MRC Laboratory of Molecular Biology

Inside the Black Box of Data Processing

Harry Powell

MADaC-2015, Soleil

Workshop on Advanced Data Collection with Multi-Axis Goniometry

12th November 2015

MRC | Medical Research Council

This lecture provides an introduction to data processing of diffraction images obtained *via* the rotation method, which is the most widely used way of collecting data X-ray data from single crystals, both for macromolecules and small molecules.

Overview - Data processing

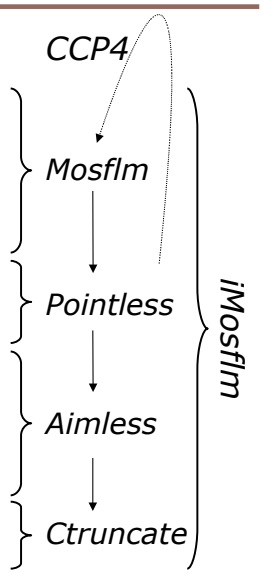
May be divided into stages:

Data reduction:
- Indexing (Bravais lattice)
- Parameter refinement
- Integration

Check symmetry
- Laue group
- maybe space group

Scaling and merging
- merging partials to form complete reflections
- merging symmetry equivalents

Truncation
- analyse intensity distribution
- convert $|F|^2$ to $|F|$)



2/25

The process of converting the spots on a diffraction image to indexed and measured diffraction data that may be used in structural analysis consists of four basic parts, though in modern programs these tend to merge into a single workflow.

Measuring the intensity of spots on the images is "integration". This can only be done well if the program knows the spot location, which is found approximately by indexing and then accurately by refinement of the crystal and detector parameters.

Once the measurements have been made, they are corrected for a variety of effects; purely geometrical effects are normally done by the integrating program – usually only Lorentz and polarisation effects. Other corrections, *e.g.* absorption by the crystal, differences between images (effective exposure, radiation damage, *etc.*) are either handled by the scaling and merging programs or by specialist programs devoted to particular aspects of the data.

Merging includes not only merging measurements of reflections that are equivalent by crystal symmetry, but also merging together the different components of reflections that are partially recorded over a number of adjacent images. This may be done either by the integration program (if it implements 3D profile fitting) or the scaling program (if the integration program performs a 2D analysis). Scaling attempts to put all of the observations onto a common scale, by accounting for errors and inconsistencies caused by the instrument or the crystal.

Truncation produces |F|s from these partially corrected $|F|^2$ measurements by taking account of expected statistical errors in measurement; analysing this process gives many of the diagnostics about twinning and also the Wilson statistics.

Indexing

Provides

- indices for reflections (*hkl*)
- unit cell dimensions (*a, b, c, α, β, γ*) } orientation matrix
- crystal orientation
- information about the crystal symmetry

Knowledge of these allows us to predict the positions of the diffraction spots on the image.

Unit cell dimensions are used in structure solution, refinement, model building, analysis - so we need accurate values.

3/25

Indexing is the process which gives indices for the reflections - they are often called “Miller indices”, but strictly speaking these refer to the lattice planes perpendicular to the scattering vectors (which correspond to the reflections). Indexing provides us with the information required to integrate the images in a dataset; the unit cell parameters and orientation of the crystal (in combination with known instrument parameters such as crystal to detector distance, wavelength of radiation, *etc.*) tell us where the diffraction spots occur on the detector for each image.

Further, the unit cell dimensions are used in many of the subsequent calculations in structure determination and refinement. Accurate values (obtained after refinement) will mean that the derived results have higher significance.

If we can determine the Bravais lattice, symmetry constraints can be applied in refinement to make the process more stable. Further, if we can determine the symmetry (or at least eliminate low symmetry solutions) we can run data collection strategy software and make sure we collect complete data with as small a rotation range as possible; in the case of crystals that suffer significantly from radiation damage this can be very important.

Indexing – overview

- Find spots on the image
- Convert 2D co-ordinates (image) to scattering vectors (corresponding to 3D RL co-ordinates)
- Index
- Cell reduction
- Apply Bravais lattice symmetry
- Pick a putative solution
- (Estimate mosaic spread)

Note that indexing only gives an approximate solution; we *hope* it will be good enough to proceed.

4/25

Indexing involves several distinct processes, the main ones of which are listed here. They start with "spot finding", or locating likely diffraction spots on the image or images (indexing tends to be more robust when information from several images separated in phi are used, rather than just from a single image).

The two-dimensional co-ordinates can be mapped (using the Ewald sphere construction) to scattering vectors that correspond to (approximate) 3D reciprocal lattice co-ordinates.

Indexing itself within *Mosflm* uses a "real-space" method (*i.e.* the real space unit cell dimensions are obtained directly, rather than via the reciprocal space unit cell) using an FFT-based method suggested by Gérard Bricogne in 1986 and implemented with a large set of 1D transforms by Steller *et al* (1997). An alternative formulation using a single 3D transform is used in HKL. XDS uses a method based on "difference vectors", which will not be discussed further here.

