MRC Laboratory of Molecular Biology

Integration and Scaling

Harry Powell

MRC LMB Crystallography Course

10th May 2013

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.

Preamble – rationale for the experiment

What are we doing, and why are we doing it?

Measuring intensities of diffraction spots to obtain structure factor *amplitudes*

$$|F_{hkl}| = \left(\frac{KI_{hkl}}{Lp} \right)^{1/2} \quad (1)$$

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx+ky+lz)} \quad (2)$$

Careful data collection and careful measurement of intensities can be used to recover the phases (which are otherwise lost)

2/44

The intensity of each reflection is related to the "structure factor amplitude" by equation (1) above. "*L*" is the Lorentz factor (which will be discussed later, but depends on, among other things, the data collection method), and "*p*" is the polarisation factor, related to the method of monochromation and the X-ray source. Both *L* and *p* depend on the diffraction angle of each reflection. "*K*" is usually a constant for a given crystal in an experiment, and depends on the crystal size, beam intensity and a number of other fundamental constants; since it is the same for every reflection in a dataset, it is usually applied as an overall scale factor to the measurements.

The "Structure Factor equation" (2) demonstrates why it is important to collect and measure the intensities as well as possible, since the electron density that gives us our structural model depends on the values we obtain for *F*. The electron density at every point in the cell depends on the intensity of every single reflection. Any badly measured or missing reflection will affect the maps we calculate.

Note that we are working with X-ray waves, and each diffracted ray has both an amplitude and a phase. The structure factor equation uses the structure factors *F*, not just the amplitudes |*F*|, but the phase information is lost in the data collection process. However, careful data collection and processing can allow us to obtain the phase information, usually by analysis of small differences in |*F*| between related reflections, *e.g.* in anomalous dispersion experiments like SAD or MAD, or in the classic heavy atom methods.

Optimization of Data Collection

Pre-process at least one image *before starting the full data collection* (preferably two at 90º to each other) to obtain:

- Cell parameters, crystal orientation and putative Laue group
- Estimate of mosaicity
- Effective resolution limit }
- Optimal crystal to detector distance } *e.g.* use *BEST*
- Exposure time }
- Strategy for data collection }

Remember! This is the last experimental stage - if you collect bad data now you are stuck with it. No data processing program can rescue the irredeemable!

Don't necessarily do what your PI or post-doc (or even the beamline scientist) says – think! At Diamond or ESRF use Edna

3/44

It is always worthwhile spending some time prior to the full data collection to determine sensible parameters for the data collection. For example;

• are you using the full area of the detector?

• does useful diffraction go beyond the edge of the detector? Does it stop halfway to the edge?

• check for overloads - are there a lot? Are you using the full dynamic range of the detector? Consider a low and a high-resolution pass. You may need to increase or decrease the exposure time.

• is the rotation angle too big or too small? As a first approximation, aim at half the mosaic spread for CCD detectors, or 1/4 - 1/5 for unshuttered data collection with Pilatus. Instrument instabilities and detector “dead-time” limit the minimum rotation range and time per image.

• check that the predicted spots really coincide with their positions on the image(s); is your initial estimate of the mosaicity realistic?

Remember to use prior information! If you have experience of your particular sample or experimental setup, use your knowledge. If something looks odd, investigate it.

Whatever integration program you are using, there will be an option (or an external program) which can calculate the optimum data collection strategy for you.

Don’t just use a standard recipe for data collection; it is almost always possible to collect a better dataset with a little forethought. Avoid the “American method” (“shoot first, and ask questions later...”).

First things first - look at the images

Questions:

- are there any spots on the image?
- has the detector been used efficiently?
- do the spots look reasonable – split? large? above background?
- can you see separate lunes?
- is there a single lattice?
- should I throw the crystal away now and collect a dataset on another crystal instead?

Check two images at 90° to each other – some pathologies are not apparent from a single image.

4/44

There are automated procedures for processing diffraction images, but they are not much use if your images display some kind of pathology, *e.g.* the crystal is split, it only diffracts to low resolution, or the mosaic spread is so high that the lunes merge into each other, making it impossible to determine the indices of each reflection.

Get into the habit of checking your images early; if you do this before starting the full data collection, or while still at the beamline, then you have the chance to collect data from a new, better crystal – after your return home it will not be quite as easy!

Sometimes the best decision is to throw the crystal away and not waste time on it. When difficulties are encountered in processing, they rarely occur with high-quality crystals.

Before starting to process

Use the program tools to mask backstop, cryostream, other shadows.

- Set resolution limit to about 0.2Å higher than visible spots.
- Make sure beam position is more-or-less correct.
- Make sure other parameters (distance, wavelength, rotation angle) are what you expect (do they correspond to what is in your notebook?).

5/44

Although integration programs can make good attempts at measuring spots that are partially masked by obstructions such as the backstop, backstop arm or the cryostream, these reflections can cause severe difficulties in scaling.

For example, if there are only two measurements of symmetry-related equivalents of a reflection, and one is weak and the other strong, the scaling program cannot tell if one is masked or the other contains a zinger.

Usually, there is measurable intensity beyond the resolution limit visible to the eye; I find it is reasonable to integrate to about 0.2Å higher resolution.

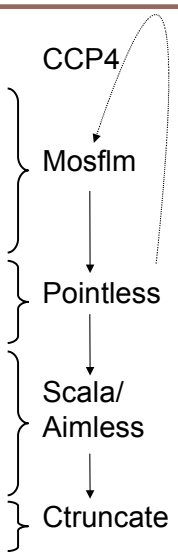
The beam position is critical to successful indexing and further processing. It should be correct to less than half the minimum spot separation, or the calculated indices of the reflections could well be out by one or more, even if the cell is approximately correct.

Finally, make sure the parameters used by the program are what you expect, or remember from the data collection. Don't necessarily believe the information in the image headers - some beamline staff are less thorough in updating their set-up files than others.

