4 s exposure and displayed low $(I/\sigma)^{\rm asymptotic}$ values. Long exposure times in data sets 2 and 3, which were also collected at the peak wavelength, increased the $(I/\sigma)^{\rm asymptotic}$ value considerably. However, data set 3, which was collected last, exhibits a lower $(I/\sigma)^{\rm asymptotic}$ value than data set 2, presumably owing to radiation damage.

However, two caveats have to be considered. Firstly, any increase in dose rate leads to increased radiation damage and thus a decrease in spindle speed usually has to be compensated for by an increase in beam attenuation. Secondly, the improvement of $(I/\sigma)^{\rm asymptotic}$ depends on the mosaicity of the crystal: according to Table 2, a higher mosaicity may mask instrument error. In other words, instrument errors lead to less

error in the intensities of high-mosaicity crystals than of low-mosaicity crystals. This conversely means that to quantify instrument errors, low-mosaicity crystals such as those commonly used for beamline calibration (*e.g.* thaumatin, lysozyme and insulin) are more suitable than high-mosaicity crystals.

Another instrument error is detector nonlinearity. This source of systematic error may be quantified by monitoring $(I/\sigma)^{\rm asymptotic}$ in a series of frames with increasing flux and may be analyzed by comparing the resulting reduced data sets.

5. Conclusions

Crystallographers tend to believe that the largest contribution to σ is from counting statistics, that an increase in dose should raise the maximum I/σ value and that the maximum I/σ value of their data sets results from a limitation of their crystals. However, this work demonstrates that this is usually not the case. Rather, the data and simulations presented here suggest that high multiplicity (*e.g.* for experimental phasing) is only required because systematic instrument error usually prevents single observations to be measured sufficiently accurately.

The experimental data used in this study were a purely random selection from the JCSG data-set archive. As Fig. 1 shows, the accuracy of the data (measured as I/σ) obtained from these macromolecular crystals is limited by $(I/\sigma)^{\rm asymptotic}$. The saturation of I/σ of the strongest reflections in seven of the eight plots suggests that a better data-collection strategy might have been to increase the multiplicity while maintaing the total dose to avoid the inflation of the systematic error term which limits the I/σ of individual observations. This would have resulted in higher I/σ values after averaging.

This gedankenexperiment demonstrates that knowledge of the amount of systematic error arising from the instrument would allow the determination of better data-collection strategies. Thus, a program such as *BEST* (Popov & Bourenkov, 2003; Bourenkov & Popov, 2006), which in its published form assumes a fixed 3% systematic error contribution [equivalent to a $(I/\sigma)^{\rm asymptotic}$ of 33.3], may be improved by taking an experimentally determined $(I/\sigma)^{\rm asymptotic}$ value into account.

At low multiplicity, high values of $(I/\sigma)^{\rm asymptotic}$ are not a sufficient condition for solving a structure. Rather, they are a required condition; in addition, the random error component needs to be low enough and the phasing power high enough. Low $(I/\sigma)^{\rm asymptotic}$ may be compensated by high multiplicity at the expense of radiation damage.

The $(I/\sigma)^{\rm asymptotic}$ value may be measured for a range of experimental situations using test crystals. As shown in §3, these values are influenced, for certain kinds of instrument error, by the mosaicity of the crystal. Furthermore, the $(I/\sigma)^{\rm asymptotic}$ values strongly depend on the amplitude and frequency of any vibrations or fluctuations (modulations) in the experimental setup. Systematic measurements may therefore help to pinpoint and eliminate or reduce certain types of instrument errors at a beamline.

An important practical conclusion from these results is that an experimenter may, in the unavoidable presence of modulations, optimize the data collection by lowering the spindle speed and attenuating more strongly and thereby transform all modulations arising from the experimental setup into the regime with high $(I/\sigma)^{\rm asymptotic}$ and high anomalous correlation values on the right-hand side of Fig. 2.

In summary, $(I/\sigma)^{\rm asymptotic}$ conveys an important aspect of data quality for both the beamline scientist and the beamline user and the values of $(I/\sigma)^{\rm asymptotic}$ for individual data sets should be routinely evaluated, recorded and reported in papers about X-ray structure solutions.

The author thanks the JCSG for making their data sets available.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Bevington, P. R. (1969). *Data Reduction and Error Analysis for the Physical Sciences.* New York: McGraw–Hill.
Borek, D., Minor, W. & Otwinowski, Z. (2003). *Acta Cryst.* D**59**, 2031–2038.
Bourenkov, G. P. & Popov, A. N. (2006). *Acta Cryst.* D**62**, 58–64.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z. (2010). *Acta Cryst.* D**66**, 389–392.
Diederichs, K. (2009). *Acta Cryst.* D**65**, 535–542.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Henrich, B., Bergamaschi, A., Broennimann, C., Dinapoli, R., Eikenberry, E. F., Johnson, I., Kobas, M., Kraft, P., Mozzanica, A. & Schmitt, B. (2009). *Nucl. Instrum. Methods Phys. Res. A*, **607**, 247–249.
Jaroszewski, L., Li, Z., Krishna, S. S., Bakolitsa, C., Wooley, J., Deacon, A. M., Wilson, I. A. & Godzik, A. (2009). *PLoS Biol.* **7**, e1000205.
Kabsch, W. (2010*a*). *Acta Cryst.* D**66**, 125–132.
Kabsch, W. (2010*b*). *Acta Cryst.* D**66**, 133–144.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696–1702.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.
Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* D**61**, 1227–1237.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Stout, G. H. & Jensen, L. H. (1968). *X-ray Structure Determination. A Practical Guide*, p. 402. London: Macmillan.