The initial cell obtained may not be the "reduced cell", *i.e.* with angles closest to 90º and the shortest cell edges, so "cell reduction" is performed next. At this point, the cell has triclinic symmetry; it can be transformed via a set of operations (listed in International Tables for Crystallography Vol. A) to 44 characteristic lattices (each of which corresponds to one of the 14 Bravais Lattices), and a distortion penalty calculated for each lattice. It is important to remember that the 44 solutions correspond to the single triclinic lattice obtained from indexing.

Having chosen a solution, the user should obtain an estimate of the mosaic spread of the crystal, prior to refinement. *Mosflm* uses an iterative integration routine to calculate a starting value.

Indexing

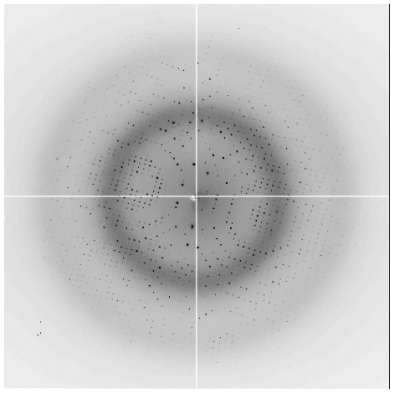
The 2D image co-ordinates of the spots can be converted to scattering vectors (that correspond to reciprocal lattice points):

$$s = \begin{pmatrix} D/r - 1 \\ X_d/r \\ Y_d/r \end{pmatrix}$$

$$r = \sqrt{D^2 + X_d^2 + Y_d^2}$$

D = crystal to detector distance

X_d, Y_d = spot co-ordinates on image relative to beam centre



n.b. wavelength, crystal to detector distance and beam centre must all be known

5/25

Here, D is the crystal to detector distance, X_d and Y_d are the spot co-ordinates relative to the beam centre on the image, and r is derived above (usually these are all in mm). In this calculation, $\boldsymbol{s}$ is in dimensionless reciprocal lattice units and the radius of the Ewald sphere is unity. The reciprocal lattice obtained is somewhat distorted, partly because the beam centre and the crystal to detector distance may be incorrect, and the detector may not be planar and truly orthogonal to the X-ray beam. The normal procedure is to assume that the phi value for each spot is the mid-point of the rotation for this image; plainly, this will not be true for spots which appear early in the rotation or for those at the end. However, provided that the rotation range for each image is not too great, however, the error is acceptably small.

Remember that all the spots that are visible on the image correspond to reciprocal lattice points that are on the Ewald sphere at some point during this individual exposure.

Note that this relationship only holds when the detector is in the "symmetrical" setting, *i.e.* the two-theta swing angle is zero, and the beam is perpendicular to the detector; the two-theta swing can be accommodated by a simple modification to this formula, but other variations can be dealt with by a more complete description of the detector geometry (this will not be dealt with here).

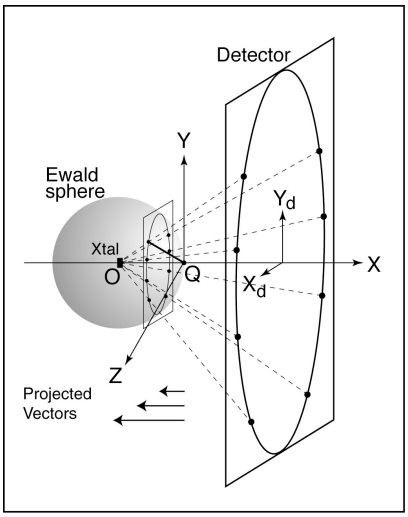
The reciprocal lattice produced must also be oriented to reflect the orientation of the crystal; this can be done by applying a simple rotation about the origin to each of the lattice points calculated

Even in the simple case presented here (which is a very good approximation to the vast majority of actual cases), the importance of knowing the wavelength of radiation used, and of determining the beam centre and crystal to detector distance accurately is obvious.

One-dimensional FFT Indexing

If the scattering vectors calculated are projected along a reciprocal space axis direction (such as a*, b* or c*) all the projected vectors for spots in the same reciprocal space plane will have the same length, as will all those spots in the next plane, *etc*.

This will give a large peak in the Fourier transform.



6/25

Probably the most reliable method for auto-indexing is based on the Fourier transform of the calculated reciprocal space co-ordinates of the diffraction spots.

In the diagram above, the spots recorded on the detector are projected onto a representation of the Ewald sphere (since all reciprocal lattice points will only give rise to diffraction spots when they are in contact with the Ewald sphere). If the scattering vectors (from the origin of the Ewald sphere to the surface of the sphere) are projected onto a vector corresponding to a reciprocal lattice axis, the projections can be summed to reinforce each other.

For reciprocal lattice planes that have a simple relationship to each other, the projected vectors will also have a simple relationship. For example, the vectors corresponding to the 1kl, 2kl, 3kl planes will have lengths in the ratio 1:2:3 (see next slide). The projections which have more contributing planes will have more regularly spaced peaks, and so give rise to Fourier Transforms with peaks which are more distinct from the background.