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 $|F|$)



6/44

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 (approximations for)

- unit cell dimensions } combined in the
- crystal orientation } "orientation matrix"
- (first estimate of the Bravais lattice)

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

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

7/44

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.

8/44

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.

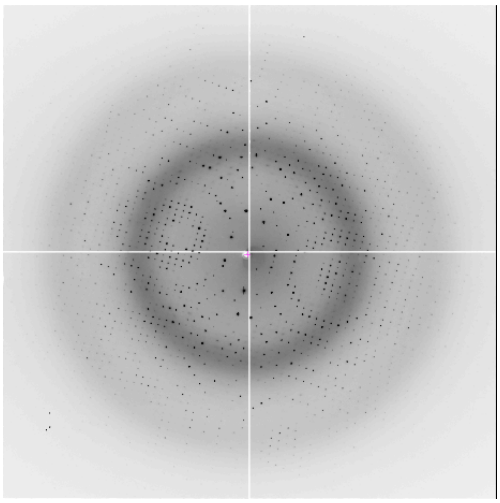
Indexing

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

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

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

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



9/44

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).

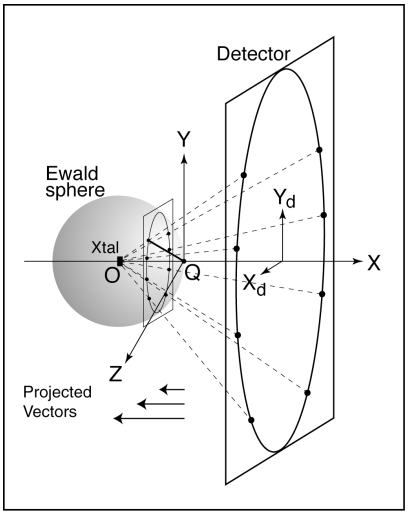
The reciprocal lattice produced must also be oriented to reflect the orientation of the crystal; this can be done by applying a simple rotation about the origin to each of the lattice points calculated

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.

Indexing

If the scattering vectors calculated are projected along a real 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.



10/44

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.

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.

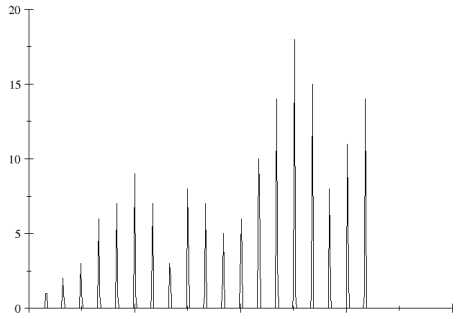
It should be remembered that generally, the crystal will not be aligned with a reciprocal space axis parallel to the X-ray beam, so the chance of obtaining the above construction is small; by calculating the projections in many directions, we increase the chances greatly (to near certainty) that some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

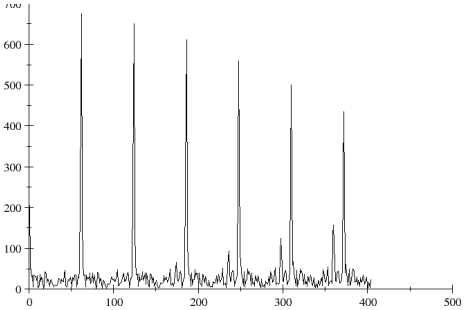
Indexing

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 real 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.

The unit cell geometry 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

12/44

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.*

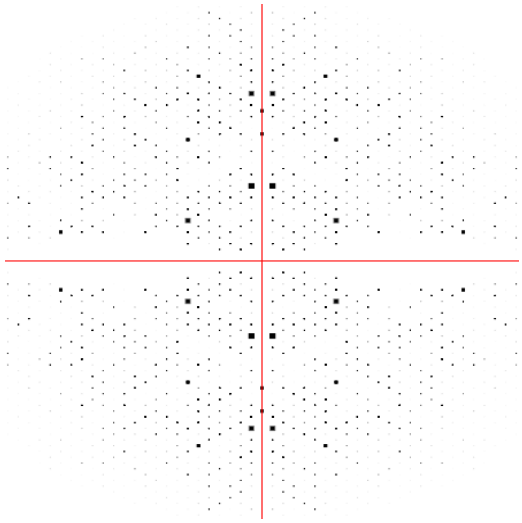
The 44 characteristic lattices and the transformations from the basis triclinic solution that correspond to the reduced bases are tabulated in International Tables Volume A pp 750 - 755. Each characteristic lattice (or lattice character) is associated with a Bravais lattice, *e.g. aP* is primitive triclinic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

There are also two incorrect C-centred orthorhombic solutions

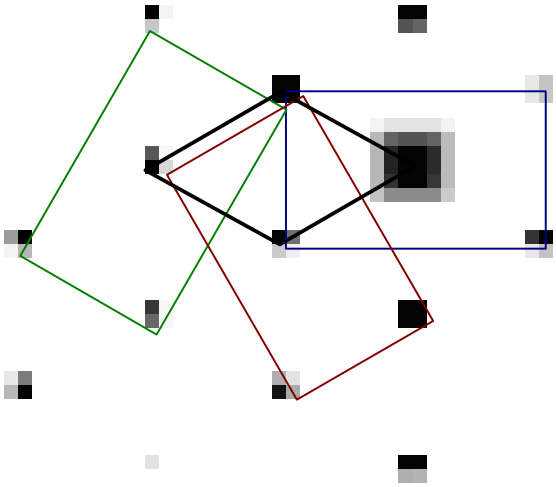


13/44

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

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

14/44

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.

Positional refinement and post-refinement

Positional refinement

- uses the spot positions on each image, so it can be done for each image without reference to the others. Both fully and partially recorded reflections can be used.

