A beginner’s guide to X-ray data processing

Harold R. Powell
(Imperial College London, UK)

Single crystal X-ray diffraction is currently the most popular method for accurately determining the structure of molecules and is used for samples containing a few to many thousands of atoms in both biology and chemistry. Following data collection, the individual diffraction images need to be processed into a single dataset containing intensity information about all the contributing X-ray reflections.

Introduction

After a single crystal of a protein has been used in an X-ray diffraction experiment, the scientist has a set of images that contain the raw data from which an accurate molecular structure of their protein may be determined. Up to this point, the scientist will have performed practical experiments, but after this, everything is purely computational and each step can be repeated ad nauseam using different programs. Because data collection is the last experimental step in the process, it is well worth putting thought, time and effort into obtaining the best possible diffraction data.

Although data processing is performed automatically at most synchrotrons and the scientist is provided with the information necessary for structure solution, it can be helpful to know what the different stages are and what they are actually doing.

Data processing itself consists of two basic stages, ‘integration’ and ‘scaling and merging’. Integration may be sub-divided into four steps, while scaling and merging is usually considered as a single step. While the integration step gives a lot of useful information, the scaling and merging step is what provides the most useful statistics about the quality of both data collection and data processing.

The four steps involved in integration are:
1. spot finding
2. indexing
3. parameter refinement
4. integration itself.

A brief note about crystals and diffraction

A crystal is a three-dimensional array (or ‘lattice’) containing its substituent molecules, atoms or ions. The fundamental building block, called the ‘unit cell’, is translationally repeated in three dimensions to form the crystal. Unit cells for protein crystals typically have sizes ranging from ~30 Å to ~1000 Å across, where an Ångström is 10^{-10}m; this is a useful unit for describing molecular structures because typical covalent bond lengths are in the range 1–2 Å.

When a crystal is illuminated with radiation with about the same wavelength as the unit cell dimensions (actually, within several orders of magnitude; X-rays with a wavelength of ~0.5–2 Å are used to study crystals with unit cells from <10 Å to >1000 Å across), the crystal acts as a diffraction grating and the scattered radiation will interfere constructively in a few directions but destructively in most. Diffraction images that result from this have regularly spaced spots (of different intensities) superimposed on a low background.

The unit cells that make up crystals have seven fundamental shapes that can be distinguished by the relationships between their edges; all are based on what can be thought of as a box with three edges of length a, b and c, and angles between them α, β and γ. The seven crystal systems are:

- triclinic: a≠b≠c, α≠β≠γ
- monoclinic: a≠b≠c, α=γ=90°, β≠90°
- orthorhombic: a≠b≠c, α=β=γ=90°
- tetragonal: a=b≠c, α=β=γ=90°
- hexagonal/trigonal: a=b≠c, α=β=90°, γ=120°
- rhombohedral: a=b=c, α=β=γ=90°
- cubic: a=b=c, α=β=γ=90°

where ‘=’ means ‘must be exactly’ and ‘≠’ means ‘does not have to be equal to’.

In addition to the seven crystal systems, there may be extra purely translational symmetry in the unit cell that is known as ‘centring’ (an absence of centring gives a ‘primitive’ unit cell); this gives 14 Bravais lattices. As well as possible centring, there can also be extra symmetry within the unit cell arising from rotations, rotations combined with translations (‘screws’), reflections, reflections combined with translations (‘glides’) and inversions. Taken together, these give a possible 230 three-dimensional ‘space groups’, although only 65
of these are found for crystals of naturally occurring macromolecules like proteins or nucleic acids – those that do not have either mirror or inversion symmetry.

This additional symmetry means that (in most cases) there will be more than one copy of the molecule in the unit cell, and it gives us information that can be useful when solving the crystal structure. It also means that for some crystal systems the intensities of some reflections should be identical; e.g., for monoclinic crystals, the \((h,k,l)\) and \((-h, k, -l)\) reflections are equivalent but this is not the case for triclinic. These are known as ‘symmetry-equivalent’ reflections.

### Integration

This is the process of extracting the intensity information for the diffraction spots on the images. In order to do this, the location of each spot must be known accurately on each image.

