

1 Introduction

There are many stages where intensity data from diffraction experiments can be analysed. During the experiment the data may be integrated, giving an idea of the I/σ_I of the observations. During scaling the pointgroup may be assessed and merging statistics inspected, and after scaling the amplitudes may be compared to those expected from a Wilson distribution to give an idea of the overall “quality” of the data.

One question which is not readily answered by this process is, however, “should I be using just a subset of the measurements?”. Here we describe a new tool aimed at assisting the crystallographer in answering this question, by looking for signs of radiation damage and enabling calculation of cumulative statistics as a function of integrated dose. This allows an objective consideration of which measurements are best combined to give an optimal trade off between completeness, redundancy, resolution and radiation damage. The package consists of two programs, DOSER to include dose and exposure time in the reflection files and CHEF to perform the analysis.

2 Application

2.1 Doser

The program DOSER is designed to augment the contents of reflection files from MOSFLM prior to scaling with SCALA, allowing the addition of two columns, DOSE and TIME. The input consists simply of a list of batch numbers with the associated dose and time values, which may be derived from an analysis of the image headers. Including this information prior to scaling allows for the possibility of a more sophisticated analysis of intensity values as a function of dose rather than simply batch numbers. More importantly here, this allows the program CHEF, described below, to perform all analyses in terms of the real exposure of the crystal rather than arbitrary batch numbers.

For SAD experiments, with measurements made in a single pass with uniform exposure times, the batch, dose and time are equivalent (assuming a relatively uniform beam strength, reasonable for third generation sources.) Analysis of data from more sophisticated methods of data collection, for example wedged MAD data collection or inverse-beam (CITE:Gonzalez) may benefit from the additional information. A straightforward tool to add this metadata will therefore be valuable in these cases.

Examples on the use of the program and methods for calculating the required dose and time information will be discussed below.

2.2 Chef

Decisions about the data to include for various stages in the structure determination process are usually made on the basis of the merging statistics. In many cases, the assumption will be to include data from all images and measurements to a resolution where $\langle I/\sigma_I \rangle$ reaches a certain arbitrary cutoff. Although this may give an adequate result, the effects of radiation damage and merging of reflections may mean that better results could be achieved by using *fewer* intensity measurements. CHEF is a program designed to allow you to make these decisions and analyse scaled but unmerged intensity observations in terms of dose and symmetry, and to determine appropriate standard error corrections to give meaningful σ_I values.

2.2.1 Input Preparation

CHEF requires scaled but unmerged reflections in MTZ format, ideally from SCALA with the option `sdcorrection noadjust both 1.0 0.0` selected, to ensure that the σ_I values are those estimated from the intensity measurement only. Since SCALA will produce one reflection file per logical dataset, CHEF will allow input of multiple reflection files. If the DOSE and TIME columns have not been added, DOSER may be used at this stage to update the reflection files. Including the dose information allows CHEF to analyse the intensity values on a real scale, rather than against arbitrary batch numbers.

2.2.2 Radiation Damage Analysis

Radiation damage can give rise to systematic variations in reflection intensities, through either increasing general disorder or specific radiation induced change. If this is the case, the intensities will no longer behave as a single population of measurements, with a single mean and standard deviation. We are, therefore, interested in observing when this systematic change becomes significant. This will most likely be a function of resolution, with the highest resolution reflections affected at an earlier stage.

A statistic, R_D (Diederichs, 2006), has been proposed as a measure of radiation damage. This is a pairwise R factor between symmetry related intensity measurements, considered in terms of a dose difference D :

$$R_D = \frac{\sum_{hkl} \sum_{|D_i - D_j| = D} |I_j - I_i|}{\sum_{hkl} \sum_{|D_i - D_j| = D} \frac{1}{2} |I_j + I_i|} \quad (1)$$

In cases where there are systematic changes in the intensities, the effects will be greatest when there is a large dose difference. Generally reflections measured with small dose differences should agree well. This gives a positive gradient across the graph, taken to be an indication of damage.

It is straightforward to demonstrate that this statistic is directly related to the standard crystallographic merging R , and therefore interesting to consider the *cumulative* interpretation of this as a function of dose.