Post-refinement

- needs intensity measurements for spots which are spread across at least two images; we cannot use fully recorded reflections for this
- ∴ needs at least two adjacent images (and probably more for fine-phi slicing, where the mosaic spread is more than twice the rotation angle)

15/44

Positional refinement can be done on an image-by-image basis, since all the information required is present on each image; all integration programs allow this.

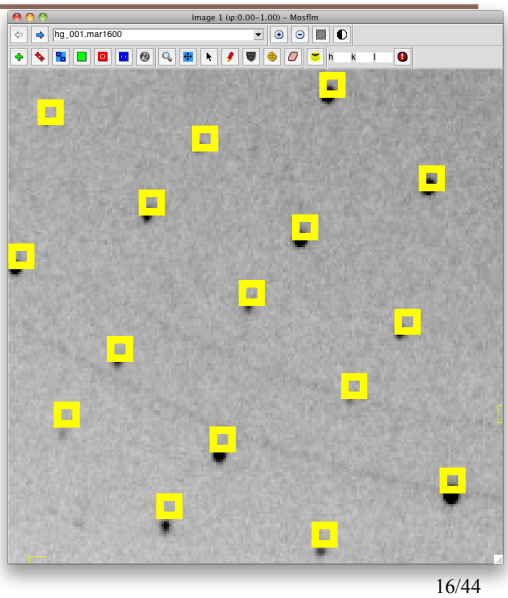
Post-refinement, on the other hand, can only be performed rigorously by using several adjacent images; the reflections used should not have missing parts. This is the reason why Mosflm needs to have several images included in each block for post-refinement, and also one why other programs leave this until the scaling and merging step, when they have data from the complete dataset.

It is necessary to be able to identify which reflections are fully recorded, and which are partially recorded for both post-refinement and integration (especially when post-refinement is performed by the integration program).

It is possible to estimate the partiality of reflections which do have missing parts (and hence carry out post-refinement), if the total intensity of fully recorded equivalent reflections are available; this will not give results which are as robust, and is usually not necessary. This method is not used in Mosflm.

Positional refinement

Minimises the discrepancy between observed and calculated ("predicted") spot positions -



16/44

We are trying to minimise the discrepancy between the observed and calculated spot co-ordinates on the detector (usually transformed to some virtual detector frame).

Positional refinement

Minimise -

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

n.b.

- rotation of crystal about phi axis has no effect on this residual so can't be refined
- cell dimensions and other parameters (*e.g.* crystal to detector distance) may be strongly correlated
- can be used to refine unit cell dimensions, crystal to detector distance, Y scale, 2 of the 3 crystal mis-setting angles, detector mis-setting angles and the direct beam position

17/44

$(X, Y)^{obs}$ and $(X,Y)^{calc}$ are the observed and calculated spot co-ordinates on the detector (usually transformed to some virtual detector frame). Pythagoras' Theorem shows why the rotation of the crystal around the phi axis has no effect here (the X and Y co-ordinates only have to lie on a circle with the beam position at the centre – *where on the circle is not defined*). 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 not using positional refinement to refine the cell dimensions.

In practice, in Mosflm, the following parameters are optimised by positional refinement:

direct beam position

crystal to detector distance

Y-scale

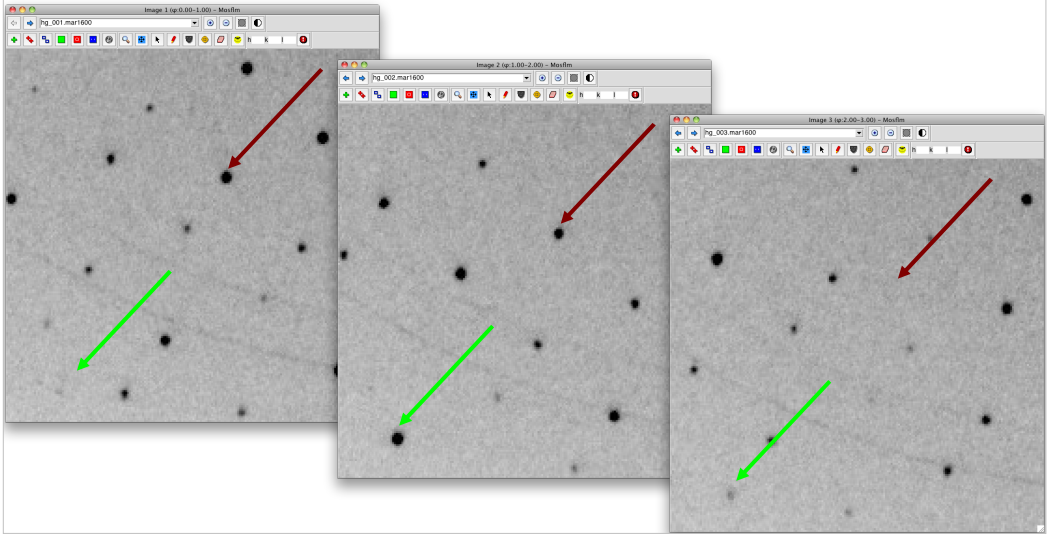
tilt & twist of the detector

(tangential and radial offsets – spinning disc detectors only)

We also report the RMS residuals of spot positions based on refining these parameters, and these values give a good indication of how stable the refinement is. Other programs also refine the cell and other detector parameters, and may use techniques such as eigenvalue filtering to improve stability.

Post-refinement or the “phi-centroid” method

Uses the intensities of reflections spread across multiple images to improve the estimates of crystal parameters.



If we have reflections that are spread across two or more images, we know that they are in the process of traversing the Ewald sphere. The relative intensities of the different parts is related closely to how close the reciprocal lattice point is to the Ewald sphere. We can use this knowledge to get more accurate information on the unit cell and other experimental parameters.

Post-refinement

Minimise -

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

n.b. we need:

- a reasonable model of the intensities for this, so it can only be done *after* integration - hence "*post-refinement*"
- a model for the "rocking curve"

can be used to refine unit cell dimensions, 2 of the 3 crystal mis-setting angles, and *either* the mosaicity *or* the beam divergence.