### Spot finding

The first step in integration is to find a selection of spots on a selection of images so that they can be indexed. Some packages (e.g., DIALS, XDS) use spots from all the images so that they have extra information available at later stages of integration but others use spots from either a single image (HKL2000) or a few (Mosflm) and still get enough information to index successfully in the vast majority of cases.

For an experienced crystallographer, even if an image is of low quality, it is usually straightforward to pick out the diffraction spots from any background and also to identify rogue spots (that are not associated with the crystal under investigation) or even to determine that there are multiple lattices represented by the diffraction. Computer programs, however, need to find the spots \emph{ab initio} every time; they use information that is common to most diffraction patterns, for example:

1. Spots near to each other will have similar shapes on each image
2. There are strong, medium and weak spots – only a sample of spots is needed to index, so an intensity cut-off can be applied to help avoid noise
3. Useful spots will be well separated from each other, so can be accurately located.

### Indexing

As the name implies, this process gives a 3D index that uniquely defines each reflection in the dataset with three indices – \(h, k\) and \(l\), commonly referred to as ‘Miller indices’.

Indexing also gives good approximations to the crystal’s unit cell dimensions, its orientation relative to the diffractometer and usually an indication of the crystal symmetry. Successful indexing relies on the direct beam position on the detector, the crystal to detector distance and the wavelength of the radiation to be known accurately; inaccuracies in any of these give rise to most failures of this stage in processing.

A further piece of information that can be obtained at this point, the mosaicity, is related to the structure of the crystal itself. Crystals are rarely 100% homogeneous and perfect; they tend to be made up of assemblies of smaller crystallites (typically 10–100 \(\mu\)m) that are slightly misaligned to each other. The consequence of this is that diffraction spots do not occur instantaneously for a single, precise orientation of a crystal, but appear and disappear through a small range of orientations. For fine-sliced data collection, all reflections will be spread over a series of consecutive images, but the partially recorded spots (or ‘partials’) arising from each reflection will share a common index. Crystals with mosaics less than about 0.5° are the most useful.

### Parameter refinement

Partly because indexing uses a number of approximations in its calculations, the parameters it reports are also approximate; because measuring the intensity of the spots requires each to be accurately located on each image, all the values for the crystal and diffractometer need to be optimized.

Two complementary methods are in use for refinement – positional and post-refinement. Convenienly, they can be used to refine parameters independently that would otherwise be closely correlated, e.g., refinement of the crystal to detector distance is closely linked to that of the unit cell dimensions in positional refinement but not in post-refinement.

Positional refinement (Figure 1a) minimizes the difference between the observed and calculated positions of spots on the detector; this is simple and easy to visualize. Post-refinement (Figure 1b) relies on the measurement of intensities of the different partials belonging to the same reflection, and using this information to determine the precise centre of the reflection.

Following refinement we should have accurate (close to the true value) and precise (small standard deviation) values for the unit cell parameters, the crystal orientation angles (with respect to the diffractometer) and the various diffractometer parameters.
Integration itself

This is the measurement of spot intensity. Because the background noise on an image is spread across the image and also occurs in the same places as the spots (often referred to as ‘being under the spots’), the pixel counts associated with a spot also include a contribution from the background; this has to be removed. The usual way
Beginner’s Guide

To do this is to calculate a background plane from pixels around each spot, then interpolate the values that should be expected underneath the spot.

There are two common ways of approaching integration and two methods of optimizing the measurement.

Integration may be carried out by measuring the intensities on each image separately and storing the information for later; this is known as 2D integration and is the method used in MOSFLM and originally in HKL2000.

Historically, it made sense to collect as many spots on each image as possible, using a wide slicing strategy (i.e., oscillation or rotation range for each image ≥ 1°). This meant that most reflections were fully recorded as spots on single images ('fulls'), with only a few partials, and so measuring the data image-by-image was the best you could do.

For narrow sliced data (rotation range per image ≤ 0.5°), all reflections are present as partials. This presents the opportunity to consider partials on consecutive images as parts of the same measurement. A 'shoe-box' can thus be extended across a series of images for a single reflection and all the measurements can be performed in a single step. The raw pixel information for each spot on the contributing images is stored, and only processed once the reflection is finished. This is 3D integration and is the method used by DIALS and XDS, as well as most commercial packages (HKL2000 has what is called a 'pseudo-3D' mode). Although 3D integration might be expected to yield more accurate results, in practice there is little difference in the results.