Here, we propose the use of R_{CP} , a cumulative pairwise R based on that defined above, as a measure of the usefulness of measurements in terms of a desired IU/σ . R_{CP} is defined as:

$$R_{CP}(D) = \frac{\sum_{hkl} \sum_{\max(D_i, D_j) < D} |I_j - I_i|}{\sum_{hkl} \sum_{\max(D_i, D_j) < D} \frac{1}{2} |I_j + I_i|} \quad (2)$$

Therefore, the statistic R_{CP} gives an indication of how well reflections agree up to dose D . Since this is related to the traditional crystallographic merging R it is straightforward to compute expected values for a given I/σ . Further, since the measurements are pairwise, there is no dependence on multiplicity (Cite: RMEAS papers).

[FIXME what follows is not correct - if the intensities are drawn from a Wilson distribution then this does not apply...]

For an I/σ of 2 a R_{CP} value of 0.56 is expected. However, for a single observation there is a standard deviation in this value of 0.42. If we, therefore, impose a limit of $0.56 + 0.42N^{-\frac{1}{2}}$, where N is the number of observations, we get a sensible limit on the reflected intensities to use. Since this does not involve the standard deviation estimates on the intensities, this is a robust measure. For high multiplicity data this will rapidly tend towards a value of 0.56. FIXME illustrate this here. For higher I/σ values, similar calculations may be performed. FIXME include figure illustrating expected R_{CP} vs. I/σ .

For a typical MAD structure solution, there are three resolution questions asked: the resolutions for substructure determination, phasing and phase extension / refinement. If we require for these an I/σ 10, two or more (nearly) anomalously complete I/σ 2 and a single natively complete I/σ 2 respectively, it is straightforward to determine appropriate resolutions and dose limits.

OLD

The first stage in the intensity values alone are analysed in terms of dose difference (Diederichs, 2006) giving an indication of radiation damage. The program allows the use of different start and end doses to consider along with options for binning, which allows an investigation of the effects of changing the dose limit in terms of the R_d statistic and the completeness of the data sets. The results are presented for viewing with CCP4i (ADD: Figure).

The R_d statistic is interesting, in that it is derived from pairwise comparisons of symmetry related reflections resulting in no multiplicity dependence. Absolute values of this statistic may therefore be used to assess how well observations agree. R_d is defined as:

$$R_d = \frac{\sum_{hkl} \sum_{|d_i - d_j| = d} |I_j - I_i|}{\sum_{hkl} \sum_{|d_i - d_j| = d} \frac{1}{2} |I_j + I_i|} \quad (3)$$

R_d values of ~ 1 indicate that the difference in intensity values is approximately equal to the values themselves. This would indicate either a low I/σ cutoff or some systematic variation in the intensity values. For small dose differences (d) it would be reasonable to expect that the systematic differences are small. It is therefore interesting to linearly extrapolate R_d to $d = 0$ as a function of resolution, as an indication of how trustworthy the measurements are.

In the event of radiation damaged data, there will be a connection between the amount of data included and the trustworthiness based on the extrapolated R_d statistic. Thus, actually *reducing* the amount of data may improve the overall agreement of the data set. For relatively high space-groups it may be possible to get a well agreeing nearly complete data set which will give rise to a better refined structure than including all measurements above a given I/σ_I .

Illustrate this with 2ISB data - show that the phasing statistics from two lots of 60 degrees are better than from three lots of 90. Also merging statistics and so on...

2.2.3 Symmetry Operators

There are a number of tools available for analysing unscaled intensities in an effort to obtain pointgroup information (POINTLESS, XTTRIAGE and LABELIT.RSYMOP.) These programs aim to give a summary of the agreement of reflections based on their symmetry operations. CHEF includes a similar summary based on the scaled but unmerged intensities. Symmetry operators with unusually high R values can then be investigated fully, perhaps indicating that the symmetry used in scaling was not appropriate.

Include here an illustration with 1VR9, scaling the measurements from the native in I222 then looking at the stats. This should be pretty straightforward as that is what pointless "wants".

Since this statistic is biased by the number of reflections used to determine the value, the number of observations is also reported.

2.2.4 Sigma Scaling

Once we have a data set which we believe is not badly affected by systematic variation, it is appropriate to determine appropriate multipliers for the σ_I values, to give a uniform reduced $\chi^2 \sim 1$.