19/44

The "rocking curve" describes how the intensity of a reflection varies with the crystal orientation. Mosflm uses a rocking curve based on a cosine function, but other (usually symmetrical) functions could be used, *e.g.* Gaussian, hyperbolic tangent or cubic. In practice, because we are only using the intensities of reflections split into a few parts, and we are only using strong reflections, the exact nature of the function does not seem to affect the calculations greatly. A good model for the rocking curve is most necessary where we have little direct information about it, *i.e.* for data collected with coarse phi slicing. For fine phi sliced data it is easier to derive it empirically from the intensities of the partials.

R^{calc} - 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.

It is not possible to refine detector parameters using post-refinement. In Mosflm, we refine the following *via* post-refinement:

crystal "mis-setting" angles

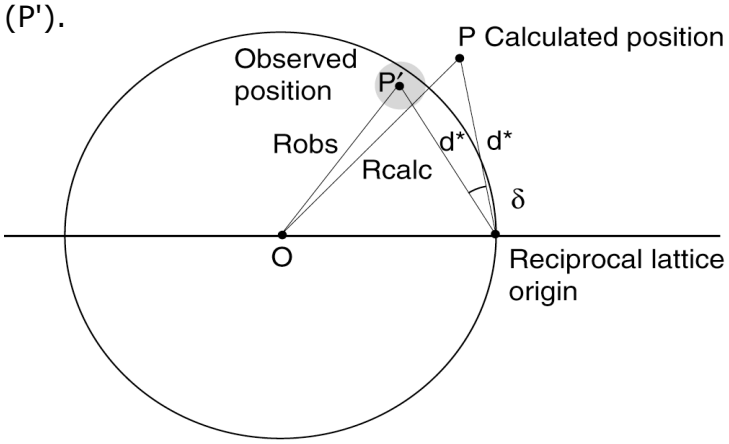
crystal cell dimensions (a, b, c, α, β, γ)

mosaicity or beam divergence

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').



20/44

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

21/44

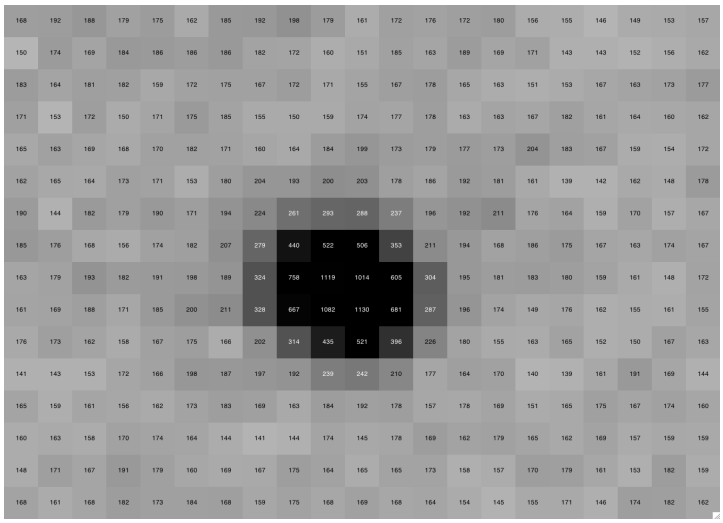
Integration is performed once the crystal and instrument parameters have been optimised by refinement.

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.



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.

23/44

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.

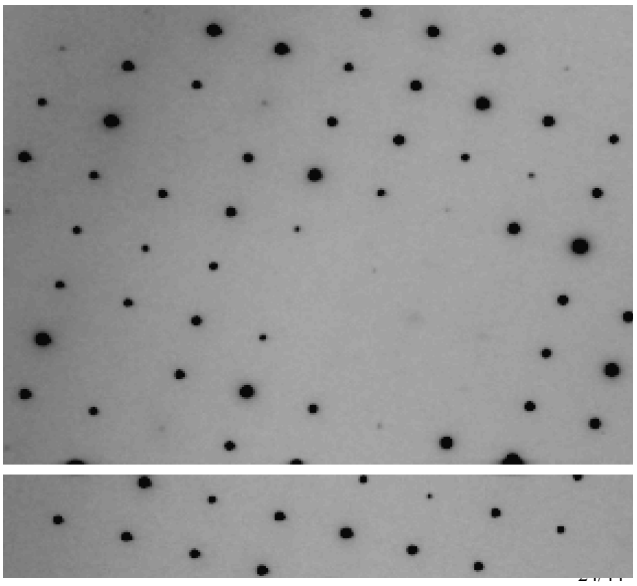
With some newer detectors that have very low intrinsic noise levels and small point-spread functions, it is probably correct to integrate using summation integration (at least for the strong reflections), especially when the background is low. However, weak spots will still have their shoulders hidden 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 assumption that spots corresponding to fully recorded reflections in the same region of the detector (and on images nearby in phi) have similar profiles.

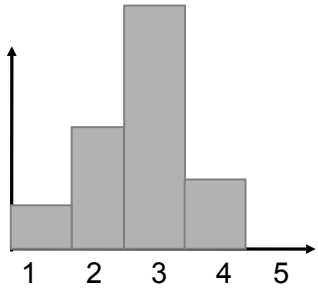


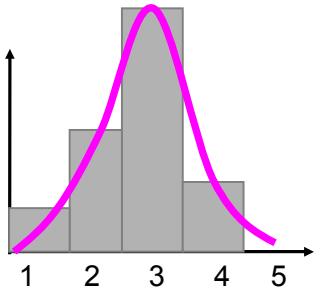
The spot shape on a detector (including its intensity profile) is a function of several physical factors – the cross-section and divergence of the illuminating radiation, the size, shape and mosaic spread of the crystal (and its orientation relative to the beam), the direction the diffracted beams exit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





