Data reduction: Pointless, Scala, Truncate

Phil Evans, MRC Laboratory of Molecular Biology, Cambridge November 2006

*Version: this document refers to programs in CCP4 version 6.0.1, **scala** version 3.2.19+*

Here are some brief notes on the data reduction steps following integration of images in eg ***mosflm***, ie processing from integrated intensities *hkl, I, σI* to merged *hkl, F, σF*

This document tells you how to run ***pointless, scala*** and ***truncate*** in the ***ccp4i*** GUI, and how to interpret the output.

Input file of reflections

We need a sorted MTZ file of unmerged intensities, usually from ***mosflm***. The file can come from other integration programs, from ***denzo***/***scalepack*** via ***combat*** (though this may restrict the scaling options in ***scala***), or from ***d*trek*** via ***dtrek2scala***.

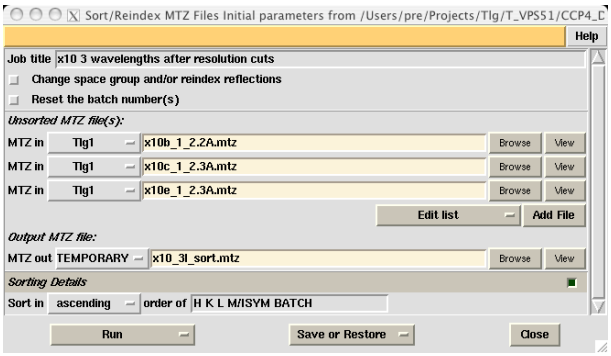
If you have a single MTZ file and you know it is in the correct space group, you can put it directly into the "Scale and Merge Intensities" task, in the "MTZ in" box. The file will be sorted (using ***sortmtz***) before running ***scala***.

If you have more than one MTZ file, then you must run the "Sort/Modify/Combine MTZ Files" task first. You can always use this task if you wish, and you also need it to reindex, to change spacegroup or to edit batch numbers ("batch" numbers are image numbers with an optional offset).

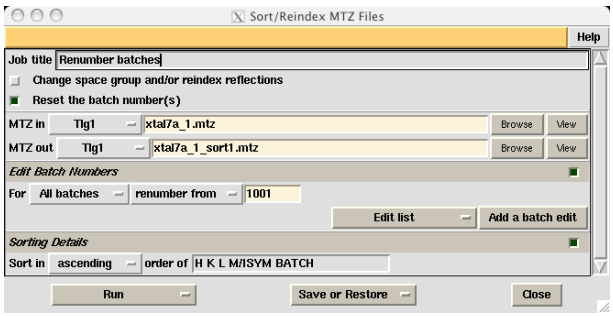
Multiple MTZ files

1. If you have multiple files belonging to the same dataset, from different runs of ***mosflm***, *eg* because you had interruptions in the image collection, they should be sorted together. Note that the BATCH numbers must be unique, so that if you had more than one image series numbered from 1, you should use the "ADD" option in ***mosflm*** to number them from say 1001, 2001 etc. If you neglected to do this, you can renumber them with the "Sort/Modify/Combine MTZ Files" task, then sort them together afterwards.

Example:



Sorting 3 files together



Renumbering batches to make them unique

2. If you have multiple datasets (eg MAD), then they should be sorted and scaled together, as multiple-wavelength data is greatly improved by scaling all the datasets together, then merging them separately. This is done automatically in ***scala*** provided that the datasets are distinctly labelled with Project/Crystal/Dataset names. This should be done in ***mosflm*** (PNAME, XNAME, DNAME commands) or ***imosflm***, making sure that the batch numbers are also unique. The dataset name is used to make column labels and filenames, so keep it short (eg 'peak', 'edge', 'remote'). Sort the files together as in the above example. Datasets can be reassigned in ***scala***, but it is much less convenient than doing it in ***mosflm***.

Determination of Laue group & space group

The indexing of the lattice in ***mosflm*** is based on the lattice geometry, with no regard for the symmetry of the diffraction pattern, which can only be determined after integration. The ***ccp4i*** task "Determine Laue group" runs the program ***pointless*** to score potential symmetry operators in the diffraction pattern, and to rank the possible Laue groups (point groups), and also to inspect axial reflections for possible systematic absences which may indicate a likely space group. Be careful, the presence of pseudo-symmetry may suggest a higher symmetry than the truth. ***Pointless*** tries to allow for this possibility, but inspection of the scores for individual symmetry elements may help to indicate the correct space group in difficult cases. *[Note that the scoring scheme in pointless is still under development and is likely to change in future versions.]*

Example:

A pseudo-symmetrical case, true space group C2, indexed in I222 (parts of the log file from ***pointless***)

```
Analysing rotational symmetry in lattice group I m m m

Scores for each symmetry element

Nelmt  Lklhd  Z-cc    CC        N  Rmeas    Symmetry & operator (in Lattice Cell)

  1   0.898   6.58   0.99   43085  0.060     identity
  2   0.876   6.36   0.95   22234  0.119 *** 2-fold l ( 0 0 1)  {-h,-k,+l}
  3   0.427   4.84   0.73   22231  0.367 *   2-fold h ( 1 0 0)  {+h,-k,-l}
  4   0.315   4.61   0.69   59250  0.388     2-fold k ( 0 1 0)  {-h,+k,-l}
```

The dyad along l (element 2) is much stronger than the other two by scores R_{meas} and correlation coefficient, and estimated "likelihood" (*Lklhd* in the table) of the correlation coefficient.

```
Scores for all possible Laue groups which are sub-groups of lattice group

Lklhd is a likelihood measure, an un-normalised probability used in the ranking of space groups

Z-scores are from combined scores for all symmetry elements
in the sub-group (Z+) or not in sub-group (Z-)
    NetZ = Z+ - Z-
Net Z-scores are calculated for correlation coefficients (cc)
The point-group Z-scores Zc are calculated
    as the Zcc-scores recalculated for all symmetry elements for or against,

CC- and R- are the correlation coefficients and R-factors for symmetry elements not in the group

Delta is maximum angular difference (degrees) between original cell
and cell with symmetry constraints imposed

   Laue Group       Lklhd   NetZc  Zc+   Zc-    CC    CC-  Rmeas   R-  Delta  ReindexOperator

> 1  C 1 2/m 1  *** 0.978   1.80  6.50  4.70   0.95  0.70   0.12  0.38   0.0 [-k+l,h,k]
= 2    I m m m  *** 0.956   5.67  5.67  0.00   0.78  0.00   0.30  0.00   0.0 [h,k,l]
> 3      P -1   **  0.699   1.40  6.58  5.18   0.99  0.78   0.06  0.30   0.0 [-h,-k,1/2h+1/2k+1/2l]
> 4  C 1 2/m 1  **  0.679   0.67  5.98  5.31   0.73  0.80   0.37  0.28   0.0 [h-l,-k,-h]
> 5  C 1 2/m 1  *   0.551   0.10  5.71  5.61   0.69  0.84   0.39  0.23   0.0 [h+k,-l,-h]
```

The scores here indicate that the true space group is C2 but the score for I222 (or $I2_12_12_1$) is close. *[future work aims to improve this discrimination.]*

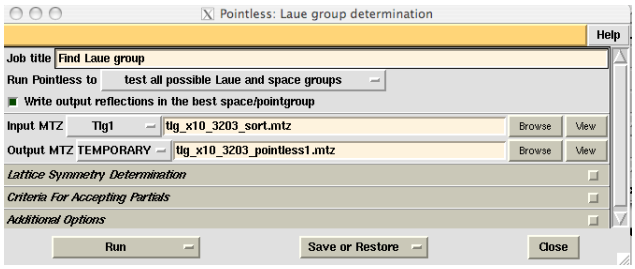
In this example there are no axial systematic absences in space group C2 to examine, but in other cases these may distinction between space groups with and without screw axes, eg between P2 and $P2_1$.

Pointless can write a new MTZ file wth the "best" space group or point group, or alternatively the space group may be changed using the "Sort/Modify/Combine MTZ Files" task. The merging operation is identical for all spacegroups in the same point-group, since spacegroup translations affect only the systematic absences at this stage. Unless you already know the spacegroup (ie for a solved structure), it is better to set the spacegroup to the one without translations (eg P222 in the orthorhombic system) since this makes the minimum assumptions.

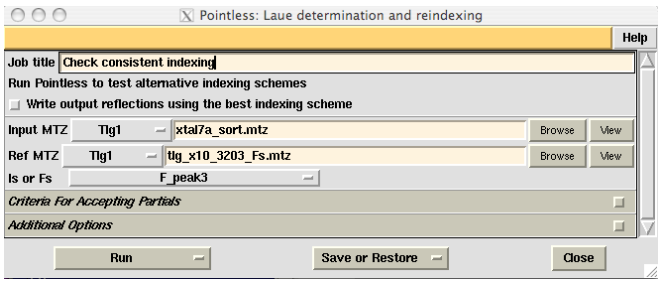
Alternative indexing schemes

In some point-groups there are more than one (two or four) valid but non-equivalent indexing possibilities. For your first crystal, you may choose any of these, but subsequent crystals must match the first. This generally arises in cases where the point-group symmetry is less than the lattice symmetry: these are the same point-groups which may lead to merohedral twinning, eg point-group P3 has four possible indexing schemes in the lattice point-group P622 (see $CHTML/reindexing.html). Within multiple datasets from the same crystal (eg MAD), you can avoid this problem by only autoindexing one set, & using the same indexing matrix for the others. Reindexing ambiguities may also arise in lower-symmetry point-groups in case of accidental coincidences or relationships between cell dimensions (eg a=b in orthorhombic).