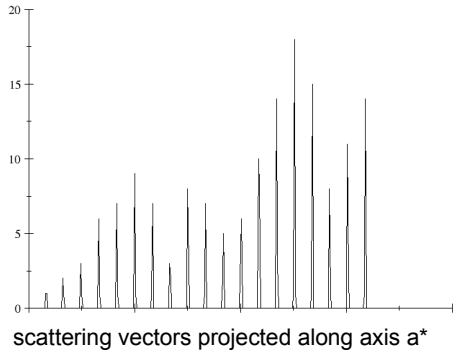
It should be remembered that generally, the crystal will not be aligned with a reciprocal space axis parallel to the X-ray beam, so the chance of obtaining the above construction is small; by calculating the projections in many directions, we increase the chances greatly (to near certainty) that some of these projections will correspond to crystal axes.

The projections are actually calculated by computing the scalar (or dot) product of the distorted reciprocal lattice points (expressed as vectors from an origin) with the vector that describes the direction of the projection, then summing the dot products.

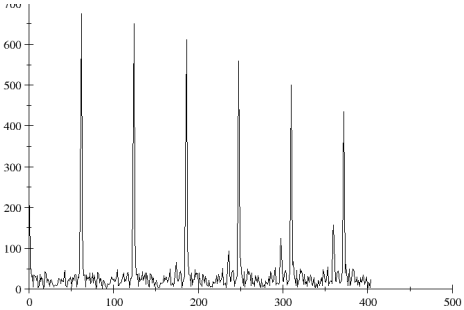
FFT Indexing

The first large peak in the Fourier transform corresponds to a real space cell edge length. In this case, ~67Å.

Provided that a single image samples enough of reciprocal space, we can get information about all three crystal axes from one image.



scattering vectors projected along axis a*



Fourier transform of scattering vectors

For directions other than reciprocal space axes, the projected vectors will have different lengths, and will not (in general) give a large peak in the Fourier transform. The indexing in *Mosflm* calculates several hundred projections, regularly spaced around a hemisphere of reciprocal space and applies a Fast Fourier Transform (FFT) to each. Although in principle, we only need to find the 3 FFTs corresponding to the three principal cell axes, they may not all be present (*e.g.* if the crystal orientation does not allow it), or we may find vectors corresponding to edges in a non-reduced cell. In practice, 30 FFTs produced which have the largest peaks are selected to determine which can be combined to give a real space unit cell which accounts for the majority of the reflections.

The unit cell determined is reduced to give a primitive cell in a conventional setting, *i.e.* one which has its three inter-axial angles as close to orthogonal as possible and the three axial lengths as short as possible. Cell reduction does not change the unit cell volume, unless there is also a change in lattice centring.

Indexing only gives the geometry of the cell

Indexing gives us a basis solution that is triclinic.

Applying symmetry transformations to give the *reduced bases* allows us to see how well this triclinic solution fits the cell edges and angles of lattices with higher symmetry, *e.g.* monoclinic, orthorhombic etc.

Mosflm and *XDS* give all 44 solutions: each of these corresponds to one of the 14 Bravais lattices (each of which may occur several times as a result of different transformations). *Denzo* and *HKL* only give the "best" 14 Bravais lattice solutions *which may not include the correct one*. *DIALS* only reports those solutions with a penalty less than some threshold value (*i.e.* the "good ones").

The unit cell geometry from indexing may not be the correct crystal symmetry, but it usually is.

The space group is only a hypothesis until after your structure is deposited in the PDB

8/25

The cell dimensions derived from autoindexing usually give a good indication of the true symmetry of the crystal. For example, in the case that $a \neq b \neq c$, $\alpha \neq \gamma \neq \beta \neq 90$, the crystal system is most probably triclinic, unless the indexing has failed. If $a = b \neq c$, $\alpha = \beta = \gamma = 90$, the crystal system may be tetragonal, but there are many examples where unit cells fit this but the true symmetry is orthorhombic or lower.

However, probably more than 95% of the time, the crystal symmetry derived from the unit cell geometry will be correct.

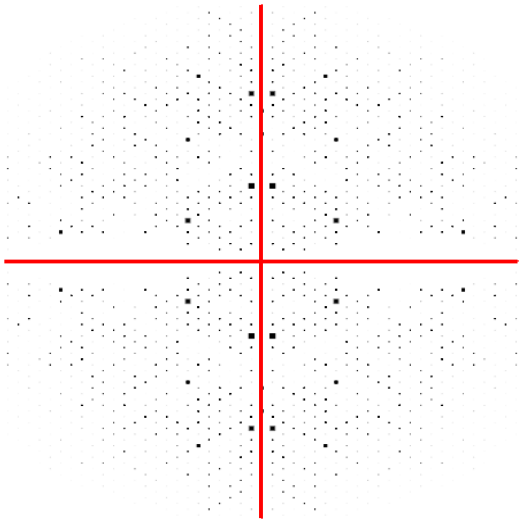
The practice of providing all 44 characteristic lattice solutions in *Mosflm* and *XDS* is to be preferred to that of *Denzo/HKL*; the latter only gives the "best guess"of each characteristic lattice as a choice. A small error in instrument parameters, or even in the choice of spots used for indexing, could easily give rise to the correct solution not being present in the list of results, *even though the program has actually calculated it.* The *DIALS* toolbox only reports the solutions which have a small penalty, so the list length can vary between samples.