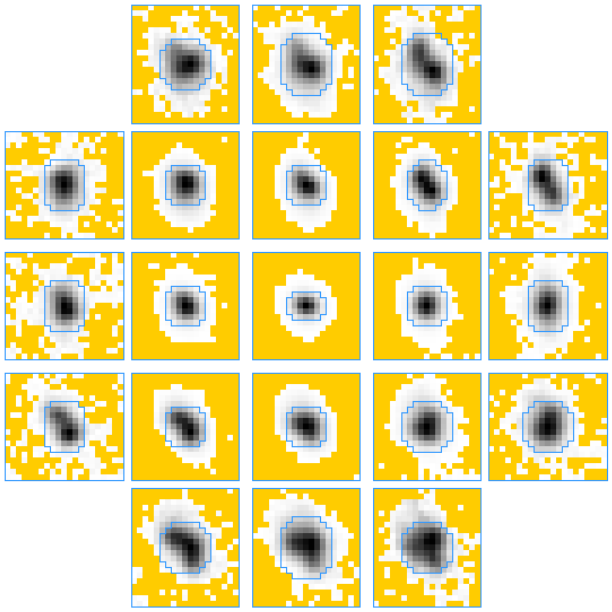
- requires accurate (sub-pixel) placement of the profile
- reduces variance for weak reflections
- should reduce random error (weak reflections)
- may increase systematic error (strong reflections)

25/44

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.

Profiles vary

- in different areas of the same image
- between images (but *not much* from one image to the next)



The profiles themselves vary from one part of each image to any other part - so the profiles on one side of an image can be very different to those on the other side.

The profiles in one region of the detector will also vary between images, but from one image to the next the change is small.

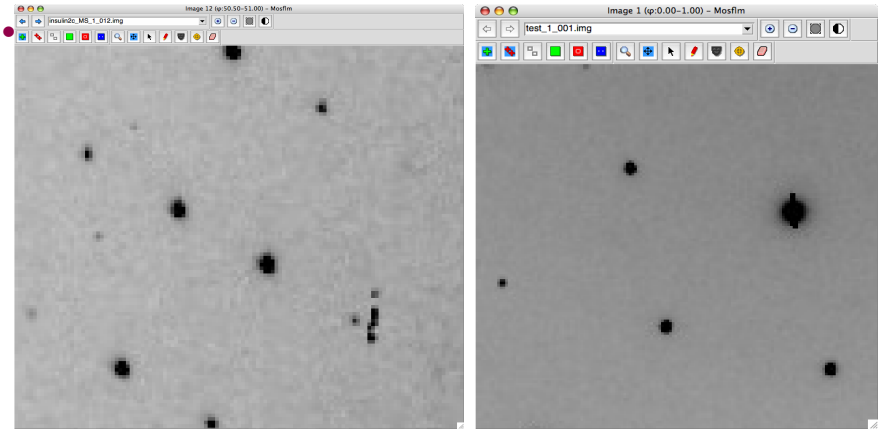
In *Mosflm*, the standard profiles (from which all profiles are calculated) are developed from spots in different parts of the detector - for low resolution datasets the image is divided into a 3x3 array, for high resolution images it is divided into a 5x5 array (for circular detectors the profile fitting areas corresponding to the non-existent corners are not calculated, as in this example).

Further, the profiles are built up across several adjacent images - usually around 10, since the profile will not change very much in this range.

The profile for each spot is calculated according to a weighted average of the standard profiles in the four (sometimes 3) closest regions, according to where in the region the spot occurs.

Other programs use different methods for calculating the individual spot profiles - *e.g. Denzo* (*HKL*) uses a "profile fitting radius" to determine which spots on the same image should be used for each spot profile.

Other improvements offered by profile fitting



identify zingers measure overloads

27/44

Because profile fitting is based on the idea of spots having essentially the same shape, it can be used to identify outliers such as cosmic ray collisions with the detector or radioactive events in the detector – these will not have the same profile.

“Zingers” are named after a Canadian statistician who studied the statistics of outliers.

Overloaded spots can also be identified (typically they have flat tops, and on some detectors “bleed” into the background), and since the expected profile is known, their intensity can be estimated - but it is better to collect data with the correct exposure and avoid the problem.

Analysing the results of integration

Check graphs - they should vary smoothly without obvious discontinuities.

- Large changes in parameters may indicate problems with the crystal or instrument.
- Look at any images corresponding to discontinuities in the graphs.
- I/σ(I) at (high resolution limit-~0.2Å) should be ≳1
- Check any warnings issued by the program; it may be best to re-process after following the advice given (all warnings given by *Mosflm* are accompanied by suggestions on how to improve the processing).

28/44

Before going on to scaling the data, it is sensible to check that the integration has not thrown up any errors. In particular, examine any graphs that the integrating program has produced. They should all vary smoothly from image to image, without any sharp discontinuities.

