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

# Some notes on choices in data collection

**Philip R. Evans**

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Correspondence e-mail:
pre@mrc-lmb.cam.ac.uk

Collecting optimum X-ray diffraction data involves a number of choices and compromises, including choice of crystal, source, rotation range, exposure time and programs for integration and scaling. This paper presents a series of questions which should be considered in planning a data-collection experiment.

Received 2 July 1999
Accepted 25 August 1999

## 1. Introduction

As a summing up of the CCP4 Study Weekend on data collection and processing, here are a few points to think about and choices to make in order to collect good data.

## 2. Which source?

The most important properties of a source are intensity, divergence, beam size and spectral distribution (wavelength and dispersion). The ideal source matches the properties of your crystal, as described by Nave.

### 2.1. Intensity

A strongly diffracting crystal does not need the most powerful beam. Many good structures have been solved with rotating anodes: a good combination is a high-resolution native collected at a synchrotron with derivatives collected at home. On the other hand, small weakly diffracting crystals may need the brightest source you can find.

### 2.2. Divergence

Low divergence is generally better but will not help with a crystal of high mosaicity.

### 2.3. Beam size

A large beam is wasted on a small crystal and will produce unwanted air scatter. The beam should therefore not be significantly bigger than the crystal. If the beam is smaller than the crystal, different volumes of crystal may be in the beam depending on the orientation of the crystal; this will adversely affect the relative intensity of reflections (the intensity of a reflection is proportional to the diffracting volume).

### 2.4. Wavelength and dispersion

A small dispersion is required for MAD, but will reduce intensity. Ideally, tune the wavelength to optimize the anomalous signal, but do not neglect the use of anomalous phasing at non-optimum wavelengths (*e.g* Cu $K\alpha$).

## 3. Which crystal?

### 3.1. Single

Your crystal should not be multiple, split or twinned [note that twinning refers to two (or more) lattices with an exact geometric relationship between them, and should not be used loosely to refer to split crystals]. Look at a few images, preferably at least two which are 90° apart, and try to index them. Multiple crystals should be apparent. Twins are much less obvious (Chandra).

### 3.2. Mosaicity

Low mosaicity is better than high mosaicity. Mosaicity may vary between otherwise identical crystals, even ones grown from the same crystallization drop.

### 3.3. Temperature

Frozen crystals (strictly, supercooled crystals) generally produce better data than those at room temperature, but the cryoprotection and freezing protocol must be optimized to minimize increases in mosaicity and ice formation. The best method, where possible, is to grow crystals in the cryoprotectant (Garman).

### 3.4. Quality

Crystals which diffract to high resolution are better (*i.e.* those with the smallest $B$ factor). Again, crystals may vary and sub-optimal freezing may damage them.

### 3.5. Background

Low background improves the signal-to-noise ratio. Unless your crystals are unusually fragile, minimize the amount of liquid around the crystal, *e.g.* by using a small loop. For large robust crystals, the loop can be smaller than the crystal. Thin fragile crystals, on the other hand, need a large loop.

### 3.6. Size

Large crystals are better than small crystals, other things being equal. However, small crystals may freeze better, though large crystals grown in cryoprotectant will often freeze well. The diffracted intensity is proportional to the number of unit cells in the beam. In difficult cases, you may have to screen many crystals to find a good one.

## 4. What strategy?

Strategy has been covered extensively by Dauter and by Pflugrath in this issue.

### 4.1. Redundancy

High redundancy produces more accurate data and allows for reliable rejection of outliers. It is an ancient principle of accurate measurement to measure something many times and take the average. With a fast read-out detector such as a CCD, collection of 180° or even 360° of data is reasonably fast and this also simplifies strategy.

### 4.2. Completeness

Completeness, both in geometrical coverage of reciprocal space and the full intensity range is very important. Systematic omission of data will distort all maps. The geometric strategy may be complicated if the detector is not centrally placed on the beam; however, strategy simulations are available in a number of programs and should be used.

### 4.3. Resolution

The maximum resolution of a data set may be reset after examination of data-reduction statistics. To collect the data, the detector may be positioned a little closer than the apparent maximum resolution, provided that the spots are resolved.

### 4.4. Exposure

Exposure time needs to be set to long enough to give reasonable statistics at the highest resolution, but not so long as to overload the detector with the strong low-angle spots, nor to give too much radiation damage. More than one pass with different exposure times may be required to catch the full dynamic range of data.

### 4.5. Width

Rotation width per image should be set to resolve the longest axis on rotation (Dauter), taking into account the reflection width. Narrower image widths may improve data quality, as discussed by Pflugrath.

## 5. Which integration program?

A number of integration programs are in use and all seem to produce good data when used properly (*i.e.* when following the instructions). A major choice is between two- and three-dimensional integration. At present, most people use two-dimensional programs, but there is a case for three-dimensional methods (Pflugrath).

There is still room for improved methods: an ideal program should consider the images as slices of a true three-dimensional reciprocal space, rather than separate samples. A full three-dimensional analysis would then allow deconvolution of overlapping spots in three dimensions and better treatment of features other than Bragg diffraction.

## 6. What scaling?

Scaling attempts to correct for systematic errors by refining a scaling model in order to make repeated measurements of symmetry-related reflections equal. The scaling model may be as simple as one scale per image or as complex as a full three-dimensional pseudo-absorption correction, and may also incorporate attempts to fix mistakes introduced by the integration program ('post-refinement'). The scaling model should reflect the experiment. Thus, if there are no discontinuities in the experiment, then a smooth correction function should be used. On the other hand, if the beam intensity may change suddenly between images (as on some synchrotrons), then a separate scale factor for each image is appropriate. The corrections applied should be physically reasonable.

For MIR and MAD phasing, intensity differences are more important than absolute intensities, and relative scaling between data sets can reduce the systematic errors in the differences.

## 7. What statistics?

Are the data good enough for the purpose required? There is no shame in throwing away bad data and recollecting it.

What is the real resolution? $R_{\rm merge}$ is a very poor guide to data quality, as it takes no account of multiplicity. $\langle I\rangle/\sigma(\langle I\rangle)$ is a much better indicator, provided that the standard uncertainties have been validated by $\chi^2$ or $T$ tests. A typical guideline is to cut the resolution where $\langle I\rangle/\sigma(\langle I\rangle) = 2.0$, but data beyond this may be used with maximum-likelihood methods.

## 8. What can go wrong?

Data collection is your final experiment, on which you will base a good deal of work. It is important to be careful and to be aware of the assumptions and of what can go wrong, even if you yourself have no control over some aspects of the experiment (as at synchrotrons).

### 8.1. The beam

The incident X-ray beam must be stable in intensity and wavelength (or at least vary only slowly in intensity).

### 8.2. The goniostat and shutter

The goniometer rotation must be regular and accurate and synchronized with the shutter opening and closing. This synchronization is particularly demanding on very intense beamlines where an error of a few milliseconds may lead to gaps or overlaps between adjacent images. The goniometer axis should be aligned properly with the beam (*i.e.* should be perpendicular).

### 8.3. The detector

The detector must be stable and calibrated, both for spatial distortion and for non-uniformity of response, and any 'dark current' must be measured and stable. Overloads should be noticed and flagged.

All of these things have gone wrong in experiments and might in yours.