The 44 characteristic lattices and the transformations from the basis triclinic solution that correspond to the reduced bases are tabulated in International Tables Volume A pp 750 - 755. Each characteristic lattice (or lattice character) is associated with a Bravais lattice, *e.g. aP* is primitive triclinic ("anorthic Primitive"), *mC* is C-centred monoclinic *etc.*

Bravais lattice – from intensities

The true Bravais Lattice symmetry can *only* be determined by analysing the intensities of symmetry equivalent reflections – *i.e.* after integration.

example of $C222_1$ with a = 74.7Å, b = 129.2Å, c = 184.3Å, which could be (incorrectly) indexed as hexagonal a = b = 74.7Å, c = 184.3Å.

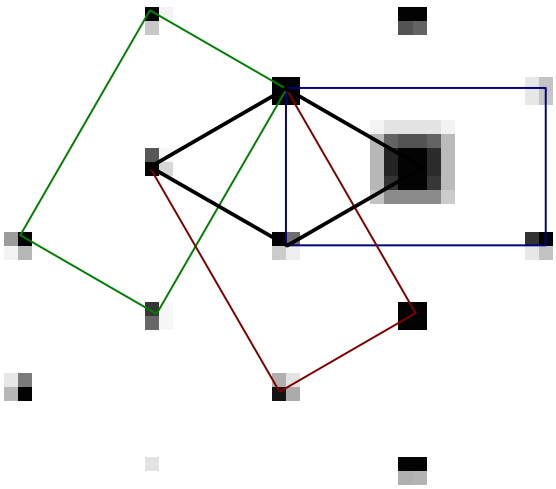


9/25

This is an example provided to Phil Evans where the metric symmetry indicated that the crystal was hexagonal, but the merging statistics showed that it was C-centred orthorhombic; the *mm* symmetry of the diffraction spots projected along the c* axis clearly illustrates this.

There are also two incorrect C-centred orthorhombic solutions at 120° to the correct solution, with identical cell parameters; again, it can be seen that the reflections that should have the same intensity by hexagonal symmetry do not match.

It is interesting to note that autoindexing gave variously the hexagonal or one of the three orthorhombic solutions, depending on the choice of spots used in indexing – or only a one in four chance of the correct answer. Differentiating between the four solutions and picking the correct one can only be done *after* integrating at least some images; *iMosflm* includes a task button in the *Integration* pane that runs *Pointless* to perform this analysis.



Refining the parameters (1)

Optimise the fit of observed to predicted spot positions, so that the measurement boxes can be placed accurately over the spots.

Specifically, improve estimates of:

- Crystal parameters
- Instrument parameters

Accurate cell dimensions are important because they are used in all subsequent stages of structure determination, refinement and analysis

Can be performed by either (or both):

- Positional refinement using spot co-ordinates
- Post-refinement using intensity measurements

10/25

Indexing is based on approximations, and the fit of observed spots to their calculated positions can be improved by refinement. These approximations include the phi position of the centroid of each reflection and various parameters like crystal to detector distance and detector mis-setting angles. Provided that there are sufficient usable data at high enough resolution, refinement not only gives better information about where on the detector the spots occur, but also gives better estimates of both the crystal and instrument parameters.

Most integration programs use a "positional refinement" based on the spot positions on the detector surface; this is simple to calculate, but care must be taken because several parameters are closely correlated (*e.g.* cell edges and crystal to detector distance), especially at low resolution.

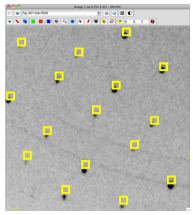
Mosflm combines positional refinement with another method, which is based on the relative intensities of the different parts of partial reflections across several images. Because this can only be done *after* the reflections have been integrated, it is called "post-refinement". Using both methods together has distinct advantages over just using positional refinement, *e.g.* it is possible to de-couple the crystal parameter refinement from that of the crystal to detector distance, and it also gives (provided there are sufficient reflections for a stable refinement) more accurate cell parameters than those available from positional refinement.

Other processing packages delay post-refinement until a step following integration, and often combine it into the scaling and merging step.

Refining the parameters (2)

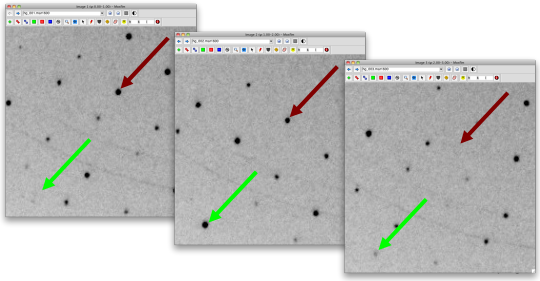
Positional refinement:

$$\Omega_1 = \sum_{i=1}^{n} w_{ix}(X_i^{calc} - X_i^{obs})^2 + w_{iy}(Y_i^{calc} - Y_i^{obs})^2$$



Postrefinement:

$$\Omega_2 = \sum_{i=1}^{n} w_i \left[\frac{(R_i^{calc} - R_i^{obs})}{d_i} \right]^2$$



$(X, Y)^{obs}$ and $(X,Y)^{calc}$ are the observed and calculated spot co-ordinates on the detector (usually transformed to some virtual detector frame).