Consistent indexing may be automated with the program ***pointless***, by giving it a merged reference file (your first choice of indexing), using the "Test Alternative Indexing" task in ***ccp4i***.



Determining Laue group



Testing alternative indexing

Options in the "Scale and Merge Intensities" task

A 3-wavelength example

Click this button if you have anomalous scatterers

Usually run ***truncate*** to convert I to F (also to put all datasets into the same file)

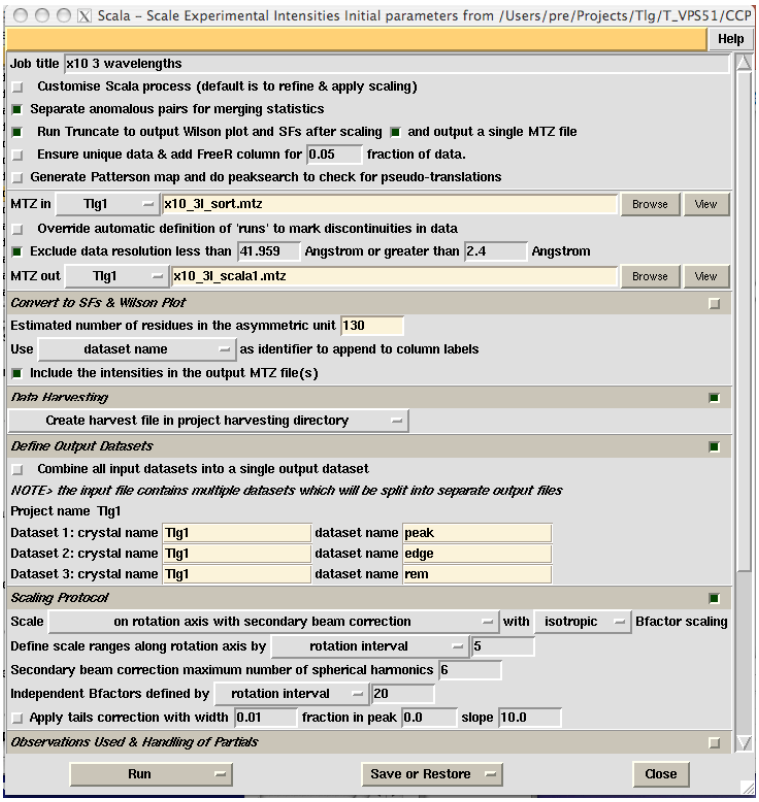
Generate FreeR set only on the 1st set, otherwise copy it

Set resolution limits (default as in MTZ file)

Information for ***truncate***

Dataset information from MTZ file: override here if required

The default scale model is recommended in most cases



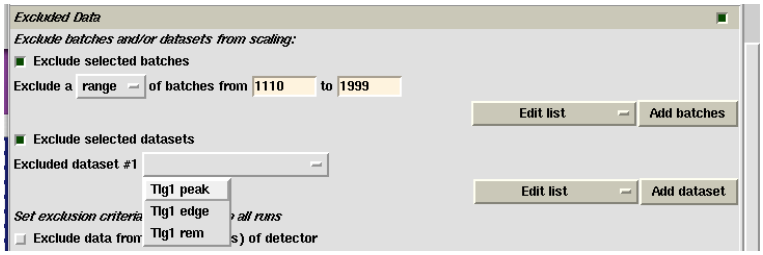
Other options

These are in panels which are closed by default.

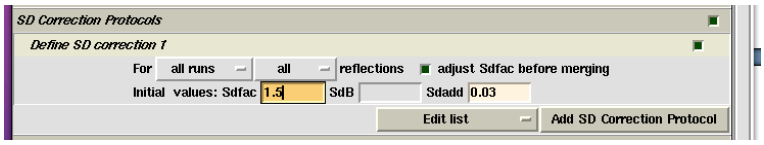
Excluding parts of data:

Batches may be excluded (list or range), or whole datasets.

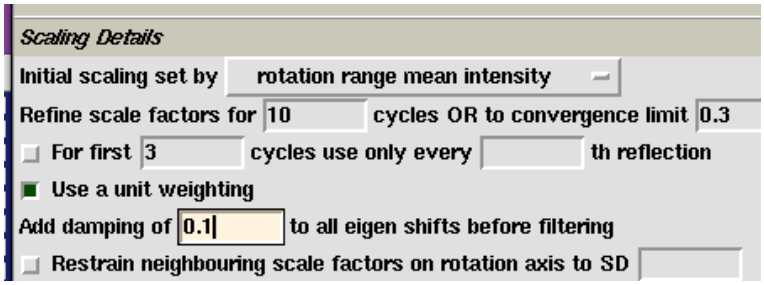
See above for resolution cut-off.