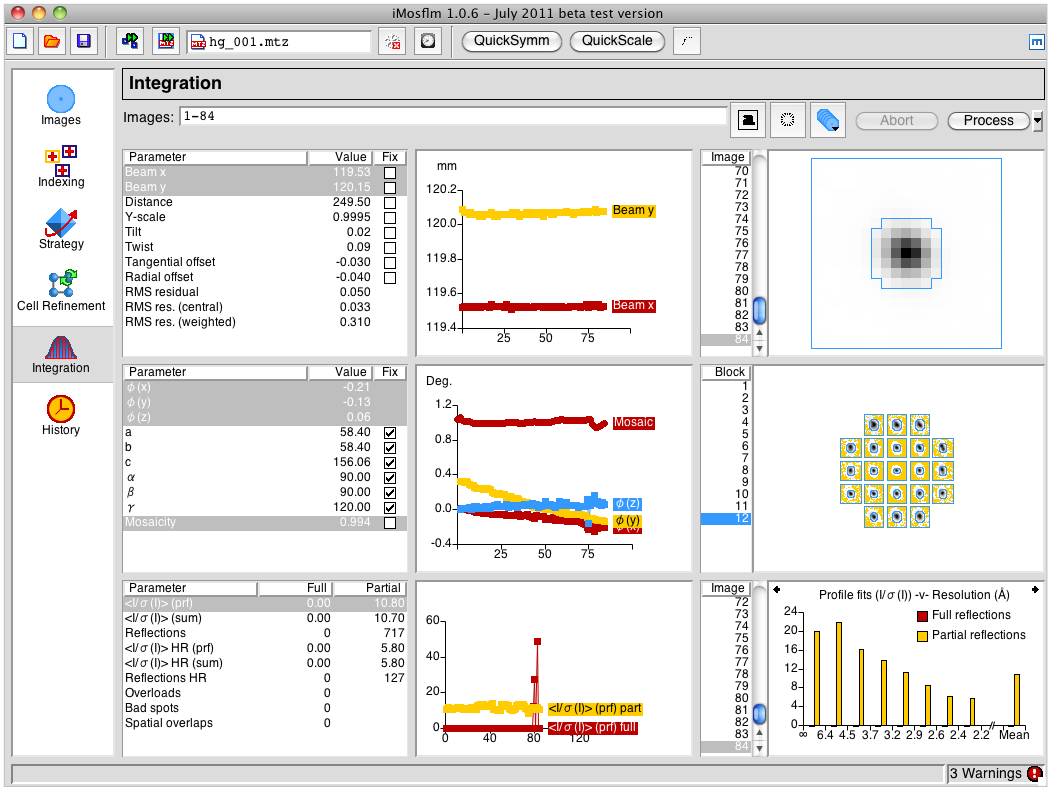
If there are discontinuities in the graphs, they often occur around the same images for different graphs. Look at any images in the region of the discontinuities and see if there is anything obviously wrong with them.

In the case that all the graphs look good until a certain point in the dataset, then the processing deteriorates, it is often an indication that too high a symmetry has been imposed on the integration, and the program cannot refine detector and/or crystal values sufficiently to keep the integration boxes well centred on the spots.

If the graphs corresponding to I/σ(I) fall gradually to lower values towards the end of the dataset, it is usually an indication that the crystal is exhibiting radiation damage.

As a first check that the data have been integrated to their resolution limit, I make sure that the average I/σ(I) for the outermost but one resolution bin is at least 1; I usually find, particularly with fine phi-sliced data where there are no fully recorded reflections, that there is significant intensity (after scaling and merging) to around 0.2Å better than the results of integration itself suggest.

Mosflm will often issue several warnings at the end of processing. Each of these is accompanied by one or more suggestions (in the main “mosflm.lp” log file) to improve the data processing.



This is an example of data processing in Mosflm where things seem to have been okay except for the last few images. There is a dip in the mosaic spread and the mis-setting angles have jumped around image 75. Also, the I/sig(I) of the fully recorded reflections has jumped from 0 to ~30 - 50 for a couple of images (because the mosaic spread for these images is lower than the rotation angle, there are actually spots identified as fulls rather than partials – all other images only have partials).

In this case it seems that there is something “odd” about the images around image 75 – it is worthwhile looking at the images near here to see if there is any obvious reason for this problem.

Scaling and merging

Scaling and merging the data is the next step following integration. It is important because:

- It attempts to put all observations on a common scale
- It provides the main diagnostics of data quality and whether the data collection is satisfactory

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

30/44

It is important to remember that integration does not provide the best diagnostics regarding integration; these are better obtained from scaling, merging and the analysis provided by the conversion from "intensities" to structure factors.

Note that none of the integration programs currently in use output raw intensities by default; all the "intensities" have been modified in some way (*e.g.* by applying corrections for the Lorentz factor and for polarisation of the X-ray beam) and are at least part-way to being more correctly termed "squared structure factor amplitudes", $|F^2|$ values.

In *CCP4*, the reflection file produced by Mosflm (or other integration programs) is best processed through

(1) *Pointless;* sorts the reflection data, analyses the Laue symmetry, and can also re-index multiple datasets to a common reference);

(2) *Scala* or *Aimless* (I currently recommend using *Aimless*, which includes many improvements not available in *Scala*); scales the intensities of equivalent reflections, merges measurements of partials, merges symmetry equivalent measurements into a single value and calculates the relevant statistics;

(3) *Truncate* or *Ctruncate*; converts the intensities into structure factor amplitudes ($|F|$) and analyses the distribution of $|F|$ values to give information on B-factor, twinning, etc..

iMosflm includes a button in its *Integration* pane to run *Pointless*, *Scala* and *Ctruncate* (using a simple default set of directives) to give an indication of data quality before leaving the integration process.

Reflections are on different scales

… because of factors related to

- the incident beam and the camera
- the crystal and the diffracted beam
- the detector

Some corrections are known from the diffraction geometry (*e.g.* Lorentz and polarisation corrections, and are applied by the integration program), but others can only be determined from the data

Scaling models should if possible parameterise the experiment - so different experiments may require different models

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

31/44

Once we have decided that the scaling and merging 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.

… incident beam and the camera

(a) variable incident beam intensity

(b) changes in illuminated volume of crystal

(c) absorption in primary beam by 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, but very small rotation angles can still cause problems with machine instabilities with similar periods to the exposure time.

32/44

The incident beam is assumed to be constant on the crystal during each image, or at least varying smoothly and slowly with respect to the exposure time. If there are large changes in the beam intensity on a time-scale similar to that of the exposure time, then this will give rise to low-quality data.

If the crystal is smaller than the X-ray beam, then the illuminated volume will remain constant provided the crystal is well-centred and does not precess out of the beam – but the overall background on the images due to air scatter will be larger than necessary.

If the beam is smaller than the crystal, then the illuminated volume will, in general, change with φ. The effects of this are indistinguishable from those caused by absorption of the primary beam by the crystal.