The cell dimensions and crystal to detector distance are strongly correlated, particularly at low resolution, and it can be hard to refine both stably at the same time. *Mosflm* avoids this by refining the distance *via* positional refinement, and the cell dimensions *via* post-refinement.

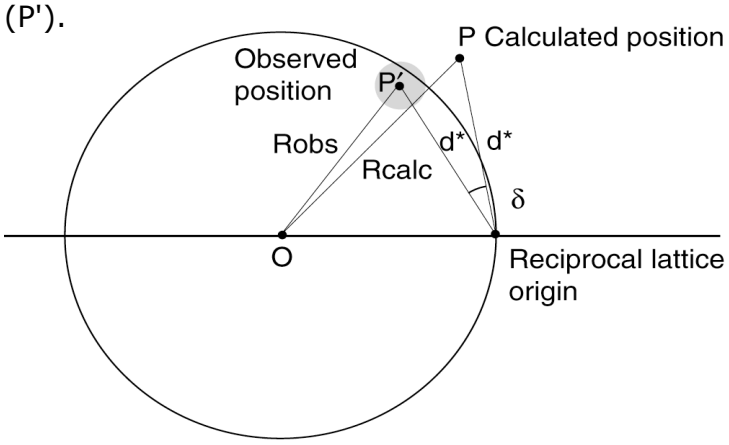
Reflections that are spread across two or more images are in the process of traversing the Ewald sphere. The relative intensities of the parts of a single reflection on consecutive images is related closely to how close the reciprocal lattice point is to the Ewald sphere. The "rocking curve" describes how the intensity of a reflection varies with the crystal orientation. We can use this knowledge to get more accurate information on the unit cell and other experimental parameters.

R^{calc} and R^{obs} are the calculated and observed distances of the phi centroid from the Ewald sphere, but may also be thought of as the calculated and observed partiality for each reflection.

The radius of convergence of post-refinement is smaller than that for positional refinement, so the parameter to be optimised must be closer to its true value for the process to be stable and accurate. Post-refinement can routinely give cell dimensions that are accurate to within a few parts in 10,000 (*e.g.* 0.03Å error in a cell edge of 100Å).

Post-refinement

We can visualise this in the Ewald sphere construction, minimising the angular residual δ. A suitable model for the rocking curve allows us to determine the "observed" position (P').



12/25

Ewald Sphere

The Ewald sphere is a useful way to visualise the conditions required for diffraction. The crystal is at "0", and the reciprocal lattice origin is at a distance $1/\lambda$ away, on the surface of the Ewald sphere. As the crystal is rotated, the reciprocal lattice rotates synchronously with it. A reciprocal lattice point is in the diffracting condition when it is on the Ewald sphere surface; with an ideal crystal with zero mosaicity and ideally monochromatic radiation, this would happen instantaneously (the surface of the Ewald sphere would have zero thickness and the reciprocal lattice points would have zero size). In practice, most crystals are not perfect, and the reciprocal lattice points have finite size. Also, the Ewald sphere surface has a finite thickness. Taken together, these mean that the reciprocal lattice points are crossing the Ewald sphere for a finite time so diffraction spots are seen through a small rotation range.

Post-refinement minimises the difference between the calculated and observed distances of reciprocal lattice points from the Ewald sphere, by minimising the angular residual δ.

Integration itself

Two basic ways -

- summation integration

simple, fast, okay for all except weak, overloaded or partially overlapping reflections

- profile fitting (only *intended* to improve weak spots)

can be sub-divided into

- two-dimensional (2D) – builds up reflections from profiles on single images (but we can use spots on different images)
- three-dimensional (3D) – builds up profiles across several adjacent images

13/25

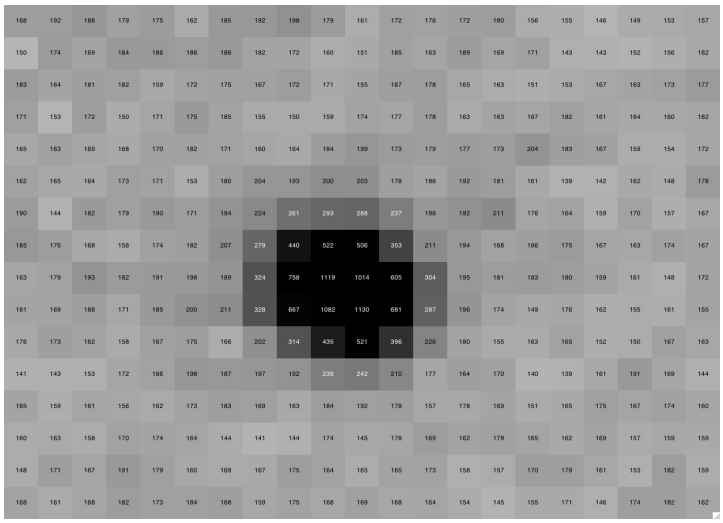
Integration is performed once the crystal and instrument parameters have been optimised by refinement.