Set SdAdd parameter to value other than default 0.02



Sometimes there are problems if the scale factors have a large range (eg strong & weak passes through the data, with relative scales of eg 10), possibly causing failure with "Negative scale" or severe lack of convergence. This panel may be used to set damped shifts and possibly unit weights, which may help.



Interpreting the output

The ***scala*** logfile is rather long with many statistics, but some of them are only useful if there are severe problems. At the end of the logfile there is a summary table which contains the information useful for deposition and publication (in ccp4i, "View logfile" and click on "Show Summary"). Note that if you scale multiple datasets together, they are each merged (averaged) separately, so the logfile contains separate statistics for each dataset, as well as some correlation statistics between them.

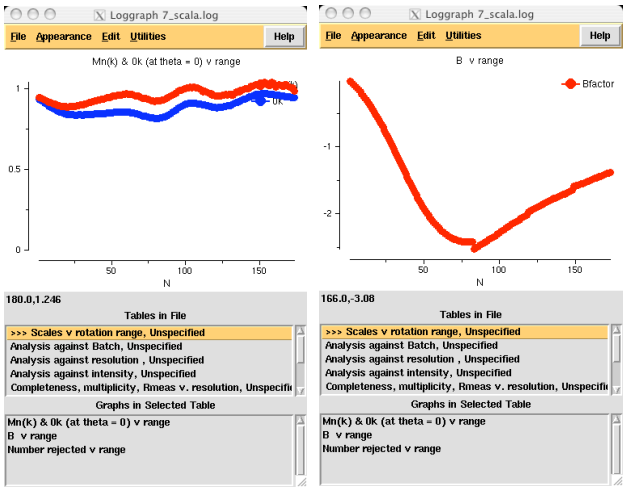
The statistics are best viewed as graphs (View Log Graphs) as illustrated below. Based on the output, you need to make some judgements about your data.

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

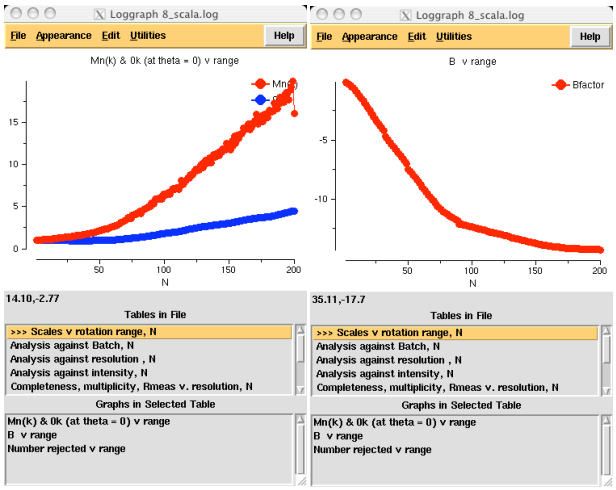
Analysis by Batch number (equivalent to time)

1. Scales & B-factor

The relative B-factor is principally a (very rough) correction for radiation damage. Data with B-factors < -10 should be treated with suspicion. Note that B is ill-determined with low-resolution data and perhaps should not be refined.



A good case: scales uniform; B-factors small

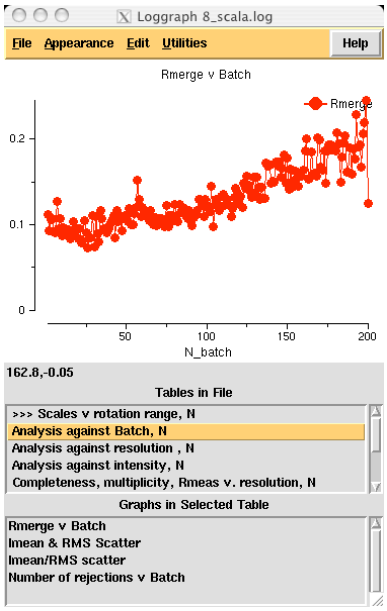


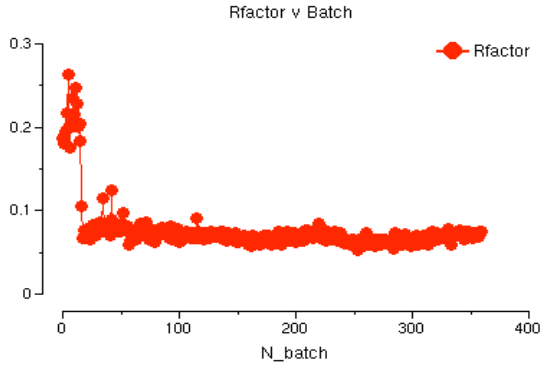
A bad case: <scale> increasing sharply; B-factors large and negative (< -10), falling rapidly. A sign of severe radiation damage

There is nothing you can do to correct for severe radiation damage, unless you have designed your experiment specifically to exploit it (extrapolation to zero-dose needs every reflection to be measured at several well-spaced time intervals).

