Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation

P.H. Zwart, R.W. Grosse-Kunstleve & P.D. Adams
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, BLDG 64R0121, Berkeley California 94720-8118, USA – Email: PHZwart@lbl.gov; www: http://cci.lbl.gov

Xtriage

A command line utility that allows the user to rapidly assess the quality and specific idiosyncrasies of an X-ray dataset has been developed. The program, called *Xtriage*, combines the twin analyses tools as described in a previous CCP4 newsletter (Zwart, *et al.*, 2005) with other data quality indicators.
In the following sections, the various steps in the characterization of an X-ray data set as carried out by *Xtriage*, are given.

Unit cell content analyses

The number of residues or nucleotides in the asymmetric unit is estimated on the basis a prior distribution of the solvent content, similar to those used by Kantardjieff & Rupp (2003). In contrast to the distributions obtained by Kantardjieff & Rupp, the distribution used in *Xtriage* is not conditioned on resolution and describes the probability of a given solvent content rather than a Matthews coefficient. Ideally, the estimate of the number of monomers in the asymmetric unit on the basis of the solvent content should be combined with an analysis of a self rotation function. The estimate provided by the current implementation, is however sufficient in most cases.

Anisotropic Wilson scaling

The anisotropic Wilson B tensor is estimated as described by Popov & Bourenkov (2004) and Zwart *et al.*, (2005). The resulting estimate of the overall B tensor is used to correct for anisotropy in the data. The likelihood based Wilson scaling routines have furthermore been shown to be less sensitive to resolution truncation than classic Wilson analyses, as can be seen from Figure 1.

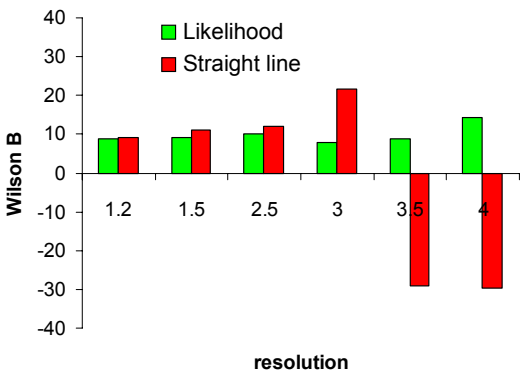

Figure 1: Comparing likelihood based Wilson scaling with classic (straight line) Wilson scaling by progressive truncation of the high resolution limit.

Outlier rejection

Possible outliers are identified using a procedure similar to that as described by Read (1999). The method is based on the distribution of the largest order statistic (Dudewiwicz & Mishra, 1988) of the intensities. The cumulative distribution of the intensities, assuming an acentric Wilson distribution of the amplitudes, is equal to

$$F(z) = 1 - Exp[-z] \qquad (1)$$

For a set of N_{refl} normalized intensities, $\{z\}$, the largest value, denoted by $z_{\max}$, out of the set $\{z\}$, is then distributed as

$$F(z_{\max}) = \left(1 - Exp[-z_{\max}]\right)^{N_{refl}} \qquad (2)$$

The *p-value* of the associated largest normalized intensity is then equal to

$$p(z_{\max}) = 1 - \left(1 - Exp[-z_{\max}]\right)^{N_{refl}} \qquad (3)$$

and quantifies the probability that the largest normalized intensity in a data set of size N_{refl}, is equal or larger than $z_{\max}$. In this manner, the outlier analysis takes into account the size of the dataset. Given for example a data set with 10000 reflections, one can be 99% sure that the largest E-value has a magnitude not greater than 3.72. However, for a data set containing 1 million observations, the latter limit is equal to 4.29. In *Xtriage*, reflections with a p-value lower or equal than 1% are considered potential outliers.

Wilson plot analyses

A Wilson plot analysis as performed by *ARP/wARP* (Morris *et al.*, 2004) has shown to be a powerful sanity check for the quality of macro-molecular data. Within *Xtriage*, an empirical Wilson plot is used that has been obtained using over 2500 high resolution data sets obtained from the *EDS* (Kleywegt *et al.*, 2004). Empirical Wilson plots for both protein and DNA/RNA structures are present, and can be used to construct expected intensity profiles for protein-DNA or protein-RNA complexes as well. Figure 2 illustrates the difference between the mean intensity of a protein structure and a DNA/RNA structure.

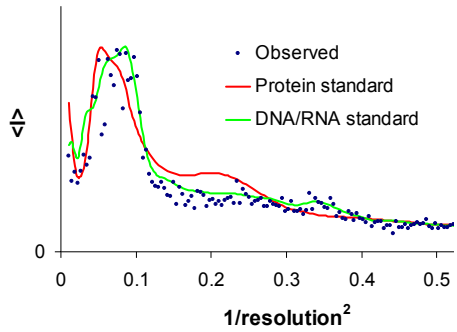

Figure 2: Mean intensity as a function of resolution. In red, the 'standard' protein mean intensity is shown. The 'standard' DNA/RNA curve is shown in green. In blue, the observed mean intensity of a sample DNA structure is shown.