The main difference between two-dimensional and three-dimensional integration is that the profiles used for partials over several images for 2D integration are the same for each part of the reflection, whereas for 3D integration, the profile for different parts of the same reflection can change significantly.

In principle, 3D profile fitting should give better results than 2D, but in practice the difference does not seem to be important, and other differences between programs (or even parts of the same program) tend to dominate.

Measuring the intensity of a spot

Identify the background & spot regions, work out what the background level is around the spot, then *assume* it is the same under the spot.



14/25

The first part of integration is to work out where the diffraction spot is, and where it ends. The assumption is made that, in the region of the spot, the background is planar and may have a slope. The background plane and its slope are calculated from pixels in the neighbourhood of the spot, once the spot pixels have been determined.

Some programs optimise the spot region, whereas others rely on the user to do this. Generally, more modern programs will do this for the user.

It can be seen from this region around a diffraction spot, that although the intensity in the background is much lower than in the spot, it is not actually flat and level; this is due to a number of reasons (*e.g.* detector noise), but our concern is how best to take this variation into account when determining the background. If we take a statistically significant number of pixels, we can get a good estimate of the background level.

Mosflm uses a rectangular mask, which is divided between an octagonal spot region and the background region. Before optimisation, the background area is chosen to be ~8x the size of the spot region, and then only the spot is optimised. If the spot region becomes larger, the overall measurements of the box are increased. If the background area drops to less than twice the spot size as a result of expansion of the spot region, the process halts and the user is prompted to intervene. This very rarely happens except with very large cells (which have many spots close together).

Summation integration

- In the absence of background, just add the pixel counts in the spot region together - but there is (always) background!
- Need to define spot and background regions - we cannot measure background directly under the spots, so we calculate a local background plane and slope from nearby non-spot pixels
- Use this to subtract the background under the spots
- Weak spots may have their shoulders under the background, so that their measurement is impaired.

15/25

If the background intensity is negligible, the program doesn't even need to be very accurate in its placement of the integration boxes when using summation integration, provided they enclose all the spot intensity.

In practice, however, there is always some background, so this needs to be taken into account. It is impossible to measure the background directly under the spot, but its intensity can be inferred by assuming it to be a sloping plane in the neighbourhood of the spot. If the plane is steeper than some threshold value (*e.g.* because the spot is near an ice-ring), *Mosflm* will issue a warning.

With some newer detectors that have very low intrinsic noise levels and small point-spread functions, it is probably correct to integrate using summation integration (at least for the strong reflections), especially when the background is low. However, weak spots will still have their shoulders hidden by the background, and summation intensity will not measure their intensity optimally.

Seed skewness – a variant on summation integration

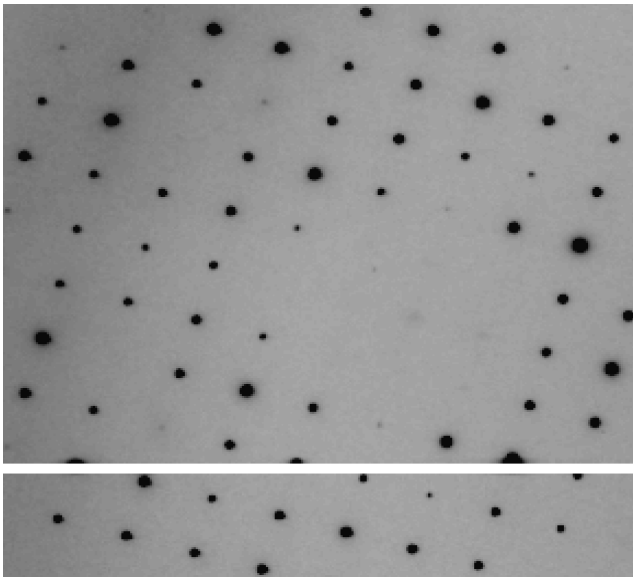
It is possible to analyse the intensity distribution of the background region pixels and use this to optimise both the shape and the size of the measurement box for each spot individually (by adding and/or subtracting pixels from the initial "seed" spot region) – this is done in the process known as "seed-skewness". This improves the spot measurement *indirectly* by optimising the measurement of the background. It is a very computationally expensive process (since it has to be performed for every single spot), and so it is slow; none of the commonly used integration programs follow this approach.

Integration by profile fitting

Based on the observation* that spots corresponding to fully recorded reflections in the same region of the detector (and on images nearby in phi) have similar profiles.

* for detectors with

- a finite point-spread function
- small pixels *cf* spot size

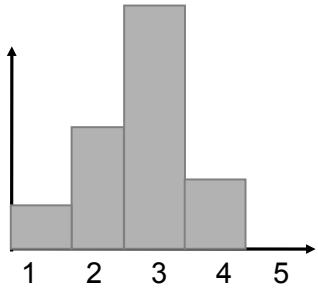


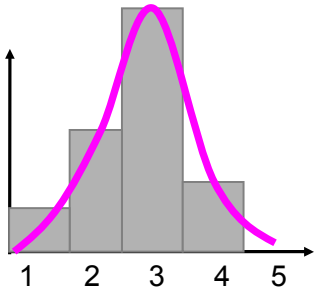
The spot shape on a detector (including its intensity profile) is a function of several physical factors – the cross-section and divergence of the illuminating radiation, the size, shape and mosaic spread of the crystal (and its orientation relative to the beam), the direction the diffracted beams exit from the crystal, scatter from air in the beam path, the size and shape of the pixels on the detector, etc.