2. R_{merge} etc

Are there any bad patches in the dataset? Examine graph of R_{merge} v. BatchNumber



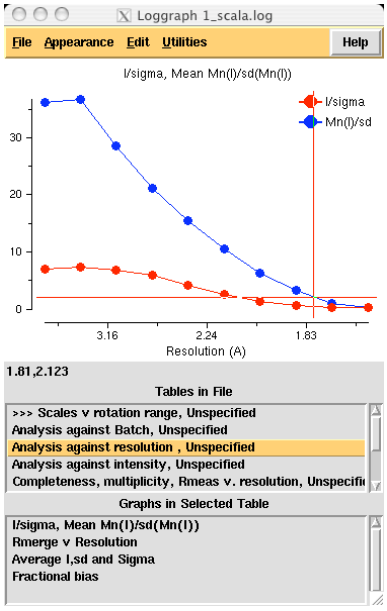


Above: something was horribly wrong at the beginning. In this case, there was poor orientation matrix at the start, fixed by re-running ***mosflm.***

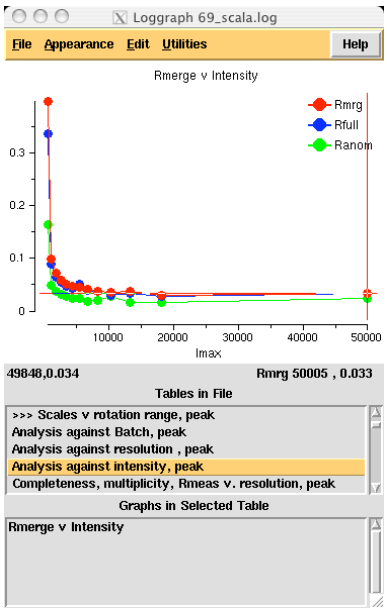
Left: steady decline in quality of data, in this case from radiation damage.

Analysis by resolution

What is the real resolution? The best guide is the average signal/noise < <I>/s(<I>) >, labelled Mn(I)/sd(Mn(I)) in table. A typical cut-off is ~2: weaker data may make a small difference in maximum-likelihood weighted refinement, but not much. It certainly will not give good experimental phases



Right: a realistic resolution cut-off is around 1.8Å

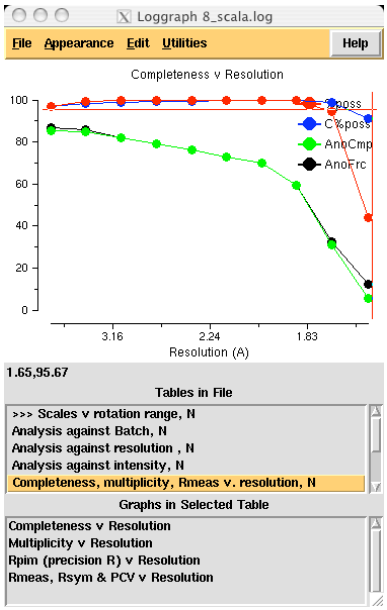


Analysis against intensity

R_{merge} is obviously always large for small intensities, but its value for the largest intensities should be in the range 0.01 to 0.04 for good data sets (0.033 in example on left). Larger values suggest that there are some worrying systematic errors.

Completeness & multiplicity

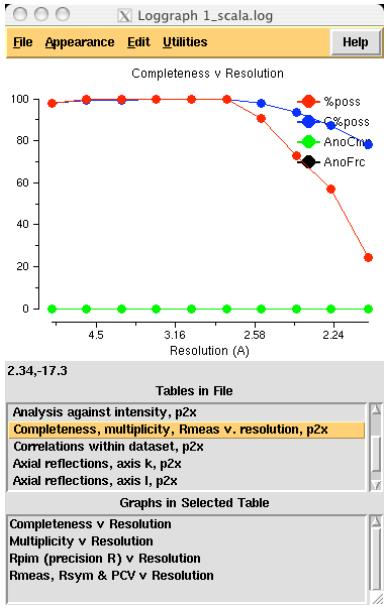
Datasets should be complete, as near to 100% as you can manage. Some loss of completeness can be tolerated in the outermost resolution bins.