Variations in rotation speed and shutter synchronisation are disastrous, since they break the fundamental assumptions of the the data collection process; we assume that the crystal rotation speed is constant, and adjacent images abut exactly in φ. Shutter synchronisation errors lead to *positive partial bias*, unlike the usual negative partial bias (largely caused by an error in mosaicity).

Shutterless data collection with a constant crystal rotation speed (*e.g.* with a fast readout detector like a Pilatus) avoids shutter synchronisation errors, but does give small gaps in the data, corresponding to the readout time of the detector. This implies there is a minimum exposure time per image even for this method. If the rotation speed varies, problems could still arise.

… crystal and the diffracted beam

(e) Absorption in secondary beam - serious at long wavelength (including CuKα)

(f) 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

33/44

Scala and *Aimless* can apply an absorption correction based on spherical harmonics which attempts to model the differences in the secondary (or diffracted beam) from a spherical crystal. This can be important at longer wavelengths and for larger crystals.

It may not be possible to correct for radiation damage if it is severe and results in a non-isomorphous structure; images in datasets strongly affected by radiation damage should probably not be treated together as belonging to the same crystal. In some cases, “zero dose extrapolation” may help to rescue a dataset.

... 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.

- for example, there are known problems in the corners of detector modules, both for CCDs and Pilatus
- Calibration should flag defective pixels (hot or cold) and dead (or otherwise unreliable) regions between the tiles.
- The user should tell the *integration* program about shadows from the beamstop, beamstop support or cryocooler because it's easier than telling the scaling program!

34/44

Most detectors have been calibrated with a “flood field”, *i.e.* an even illumination with X-rays (often from a small sphere of ^{55}Fe, commonly written as Fe-55). However, both CCD and pixel array detectors like the Pilatus behave differently in the corners of the modules when detecting sharper diffraction maxima. Some of the scaling programs correct for this effect.

“Active masks” which flag defective pixels are provided by detector manufacturers, which should allow the integration programs to identify this when processing, before the scaling step.

It can be difficult to tell in scaling why there are bad regions on a detector - it is best to tell the integration program about shadows, *etc.* so that the scaling program does not need to correct for them.

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*.

However, systematic errors that are the same for symmetry-related reflections will remain.

The X-ray data collection is the process of converting structure factor amplitudes to intensities on the images

experiment

|F| ----------> I

lots of effects ("errors")

Essentially, scaling can be viewed as inverting the experiment to obtain the squared structure factor amplitudes (the square rooting is performed in the subsequent truncating step)

model of experiment

I ----------> $|F^2|$

parametrise experiment

Scaling (done correctly) will apply and refine a suitably parametrised model.

Merging

Once the observations are on a common scale we

- merge the individual parts of partially recorded reflections into complete reflections
- merge symmetry related reflections into unique observations, *e.g.*
  - monoclinic - (h k l) and ($\bar{h}$ k $\bar{l}$)
  - orthorhombic - (h k l), ($\bar{h}$ k $\bar{l}$), ($\bar{h}$ $\bar{k}$ l) and (h $\bar{k}$ $\bar{l}$)
  - *etc*.
  - Friedel pairs (or mates) - (h k l) and ($\bar{h}$ $\bar{k}$ $\bar{l}$)

- Caution!
  - Do not merge Friedel pairs in any anomalous experiment
  - Some programs (*e.g. SHELXC/D/E*) prefer unmerged equivalents in their data

36/44

The symmetry related reflections that are merged together are those that are related by rotations characteristic of the Bravais lattice symmetry, not by reflections or inversions.

Remember that any anomalous experiment will be trying to make use of Bijvoet pairs - which are Friedel mates in the presence of an anomalous scatterer - so they have to be kept separate in the merging step.

Note that merging is performed as a separate step for 2D integration programs like *Mosflm*. For 3D programs like *XDS*, the merging of partials to form full reflections is done as part of the integration step.

Questions about the data

- What is the overall quality of the dataset?
  - How does it compare to other datasets for this project?
  - How complete are the data?
  - What is the multiplicity?
- What is the real resolution? Maybe reject high resolution data?
- Are there bad batches? Individual images or groups?
- Is the whole dataset bad? Throw it away?
- Extent of radiation damage? Exclude the later parts?
- Is the outlier detection working well?
- Is there any apparent anomalous signal?
- Are the data twinned?

While a quick scan of “table 1” can give us an idea of the answers to most of these questions, it is no substitute for reading through (at least the tables in) the extensive log files produced by the scaling programs.

Annotated log in web browser (i)

"Table 1"

Summary data for Project: Gamma Crystal: Gamma Dataset: xe1a

	Overall	InnerShell	OuterShell
Low resolution limit	16.11	16.11	1.82
High resolution limit	1.79	7.86	1.79
Rmerge (within I+/I-)	0.045	0.029	0.295
Rmerge (all I+ and I-)	0.061	0.057	0.328
Rmeas (within I+/I-)	0.060	0.040	0.392
Rmeas (all I+ & I-)	0.070	0.070	0.384
Rpim (within I+/I-)	0.039	0.027	0.256
Rpim (all I+ & I-)	0.035	0.039	0.195
Rmerge in top intensity bin	0.027	-	-
Total number of observations	45306	433	2053
Total number unique	12275	150	624
Mean((I)/sd(I))	16.0	28.4	3.5
Mn(I) correlation between half-sets	0.997	0.956	0.840
Completeness	97.5	84.6	87.2
Multiplicity	3.7	2.9	3.3
Anomalous completeness	91.3	83.5	71.9
Anomalous multiplicity	1.9	1.8	1.6
DelAnom correlation between half-sets	0.349	0.848	-0.071
Mid-Slope of Anom Normal Probability	1.189	-	-

The traditional measures of quality of a dataset are the merging R-factors; after scaling, the remaining differences between observations can be analysed to give an *indication* of data quality, *e.g.*

$$R_{merge} = R_{symm} = \frac{\sum_h \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l |\langle I_h \rangle|}$$

traditional, but increases with multiplicity even if the data improves

$$R_{meas} = R_{r.i.m} = \frac{\sum_h \sqrt{(n/n-1)} \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l |\langle I_h \rangle|}$$