Ice ring detection

If for some reason there was a significant build up of ice on the crystal during data collection, or if no suitable cryo-conditions were available, the diffraction from ice ("ice rings") can have a detrimental effect on the data. If the ice rings are sufficiently diffuse, data reduction programs are capable of subtracting the ice intensities by evaluating the background around the diffraction spots of the crystal. However, if the ice rings are too sharp, the variation in the background may be too large to determine a meaningful intensity to subtract. These difficulties in the diffraction images can result in artifacts in

the merged data, or, when suspicious observations are properly rejected, result in a low completeness in certain resolution ranges. The ice ring detection procedure in *Xtriage* monitors ice ring sensitive resolution areas (Garman & Schneider, 1997) and reports any signs (a local low completeness or a local spike in mean intensity) of ice rings.

Completeness analyses

In order to quickly assess the completeness of the data at low resolution, a completeness analyses is carried out using data up to 5 Å. This analyses is not actively used in further processing of the data, but is provided as a basis for manually judging the low resolution completeness to a greater detail than possible given output of most data reduction or scaling programs.

Analyses of anomalous signal

The presence of anomalous signal is quantified by reporting the fraction of Bijvoet pairs for which the absolute intensity difference is significantly larger than zero, a quantity denoted as the measurability (Zwart, 2005). The resolution range where the measurability lies between 6 and 3% is reported as the resolution range to where the anomalous signal is expected to be useful in substructure solution.

Detection of pseudo translational symmetry

Pseudo translations are located and identified by inspection of the Patterson function using data truncated to 5 Å. Prior information of the distribution of Patterson peak heights in data without pseudo translational symmetry (Zwart *et al.*, 2005) is used to judge on the presence of translational symmetry.

Twinning tests independent of twin laws

Classic indicators of twinning such as the intensity ratio $\langle I^2 \rangle / \langle I \rangle^2$ and the Wilson ratio $\langle F \rangle^2 / \langle F^2 \rangle$, as well as the cumulative normalized intensity distribution are provided. Besides these tests, the L statistic is computed (Padilla & Yeates, 2003), using a Miller index partitioning scheme based on the possible location of pseudo translational symmetry (Zwart *et al.*, 2005).

Derivation of twin laws

Twin laws are determined from first principles as described by Grosse-Kunstleve et al. (2005). First, the highest lattice symmetry is determined given the unit cell parameters (Grosse-Kunstleve et al., 2004). Next, the coset-decomposition algorithm as proposed by Flack (1987) is used to enumerate the twin laws, given the space group of the merged data. The derived twin laws are further classified as *merohedral* if the twin operator is a member of the Euclidian normalizer (Koch. & Fischer, 1983) of the given space group of the data, and as *pseudo-merohedral* otherwise. For example, pseudo-merohedral twin laws arise if the space group of the data is monoclinic, but the lattice symmetry is orthorhombic, as is the case for PDBID 1r3h (Rudolph *et al.*, 2004).

Twinning test dependent on twin laws

Using the twin laws as derived above, twin law dependent tests such as the H-test and Britton test are performed. The resulting estimates of the twin fractions, together with the

results from the L-test are used to provide clues as to whether or not the data is likely to be twinned, has a potential non crystallographic symmetry axes parallel to a putative twin law, or if it has been indexed in the wrong space group.

Most results of *Xtriage* can be viewed using the ccp4i program loggraph.
An *Xtriage* tutorial is available online at http://www.phenix-online.org/tutorials.

Fest: ΔF and F_A estimation.

The substructure solution engine *HySS* (Grosse-Kunstleve & Adams, 2003), has shown to be a powerful, easy to use and robust tool that allows the user to quickly solve substructure when SAD data is available. Until recently, external tools had to be used to generate ΔF or F_A values given data from substructure solution scenarios such as SIR(AS), RIP(AS) or MAD.

With the goal of furthering automation and integrating basic crystallographic tasks for solving substructures under various experimental scenarios, a number of relative scaling routines have been implemented. The relative scaling and outlier rejection tools needed for ΔF or F_A calculation will be highlighted in the remainder of this report.

Relative anisotropic scaling

An anisotropic scale factor is determined by minimizing the target function:

$$Q = \sum \frac{(I_j^{x1} - k^2 I_j^{x2})^2}{(\sigma_j^{x1})^2 + k^4 (\sigma_j^{x2})^2} \qquad (4)$$

where k is the anisotropic scale factor (see Zwart et al, 2005 eq.) to be determined. The super scripts $x1$ and $x2$ are dataset identifiers. Amplitude based target functions, as well as target functions without the use of standard deviations, are available. The optimization is carried out using second derivatives optimization methods currently under development in the *CCTBX* (http://cctbx.sourceforge.net).