Dangers:

• watch out for bad anomalous completeness (see example left). Note that in P1 you need a rotation of at least $180°+2\theta_{max}$ (better 360°) to get complete anomalous data.

• Note that by default mosflm integrates into the corners of square detectors, leading to very incomplete data at the maximum resolution. You should apply a high resolution limit (see example on right).



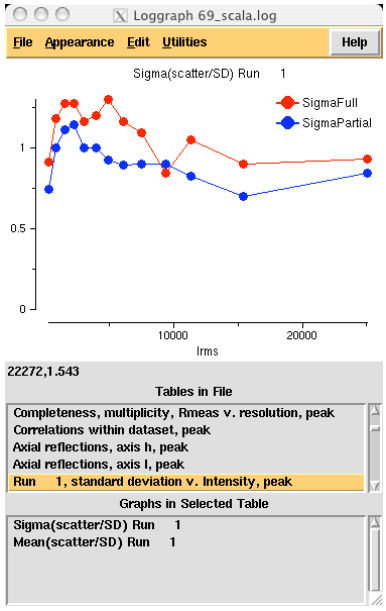
High multiplicity gives more accurate data and is essential for use of the weakest anomalous signals (eg sulphur), but note that Rmerge will tend to rise with increasing multiplicity even though <I> is more accurate. The multiplicity-weighted R_{meas} (=R_{rim}) does not suffer this problem, and R_{pim} gives a measure of the accuracy of the average I.

Analysis of Standard Deviations

Scala compares the observed scatter of symmetry-related observations around the mean with the estimated s(I), and "corrects" the s(I) by a multiplier and a fraction of <I>

$$\sigma'(I) = SdFac \sqrt{\{\sigma^2(I) + (SdAdd \times I)^2\}}$$

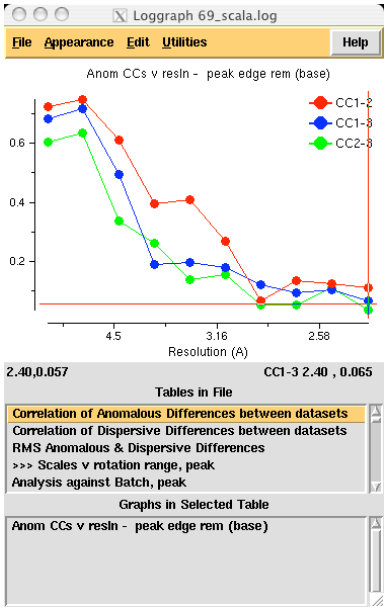
The multiplier SdFac is determined automatically (see Normal Probability Analysis below), but the factor SdAdd is not: it defaults to 0.02, which is typically fits reasonably well.



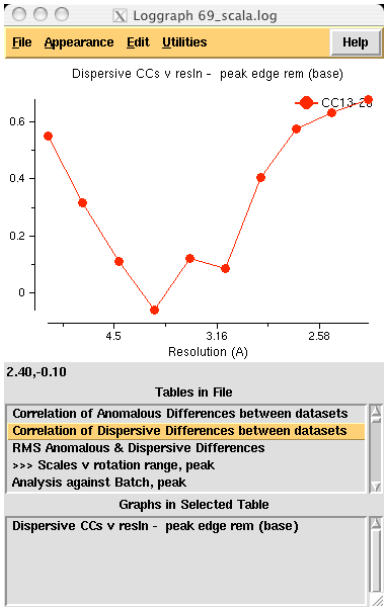
The plot on the right represents the distribution of normalised errors $\delta = (I-<I>)/\sigma(I)$ as a function of <I>. This distribution should have a standard deviation of 1.0, so the plot should be horizontal with a value = 1.0 everywhere. If it slopes upwards, then σ(I) is too small for large I, so SdAdd should be increased, or reduced if it slopes downwards. Typically, it is far from linear, since this correction is very crude, but you should try to adjust SdAdd so that the line is not too sloping.

Correlations within and between datasets

If you have multi-wavelength datasets, then the anomalous differences for each dataset should be correlated, as should the dispersive differences between them. Examination of these correlation coefficients as a function of resolution gives an indication of reliability of the signal.