"redundancy independent R-factor" but larger than R_{merge}

$$R_{p.i.m} = \frac{\sum_h \sqrt{(1/n-1)} \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l |\langle I_h \rangle|}$$

"precision-indicating R-factor"

Annotated log in web browser (ii)

Estimates of resolution limits: overall
from half-dataset correlation coefficient > 0.50: limit = 1.79A == maximum resolution
from Mn(I/sd) > 2.00: limit = 1.79A == maximum resolution

Estimates of resolution limits along reciprocal lattice axes:
Along axis a*
from half-dataset correlation coefficient > 0.50: limit = 1.79A == maximum resolution
from Mn(I/sd) > 2.00: limit = 1.81A
Along axis b*
from half-dataset correlation coefficient > 0.50: limit = 1.79A == maximum resolution
from Mn(I/sd) > 2.00: limit = 1.79A == maximum resolution
Along axis c*
from half-dataset correlation coefficient > 0.50: limit = 1.79A == maximum resolution
from Mn(I/sd) > 2.00: limit = 1.79A == maximum resolution

Average unit cell: 34.16 54.82 68.05 90 90 90

Space group: P 21 21 21

Average mosaicity: 0.90

Minimum and maximum SD correction factors: Fulls 0.06 1.88 Partials 0.62 19.20

39/44

However, the correlation coefficients $CC_{1/2}$ and CC* are more statistically meaningful than merging R-factors

$CC_{1/2}$ - “Pearson correlation coefficient between random half-datasets” -

$$\rho_{X,Y} = cov\frac{(X,Y)}{\sigma_X \sigma_Y} = E\frac{[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

CC* - Correlation coefficient of our measurements with the true intensities

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1+CC_{1/2}}}$$

These give a better estimate of true resolution, and show if weak data has real information content

Truncation

- Analyses scaled & merged data according to an expected physical model
- gives statistics on intensity distribution - *e.g.*
  - Wilson statistics
  - twinning analyses
- outputs |*F*| values for use in subsequent CCP4 programs

While scaling tries to make the data internally consistent (so that symmetry related equivalents are put onto a common scale), truncation attempts to fit the overall distribution of intensities to what we expect for our crystal.

The deviations from our expected model can be analysed to indicate the overall isotropic B-factor for our structure, and also for crystal pathologies such as twinning.

Wilson plot

- Plot of

$$\ln \frac{\overline{I_{hkl}}}{\sum_i (f_i^0)^2} \quad vs \quad \frac{\sin^2\theta}{\lambda^2}$$

- For a structure with randomly distributed atoms is a straight line with a slope -2B
- Gives an estimate of the "isotropic temperature factor" of the structure
- Protein crystals do not have randomly distributed atoms

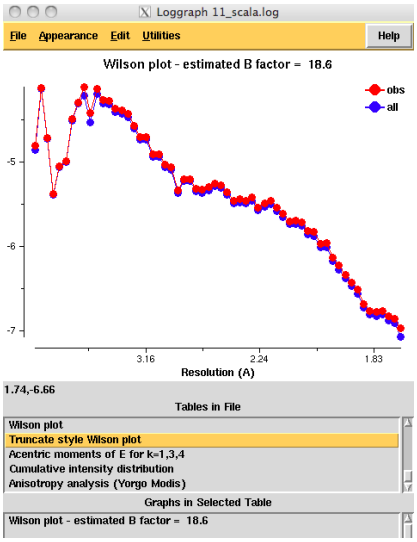
The data is divided into resolution bins;

I = average intensity of reflections in each bin

f_i^0 = atomic scattering factor squared for each atom "i"

$\sin^2\theta/\lambda^2$ is simply a convenient way of expressing the resolution which would make the Wilson plot a straight line if the crystal was composed of equal randomly distributed atoms - which is almost the case for small molecule crystals, but not for macromolecules like proteins or nucleic acids.

Intensity statistics - Wilson B factor



Average intensity falls off with increasing resolution and is associated with disordered atoms. More disorder gives a faster fall-off with intensity, a steeper slope and a larger Wilson B.

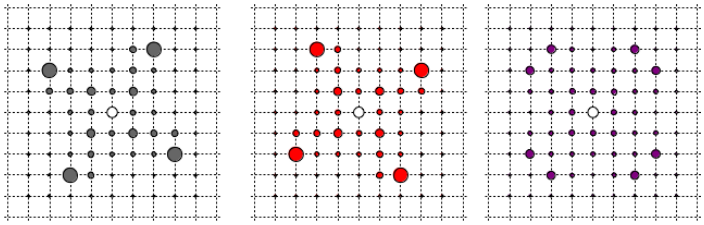
For the purpose of looking at crystal pathologies, we can ignore the variation with resolution, so we can use "normalised" intensities which are independent of resolution

42/44

Wilson statistics were developed by Arthur Wilson and reported in Nature in 1942.

Twinning

- Multiple crystal components related by geometrical operations (the twin operators)
- Can give rise to odd intensity distributions *e.g.* too few strong reflections



- Can be seen in
  - cumulative intensity plots
  - plots of $\langle I \rangle^n / \langle I^n \rangle$ and $\langle E \rangle^n / \langle E^n \rangle$ against resolution - "moments"

As can be seen here, if we have two lattices that are superimposed over each other, the probability that a strong reflection from one lattice will be superimposed over an equally strong reflection from the other is small - so in a merohedrally twinned crystal like this with two roughly equal components, there will be fewer very strong reflections than we expect (and also fewer very weak reflections). The cumulative intensity distribution plot will become more sigmoidal, but the plots of moments are more diagnostic.

The "E"s referred to above are the normalised values for F ($\sim\sqrt{I}$), taking into account the B-factor.

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

44/44