For strong reflections, only a tiny proportion of the intensity will be obscured by the background, so the size of strong, well-measured spots is a very good indication of the true size of all the spots in that part of the image. As a result, the best measurement is to add together the values of the pixels contributing to a reflection and subtract the background that is inevitably underneath the spot. This is known as 'summation integration' and is both fast and accurate.

Weak reflections, on the other hand, may have their shoulders hidden under the background, so summation integration will probably give an underestimate. Using the information that the spot shapes of neighbouring reflections are similar because of the physics of the experiment, it is possible to construct an expected shape for the spots and fit this to the measured pixel counts – this is 'profile-fitting' integration. For strong spots, this should give the same measurement as summation integration, but for weak spots the improvement in measurement can be substantial (Figure 2).

Both 2D and 3D programs use both summation integration and profile-fitting integration.

Scaling and merging

This gives the first reliable indication of the quality and resolution of the data collection itself, the data processing, and a good indication of the true symmetry of the crystal. It provides the best statistics on the data processing which are commonly summarized in what crystallographers often refer to as 'Table 1', because it is often the first table that appears in papers describing X-ray structures.

Scaling and merging is the process of putting the measured intensities onto a common scale by allowing for imperfections in the experiment and merging 'symmetry-related' measurements so that subsequent programs in structure solution can deal with them more conveniently. For 2D integration, it is the stage at which the partials are combined into single measurements that can be treated with fully recorded reflections.

For a variety of physical reasons, the measurements on different images in a dataset will be on different scales, for example:

1. The changing volume of crystal in X-ray beam as it rotates
2. The crystal may absorb X-rays differently in different directions (e.g., if the crystal is a plate)
3. The crystal may be decomposing due to radiation damage
4. The intensity of the X-ray beam changes during the course of the experiment
5. The dataset may be composed of partial data collections from different crystals.

Figure 2. Profile fitting. The physics of the experiment means that all these spots have the same shape and size on the detector; the weaker spots look smaller because their shoulders are hidden beneath the background. Their intensity can be measured more accurately if we apply an intensity profile calculated from the shapes of the strong, well-measured spots.
Provided that these effects vary smoothly through the course of the data collection, it is relatively straightforward to calculate correction factors to put the measurements on a uniform scale. Problems occur when there are sudden changes between adjacent images, and it is not easy to correct for these.

**Conclusion**

Single crystal X-ray diffraction remains, nearly 110 years after its initial development, the primary method of determining the structure of molecules ranging from a few to many thousand atoms; for example, at the time of writing, the vast majority of the ~1450 structures of proteins from the novel coronavirus have been determined by X-ray crystallography (~1150) rather than other methods like Cryo-EM (~250) or NMR (~35). Once you have the crystal, data collection at a synchrotron can take seconds (at a home source minutes to hours). Processing the dataset and solving the structure can be completed in a few minutes on a laptop; this is simply not possible at the moment for structure solution from cryoelectron microscopy.

**Further reading**

- By far the best way to find out more about data processing is to attend one of the courses run at synchrotrons and other sites during the year; these are advertised on the ‘CCP4 Bulletin Board’ (an internet search is the best way to find links to this). CCP4BB is also a particularly good place to ask for advice on any aspects of crystallization, data collection and processing.
- Although over 20 years old, the proceedings of the 1999 CCP4 Study Weekend on Data Collection and Processing in Acta Crystallographica Section D are still particularly well worth reading; https://journals.iucr.org/d/issues/1999/10/00/
- It may also be useful to read Powell, H.R. (2017) X-ray data processing. *Biosci. Rep.* 37, 1–14 DOI: 10.1042/BSR20170227
- A useful introduction to cryoelectron microscopy and its principal developers in structural biology can be found in Krämer, K (2017), Super Cool Science, *Chemistry World*, 14(11) 14-19

Harold Powell trained as a synthetic inorganic chemist but has worked in crystallography since the late 1980s and spent nearly 18 years at the MRC Laboratory of Molecular Biology in Cambridge developing the integration program Mosflm and its user interface iMosflm. Since 2016 he has run his own crystallographic consultancy and worked for the European Bioinformatics Institute on the PDBeKB project; in 2020 he moved to Imperial College London and is currently developing the homology modelling server Phyre2.