Systematic variation as a function of dose or time in reflection intensities, due for instance to radiation damage, will give a characteristic shape to the

χ^2 curve as a function of dose. If we consider a linear variation in I values as a function of dose, FIXME determine this for reduced chi squared.

$$\chi^2 = \sum \frac{(I - \langle I \rangle)^2}{\sigma_I^2} \quad (4)$$

will take the form of a quadratic curve, with a minimum at a dose equal to half the total dose, as these will give the closest agreement with the rest of the data as a whole. The σ_I values should therefore be optimised to give a value of $\chi^2 \sim 1$ at this point, and the values at the dose extremes will give an indication of the cumulative radiation damage.

3 Modelling the Behavior of Cumulative-Pairwise R

To make robust decisions based on the values of R_{CP} it is necessary to have a good understanding of exactly how the statistic behaves as a function of I/σ , redundancy and number of reflections binned for undamaged data. To determine these relationships, synthetic data sets were generated with reflections drawn from a Wilson distribution with a given I_0 . N_m symmetry related intensity values were drawn from a normal distribution around I_0 with the given I/σ to represent multiple, symmetry related measurements of the same intensity, with N_u unique I_0 values from the same intensity shell.

For each collection of $N_m \times N_u$ reflections, R_{CP} was computed as defined above, with the effects of dose ignored. This calculation was repeated N_r times to obtain an estimate of $\langle R_{CP} \rangle$ and $\sigma(R_{CP})$. This process was itself repeated N_r times to obtain confidence limits on these values. The exact value of N_r was found to be unimportant, as long as it was “large”, so 100 was used.

3.1 Dependence on Intensity

The expected values for R_{CP} as a function of I/σ can be determined in a straightforward analytical manner as $\frac{1.128}{\langle I/\sigma \rangle}$, or approximately 0.56 for data with an $I/\sigma \sim 2$. This is independent of N_m and N_u . The standard deviation in this value (S_{CP}) does, however, depend on these properties. Nevertheless, for a given N_m and N_u , the ratio of $\langle S_{CP} \rangle$ to $\langle R_{CP} \rangle$ is found to be constant, for $I/\sigma > 2$. This is illustrated in Figure FIXME.

FIXME - rerun Test 6 with closer I/σ spacing to get an idea of how this behaves for very weak reflections.

$\langle S_{CP} \rangle$ is perhaps the most important measurement. For a given required I/σ and observed N_m and N_u , we can determine the significance of deviations away from the expected value of R_{CP} for a given set of reflections. Clearly, for poorly sampled measurements there will be a large random error,

while for high redundancy data the errors are likely to be negligible. The ratio of $\langle S_{CP} \rangle$ to $\langle R_{CP} \rangle$ is therefore critical to the usefulness of the R_{CP} statistic. The value of this ratio, and hence the value of $\langle S_{CP} \rangle$, is the focus of the rest of this discussion.

3.2 Dependence on Multiplicity

For a given reflection, the number of observations of R_{CP} will be equal to $\frac{N_m(N_m-1)}{2}$.

FIXME: Need to comment here on crystal symmetry and “typical cases” for multiplicity ranges.

Note well that this dependence on multiplicity will give rise to a dependence on the crystal symmetry. Give some scope for this e.g. from P1 to P222 to P23. Have examples in P1, etc.

3.3 The Final Model

The expected multiplier for the standard deviation was found to be $\sim N_u^{-\frac{1}{2}}$, with the dependency on the number of pairs as $\sim \left(\frac{1}{2}N_m(N_m-1)\right)^{-0.3}$. The dependency on N_u clearly fits with the expected behaviour for N independent observations of a parameter. Since the differences between symmetry related reflections are not independent observations, predicting the results will not be straightforward.

The final estimate for the standard deviation multiplier was found to be:

$$\frac{\langle S_{CP} \rangle}{\langle R_{CP} \rangle} = \left(\frac{1}{2}N_m(N_m-1)\right)^{-0.3} N_u^{-0.5} \quad (5)$$

To validate this, a range of N_u and N_m values were tested, with the weighted difference computed blah!

4 Program Structure

Both DOSER and CHEF are written in FORTRAN and make use of the CCP4 libraries. DOSER is a straightforward program which extends the reflection file with two additional columns of data, values taken from the command line. CHEF is rather more complicated, as there are several different analysis routines with a number of supporting routines incorporating “standard” calculations for e.g. completeness and R_{merge} . In writing these programs we took the opportunity to structure the code such that a novice who wanted to understand a procedure could do so with a straightforward inspection of the code. We therefore hope that the programs will represent a useful resource in two ways - as both a tool for users and an illustration of how to write data analysis routines.