For a given image (or short series of images) most of these may be assumed to be constant in the diffraction experiment (or nearly constant); the biggest change between nearby (fully recorded) spots is in the direction of the diffracted rays from the crystal, and if the angle between these rays is small, this major difference is also small, so the idea that spots close to each other on the detector (even on different images) have similar profiles has some validity. However, if the physical spot size (determined by the cross-section of the diffracted rays) is similar to the pixel size on the detector, and the detector has a point-spread function that is small compared to the pixel size, this may not be true. There are other complicating factors which may occur to the reader!

Profile fitting integration – standard profiles

Use a profile determined empirically from well-measured reflections to measure the intensity of weak reflections (whose shoulders disappear below the background), by fitting a learnt profile to the observed reflection:





- requires accurate (sub-pixel) placement of the profile
- reduces variance for weak reflections
- should reduce random error (weak reflections)
- may increase systematic error (strong reflections)

17/25

If the centre of each reflection on the detector is not calculated accurately, the profiles calculated using the spots will be broader than the true profile because the centres of the measured profiles will not coincide exactly. This can give rise to systematic errors that are largest for the strongest reflections, even for detectors with relatively large PSFs. Modern programs do locate the centres very accurately, so generally this is not a big problem, but it should be borne in mind when analysing results; in some circumstances it may be appropriate to use summation integration for the strongest reflections and profile fitting for the weaker ones. *Mosflm* records both measurements in the output MTZ reflection file, and *Scala* or *Aimless* can perform the appropriate combination.

Questions about the data

- What is the overall quality of the dataset?
  - How does it compare to other datasets for this project?
- What is the real resolution?
  - Should you cut the high-resolution data?
- Are there bad batches
  - individual duff batches or ranges of batches?
  - Is the whole dataset bad?
  - Should it just be thrown away?
- Was the radiation damage such that you should exclude the later parts?
- Is the outlier detection working well?
- Is there any apparent anomalous signal?
- Are the data twinned?

18/25

Once we have decided that the fast scaling and merging (*e.g.* provided by the *QuickScale* option in *iMosflm*) have proceeded without too much incident, we can start to look at the output more closely to make sure that the dataset itself is of sufficient quality to proceed. As with integration, if serious problems are encountered, it is always worth asking if it is worthwhile struggling to use a bad dataset (and get the best out of it), or if it should be discarded and a new dataset collected on a new crystal.

A further question is "are the data any good for the experiment we want to perform?", *e.g.* we don't need atomic resolution data for a SAD experiment, and we don't need an anomalous signal for refinement. Therefore, concentrate on those diagnostics that are relevant.

Scaling and merging

This is the next step following integration. It is important because -

- It attempts to put all observations on a common scale (which is necessary for subsequent structural analysis)
- It provides the main diagnostics of data quality ($CC_{½}$, I/σ(I), resolution, *etc.*) and indicates whether the data collection and data integration were satisfactory

Because of this diagnostic role, it is important that the data are scaled as soon as possible after data collection – it is best to do it during data collection, preferably while the crystal is still on the camera.

- Do not leave integration and scaling until you get home after a synchrotron visit!

19/25

Scaling and merging follows integration and together provide the main diagnostics concerning the quality of the data collection and the data processing.

Scaling attempts to put the observations onto a common scale, allowing for variations which have occurred in both the sample and the instrument used for data collection (the "camera", which comprises the detector and other diffractometer components like shutter, goniostat, X-ray source, etc).

The term "merging" covers two quite different processes, i.*e.*

(1) merging together of the parts of reflections that are partially recorded over multiple images to form complete observations and

(2) merging together symmetry-related copies of these complete observations into single measurements.

For some purposes, the second of these two steps is performed outside the normal scaling and merging procedure, for example programs like the *SHELX* suite that use their own internal merging tests on "unmerged data".

In the process of scaling and merging, all the normal quality criteria such as the various merging R values and correlation coefficients are calculated, and for this reason it is important to perform this step as soon as possible after the data collection is finished - preferably while the crystal is still mounted on the camera, so that better and/or more data can be collected if necessary.

On a synchrotron visit, process *all* your data while you are at the beamline! It is always possible to re-process again later if necessary, but this first processing gives confidence in the quality of the experiment.

Scaling

We try to make symmetry related and duplicate measurements of a reflection equal by modelling the diffraction experiment, principally as a function of the incident and the diffracted beam directions in the crystal.

Scaling attempts to make the data internally consistent, by minimising the differences between the individual observations I and the weighted mean of all the symmetry-related equivalents of reflection I.

Provided that there has been no radiation damage to the sample, all reflections related by the space group symmetry should have equal intensities (and hence structure factor amplitudes). However, because we live in the real world where our sample and instruments are not perfect, this will not be true. Therefore, we have to try to model the likely causes of the differences and apply appropriate corrections to individual measurements.