Local scaling

Local scaling is an effective tool to reduce differences between data sets due to absorption or other sources of systematic errors (Matthews & Czerwinski, 1975) and can play an essential role in obtaining suitable isomorphous, dispersive or anomalous differences needed for a successful substructure solution. Three local scaling functions have been implemented:

1. local least squares target using either intensities or amplitudes (Matthews & Czerwinski, 1975, expression 5):
   $$k_{local} = \sum_{\mathbf{h}} X_{\mathbf{h}}^{x2} X_{\mathbf{h}}^{x1} \Big/ \sum_{\mathbf{h}} \left(X_{\mathbf{h}}^{x2}\right)^2 \qquad (5)$$
   where X denote ε corrected intensities or amplitudes.
2. The ratio of local intensity or amplitude moments (Terwilliger, personal communication) :
   $$k_{local} = \sum_{\mathbf{h}} X_{\mathbf{h}}^{x1} \Big/ \sum_{\mathbf{h}} X_{\mathbf{h}}^{x2} \qquad (6)$$

3. The scale factor estimation procedure developed by Nikonov (1983), implemented in a local scaling framework:

$$k_{acentric}^{-1} = \frac{2}{N_{acentric}} \sum_{\mathbf{h}} F_{\mathbf{h}}^{x2} / F_{\mathbf{h}}^{x1} - \left(\frac{1}{N_{acentric}} \sum_{\mathbf{h}} \left(F_{\mathbf{h}}^{x2}\right)^2 / \left(F_{\mathbf{h}}^{x1}\right)^2 \right)^{1/2} \quad (7)$$

$$k_{centric}^{-1} = \frac{1}{N_{acentric}} \sum_{\mathbf{h}} F_{\mathbf{h}}^{x2} / F_{\mathbf{h}}^{x1} \quad (8)$$

and

$$k_{local} = \left(N_{acentric} k_{acentric} + N_{centric} k_{centric}\right) / \left(N_{acentric} + N_{centric}\right) \quad (9)$$

This scaling procedure is designed to take into account the effect that the expected intensity for a derivative is not exactly equal to the expected intensity of the native.

The summation in the three expressions above is carried out around a local sphere around each reflection. Preliminary tests indicate that for reasonable to good data, no specific local scaling method is preferred. However, further analyses are required to make any definite conclusions.

Outlier detection

The presence of outliers in a difference dataset can have a detrimental effect on the success of substructure solution. A few large, but incorrect F_A or ΔF values can dominate the Patterson function, E-maps or triplet relations used to determine the substructure.
Outlier detection can be performed using various protocols and criteria. Two following outlier detection schemes are available to the user, both giving similar results:

1. Reject reflections for which $|\Delta F| < C_\sigma \sigma_{combined}$ where

   $$\sigma_{combined}^2 = \left\langle \Delta F^2 \right\rangle_{d^{*2}} - \left\langle \sigma_{\Delta F}^2 \right\rangle_{d^{*2}} + \sigma_{\Delta F}^2 \quad (10)$$

   The subscript d^{*2} denotes that an averaging over a resolution shell has been carried out. C_σ is a constant set to 3, but can be changed by the user.

2. Reject reflections for which

   $$F_{\mathbf{h}}^{x1} < C_\sigma \sigma_{\mathbf{h},obs}^{x1} \text{ or } F_{\mathbf{h}}^{x2} < C_\sigma \sigma_{\mathbf{h},obs}^{x2} \quad (11)$$

   and

   $$|\Delta F| > C_{rms} \left(\left\langle \Delta F^2 \right\rangle_{d^{*2}} \right)^{1/2} \quad (12)$$

   C_σ is a constant set to 3, C_{rms} is set to 4. These values can be changed by the user, if desired.

After potential outliers have been identified, they are removed from the reflection list, and the data is rescaled. This scheme can be iterated until no more outliers are detected, or can be carried out a fixed number of times. Optionally, the value of C_{rms} can be increased for subsequent outlier rejection cycles.
The first outlier rejection scheme is reminiscent to the once used in Solve (Terwilliger, personal communication). The second set of criteria is similar to those present in *CNS* (Grosse-Kunstleve & Brunger, 1999) and *HySS* (Grosse-Kunstleve & Adams, 2003).