5 Usage

5.1 Doser

The program DOSER follows the usual CCP4 convention of specifying HKLIN and HKLOUT on the command line, and has a single command:

```
batch N time T dose D
```

The batch refers in simple cases to the frame number, usually offset by some amount for MAD data sets. The time T should be the wall clock time, in some uniform time unit, since the start of the experiment. The dose will likewise be the cumulative exposure, either in terms of measured dose, simple accumulated seconds of exposure or some product of second of exposure, attenuation and beam strength - the choice is for the user. The only real requirement is that the dose indicates the order of image collection (e.g. for wedged data collection) and gives a reasonable indication of the integrated exposure of the crystal.

This program will work correctly with any reflection file which contains a column BATCH - the TIME and DOSE columns will be set to the corresponding values given in the input. Missing values will be set to -1. The data may be accumulated by inspecting all of the image headers recorded from the crystal. Since these usually contain timestamps and exposure times, it is straightforward to derive a cumulative dose as a function of timestamp and assign this to the appropriate batch numbers.

5.2 Chef

The input to chef is rather more complicated, and is designed to be run interactively. Once again, the input reflection file should be specified as HKLIN. The following options are available:

```
labin BASE=BATCH|TIME|DOSE
```

```
range width widthvalue max maxvalue  
resolution high low
```

```
anomalous on
```

```
bins nbins
```

```
print rsym rd rmerge
```

The first option defines the baseline to use for analysis - BATCH will usually be fine for straightforward single-pass SAD data, but for MAD, multi-pass and inverse-beam data using the DOSE is more appropriate.

The next options define the area of measurements to consider - to a given dose limit, in given width bins, within a given resolution range. This is used to experiment with different subsets of measurements. The Friedel pairs may be considered separately or together, as determined by the anomalous flag.

The last two options determine what is printed - the bins give the number of resolution bins to consider, uniformly spaced in $\frac{1}{d^2}$, while the print command determines which analyses are performed - rd and rsym are appropriate in early stages, rmerge more useful for determining useful error inflation parameters once the subset of measurements has been chosen.

FIXME add a cumulative completeness for different data sets as a function of BASE - would this be straightforward?? E.g. print completeness vs. BASE for each input reflection file. Challenging... actually not really - just need to use the bins width and max, and accumulate the completeness on this based on that!

6 Examples & Results

Here give a negative before providing a positive - discuss the results for Chris Nielsen's Myoglobin data and show how there is essentially no radiation damage and that the measurements agree well to a high resolution.

6.1 Radiation Damage

JCSG structure 2ISB for this one.

6.2 Symmetry Operations

JCSG structure 1VR9 for this one.

6.3 Chi Squared

???

7 Discussion

7.1 Application to Automation

The programs described here have been developed with automation in mind, to enhance the expertise in the automated data reduction pipeline XIA2.

7.2 Comparison to Zero-dose

FIXME CITE DIEDERICHS and or RAVS.

While zero-dose extrapolation exploits some of the same techniques which are discussed here, there is a fundamental difference. CHEF provides advice on the best subset of unmodified intensity values to exploit for your experiment, without modifying them in any way. Further, if there are enough measurements to get reliable extrapolation then it would suggest that a complete data set could be assembled without including all of the measurements - and therefore without making any assumptions about the form of the decay curve. Finally, appropriate subsets of data can be used for a range of tasks, for example substructure determination, phase calculation and refinement, by optimising anomalous and dispersive differences or native intensities separately.

7.3 Symmetry

The usefulness of CHEF as an analysis tool increases with symmetry. For triclinic cases there are usually too few symmetry recorded measurements to give a useful comparison, and completeness is a substantial issue. For high symmetry spacegroups the quantity of data can frequently be substantially reduced with limited consequences for the data completeness. In such cases this analysis is more valuable.

8 Acknowledgements

GW would like to acknowledge BBSRC Grant ... and EU Framework 6 ... BioXHIT for support in developing these tools, as well as Steve Prince for the inspiration for the program CHEF by comments about “cooking the books”.

A Procedures

A.1 Cumulative Completeness

For each HKLIN file an two integer arrays were available with a sufficient space for all bins. During accumulation of reflections, the lowest BASE recorded for each h, k, l was stored. At the end of reading all symmetry mates for this reflection the number of reflections for each bin above the minimum was incremented. Dividing the total in $I+$ and $I-$ by the number of possible reflections (accounting for centric reflections) gives rise to the cumulative completeness.