In the absence of other information, we can only try to make the make internally consistent, *i.e.* try to make symmetry equivalent data agree. Scaling does not attempt to scale non-symmetry related data, so any systematic errors which are the same for symmetry-related data will remain.

Why are reflections on different scales?

Some physical factors vary during the experiment, including those associated with

(1) the incident beam and the camera

(2) the crystal and the diffracted beam

(3) the detector

Scaling should model the parameters contributing to each of these groups appropriately; since experiments differ, each experiment may require a different model

Understanding the effect of these factors allows a sensible design of correction and an understanding of what can go wrong

Various physical factors lead to observed intensities being on different scales. Some corrections are known at the time of data integration because these are related to the instrument, the X-ray source and the method of data collection (*e.g.* Lorentz and polarisation corrections), but others can only be determined subsequently. Careful analysis of the variations in intensity differences allows the effects of the different factors to be modelled sensibly in scaling.

(1) Factors related to the incident beam and the camera

(a) Variable beam intensity

(b) Changes in the illuminated volume of the crystal

(c) Absorption of the primary (incident) beam by the crystal (indistinguishable from (b))

(d) Variations in rotation speed and shutter synchronisation. Shutter synchronisation errors lead to partial bias which may be positive, unlike the usual negative bias.

"Shutterless" data collection (*e.g.* with Pilatus detector) avoids synchronisation errors (d), but very small rotation angles can still cause problems with variations in rotation speed.

Exposures that are long compared to variations in intensity tend to reduce this problem, but very short exposures can cause a large variation in scales between adjacent images.

If the crystal is smaller than the beam and correctly centred, the illuminated volume will remain constant as the crystal rotates, but if it is larger than beam or if it moves in and out of the beam while rotating, the total diffracting volume will change, and consequently so will the scales for the images.

If the crystal absorbs X-rays heavily (*e.g.* because it contains a heavy atom) and it is not isometric, then more X-rays will be absorbed in some orientations compared with others before they diffract.

The shutter needs to completely open and close precisely at the correct rotation angles; errors associated with poor shutter synchronisation can lead to positive partial bias, *i.e.* the intensities of summed partials is greater than expected compared with the symmetry equivalent fully recorded reflections.

(2) Factors related to the crystal and the diffracted beam

(a) Absorption in the secondary (diffracted) beam – serious at long wavelength, *e.g.* CuKα on a home source

(b) radiation damage – serious on high brilliance sources. Not easily correctable unless small, as the structure is changing

Maybe extrapolate back to zero time? (but this needs high multiplicity)

The relative B-factor is largely a correction for the average radiation damage

Absorption in the secondary beam is more evident with long wavelength radiation *e.g.* (CuKα, 1.54Å), and can be noticeable when second row main group elements (*e.g.* S, P, Br) are present. The presence of heavier elements exacerbates the problem. Shorter wavelength X-rays (say, 1Å or less) are not absorbed as badly.

Radiation damage is a real problem with high brilliance sources, and is not easily correctable since the structure of the molecule being studied is actually changing (typically, CO_2 is lost from acidic amino acids and S from methionine, cystine and cysteine). The best way of dealing with this is to avoid it by using low doses of X-rays and collecting data quickly before significant damage has occurred.

If the relative B-factor drops significantly during the course of the data collection, it is a strong indication that radiation damage has occurred, as is a strong increase in the "Rcp *vs* batch" plot. Rcp is the "cumulative pairwise residual" which measures the overall merging R value of the data from the start of the dataset up to the current batch.

(3) Factors related to the detector

The detector should be calibrated properly for spatial distortion and sensitivity of response, and should be stable. If this is not true, problems are difficult to detect from the diffraction data.

(a) *e.g.* there are known problems in the corners of detector modules, both CCDs and Pilatus (some programs correct for these)

(b) Calibration should flag defective pixels ("hot" or "cold"), and also dead (or otherwise unreliable) regions between modules.

(c) The user should tell the integration program about shadows from the beamstop, cryocooler or other shadows.

Most modern commercial detectors write images which are corrected for spatial distortion and for non-uniformity of response; Bruker CCD detectors are the only widely installed devices that do not write distortion corrected images by default.

Other (minor) problems exist with "tiled" detectors, *i.e.* those made several individual modules arranged in a mosaic. Both CCD and individual PAD chips are less sensitive in the corners; this can be corrected in the scaling stage if the

Essentially all electronic detectors have a few bad pixels that consistently give readings that are too high or too low (or are "dead", reading 0); these can be located by the manufacturer, and a bad pixel map supplied to the customer. During use, more pixels may become damaged, so it may be worthwhile re-calibrating occasionally.

Integration programs do not care if predicted diffraction spots are masked by physical objects, but the scaling programs cannot be expected to scale properly recorded reflections with symmetry equivalents that are obscured (and therefore integrated incorrectly).

Finally

Remember -

- Don't expect software to correct for a badly performed experiment
- Take the time to look at your images and the results of integration and scaling
- Scaling and merging provide the best statistics on the quality of your data

25/25