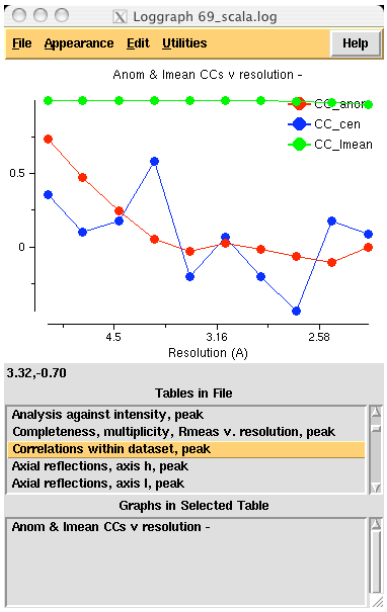


In this 3-wavelength example, the correlations on anomalous differences decline with resolution, as expected, and look good to at least 3.5Å. Bizarrely the correlation on dispersive differences goes down to zero at about 3.8Å resolution, then climb to a high value again at high resolution, which is extremely suspicious: indeed, the dispersive differences did not give much useful phasing though the anomalous differences did.

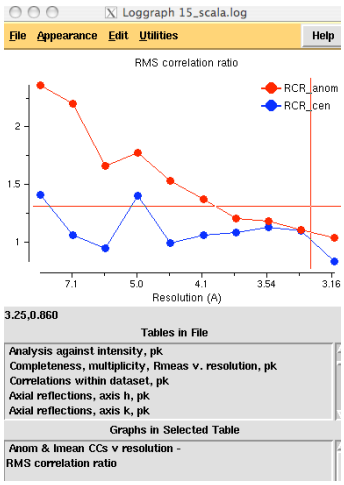


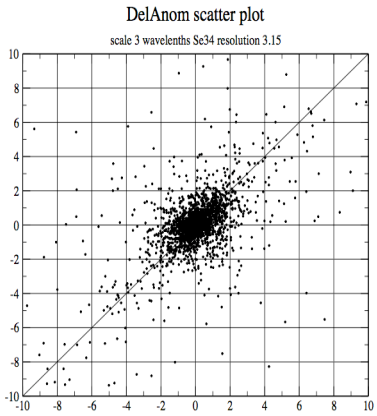
Within one dataset, an equivalent correlation analysis is done between random halves of the dataset: clearly this only works well with a reasonably high multiplicity (at least 4, preferably higher). The plot on the right is for the "peak" dataset from the same 3-wavelength set as the examples above. The anomalous correlation (red) is always lower than that for the complete dataset compared to the "edge" dataset, reflecting the lower multiplicity in the half-set. The centric data (blue) should have a correlation coefficient of zero, but the number of data is small so the error is large.

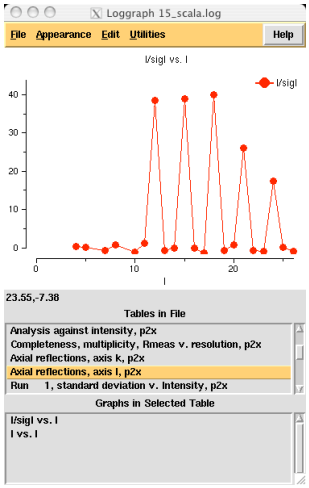
The correlation coefficient on I (as opposed to ΔI, green) is another indicator of the maximum real resolution: if it is falling rapidly at the edge, perhaps a high resolution cutoff should be applied.



Correlation of anomalous differences within a dataset are also analysed in the "correlation plot" ("..._correlplot.xmgr" in the View files from job pull-down). This is just a scatter plot of the two estimates of Δanom from the half-datasets. If they are correlated, then the scatter plot will be elongated along the diagonal (see right). The degree of elongation may be measured as the ratio of the RMS widths of the distribution along the diagonal and perpendicular to the diagonal. This "RMS correlation ratio" is plotted against resolution in the loggraph plots and like the correaltion coefficient may be used to choose a resolution cutoff for locating heavy atoms



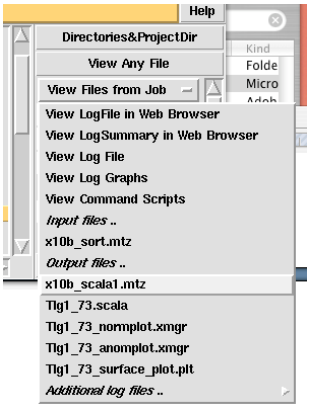




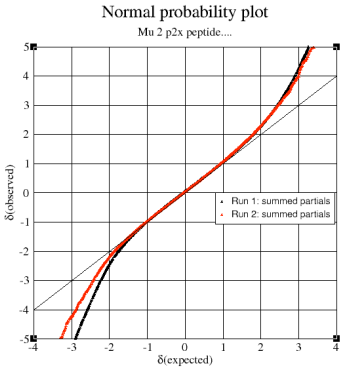
Axial reflections

These can give some indication of the systematic absences and hence the space-group, eg in the plot on the left for **00l** reflections shows them only present for l = 3n: this a hexagonal space-group, point-group P6, suggesting the space-group is $P6_2$ or (in this case) $P6_4$. Be careful of deciding the space-group too soon on the basis of insufficient data, bear in mind that you may be wrong. Non-crystallographic screw axes approximately parallel to a crystal axis may give absences which may be misleading: sometimes these may be detected in the native Patterson. Axial absences are analysed better in ***pointless***.

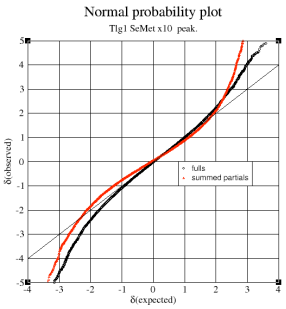
Normal Probability Analysis



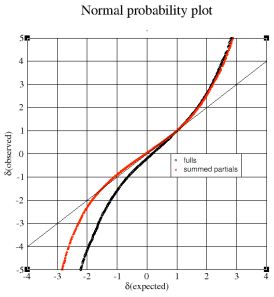
To get the normal probability plots, in "View Files from Job", click on <name>_normplot.xmgr (or <name>_anomplot.xmgr for the anomalous probability plot). These plot the normalised deviation $\delta = (I-\langle I\rangle)/s$ against what would be expected from a Gaussian (Normal) distribution, thus if the data truly followed a Gaussian distribution of errors with the estimated σ, then all the points would lie on the diagonal, with most of the observations in the centre and fewer the further away from the centre along the diagonal. Real data always has more deviant observations than predicted from a Gaussian, shown as the points curving away from the diagonal to the top right & bottom left. The SdFac correction forces the slope = 1 in the centre, but the overall truth of the error assumptions may be judged from (a) how close to the centre that the points deviate from the diagonal (b) how similar are the sub-sets of data, fulls, partials & different runs.



Above: a good example



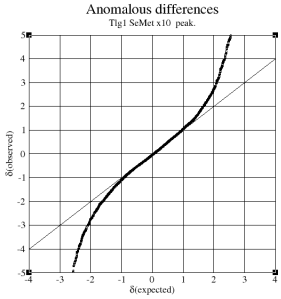
Right: not so good but tolerable



Left: poor

If the plot is poor, there is not much you can do about it, except treat it as a general indicator that the quality this dataset may not be the best.

In the case of the anomalous probability plot, $\delta = (I^+ - I^-)/\sigma$ and a slope > 1 in the centre indicates that the measured anomalous differences are greater than would be expected from the standard deviations.

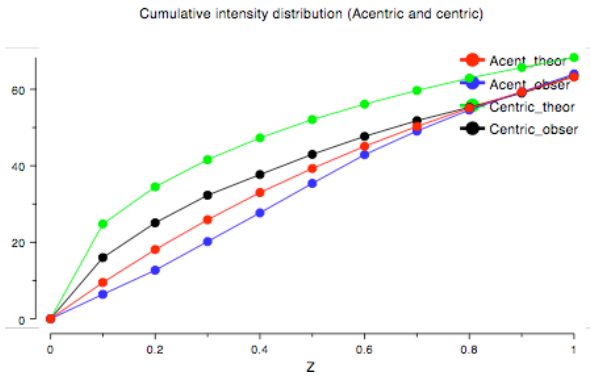


However, note the example on the left, where there appears to be no significant difference, but nevertheless there was useful phasing information, in combination with other datasets.

Conversion of intensity to structure amplitude F

This job is carried out by the program ***truncate***, which is usually done as part of the "Scale and Merge Intensities" task along with ***scala***, though there is also separate task "Convert intensities to SFs".

For perfect data, $|F| = \sqrt{I}$ but in the presence of errors, for small intensities the best estimate of F is larger than $\sqrt{I}$. ***Truncate*** estimates the "best" F based on the probability distribution of I.



It also produces some statistics, the most important of which are statistics which can be used to detect twinning. In a merohedral twin, each observed reflection is actually the sum of two different reflections related by the twin symmetry. There are apparently too few weak intensities, since a weak reflection is added to a twin-related reflection which is probably stronger (since it is unlikely to be weaker). Similarly, there are too few very strong reflections. The effect of this may be seen as a sigmoidal cumulative intensity plot (right, blue line). This plots the number of reflections with intensities less than Z, the fraction of the average intensity (a function of resolution). The same effect may be seen (less clearly) in the higher moments of E being nearer 1.0 than expected (E is the normalised amplitude, $\langle E^2 \rangle = 1.0$ at all resolutions).

The phenomenon of too few weak intensities can also arise from the spots not being properly resolved on the images: weak reflections are then over-estimated due to the tails of neighbouring strong reflections running into them.

Intensity statistics are distorted for very anisotropic data, since they compare intensities to the mean intensity, which is assumed to vary only with resolution. In that case, the cumulative intensity plot is usually above the theoretical line, and this may obscure twinning.