F_A estimation

The scaling and outlier routines are combined in a program that allows the calculation of ΔF values or the estimation of F_A values. The program, called *Fest*, is a command line driven utility that currently supports ΔF of F_A value estimation for phasing scenarios such as SAD, SIR(AS), 2 wavelength MAD as well as RIP(AS).

```
scaling.input {
  basic {
    n_residues = None
    n_bases = None
    n_copies_per_asu = None
  }
  xray_data {
    unit_cell = None
    space_group = None
    reference {
      file_name = None
      labels = None
    }
  }
  output {
    log = fest.log'
    hklout = 'fest.mtz'
    outlabel = '_A'
  }
}
```

Figure 3: basic input parameters for *Fest*.

The F_A value estimation algorithm for MAD data is currently based on the work of Singh & Ramasheshan (1968) and Kingston (2001) and is limited to 2 wavelength MAD. The CNS based method of averaging isomorphous and anomalous difference Pattersons for obtaining F_A estimates, is also available (Grosse-Kunstleve & Brunger, 1999). Other methods are currently under development.

A large number of options in *Fest* are accessible to the user via a comprehensive parameter file, using the *PHIL* module of the CCTBX (Grosse-Kunstleve *et al.*, 2004; see also figure 3). The control over the behavior of *Fest* is very similar to the macromolecular structure refinement module in *PHENIX* (Adams *et al.*, 2004). The user has full control over which relative and local scaling targets will be used, as well as over outlier detection parameters. Most of these choices are however hidden by default, unless the *expert_level* is set so as to make all customizable variables visible. The basic input parameters for a SAD experiment can be found in Figure 3. Although default settings should be sensible for all scenarios, the full control the scaling procedure and subsequent F_A calculations could be crucial in unusual cases.

Automatic re-indexing

For each data set supplied (other than the first data set given), all possible re-indexing matrices are derived as described by Grosse-Kunstleve et al. (2005). Each data set is automatically re-indexed using the matrix that maximizes the correlation of amplitudes.

Examples: RIP

The data sets 2blu and 2blr, the 'before' and 'after' data sets from a RIP experiment (Nanao *et al*, 2005), were used to assess the behavior of *Fest* and *HySS* when dealing with RIP data.

The RIP scaling strategy is equal to the scaling strategy for SIR, albeit with an option to downscale the 'after' dataset by a user given constant k (controlled by the keyword *nsr_bias*), in line with the work of Nanao, Sheldrick & Ravelli (2005).

Interestingly, without downscaling the 'after' data set, *HySS* was able to locate 6 'super-sulfurs' out of 8 in total, by using data with a high resolution limit anywhere between 4 and 3 Å. The effect of downscaling on the number of sites found by *HySS* that correspond to peaks found in a difference map is shown in figure 4. These results were obtained by using only anisotropic least squares scaling on intensities using the target function shown by expression (4). The optimal down weighting factor for this particular dataset lies around 0.995, slightly larger than the value reported (0.972) by Nanao *et al.* (2005).

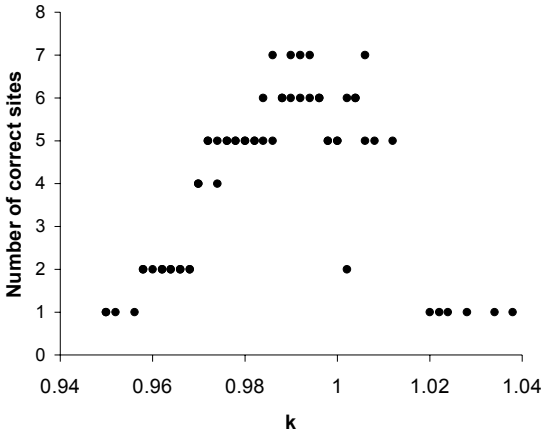

Figure 4: The influence of down weighting the 'after' dataset on the number of correct sites found. The 'optimal' down weighting factor k seems to lie around 0.995.

Note that various factors influence estimated value of the final scale. In *Fest*, these are the type of target function used, whether to scale on amplitudes or intensities or whether to use or ignore experimental standard deviations. The changes in the final scale are usually in the order of a few percent.

Given the described behavior of the scaling algorithms in *Fest* as well as the presence of a different scaling algorithm used in *SHELXC*, used by Nanao *et al.* (2005), combined with different resolution limits in substructure solution, it is not surprising that different values of 'optimal' down weighting are obtained by different programs.

Example: MAD

A MAD data set of Acyl-CoA Thioesterase II (Li et al, 2005), was used to test the implementation of the Singh & Ramaseshan F_A value estimation procedure available in *Fest*. The f' and f'' values used in the procedure were known in advance from experimental measurements conducted at the time of data collection. In order to assess the quality of the F_A estimates, *HySS* was used to determine the substructure, while disabling the automatic convergence procedure.

Furthermore, the substructure was solved using only the anomalous differences of a single dataset, as well as using dispersive differences between

Table 1: Fest and HySS results for Acyl-CoA Thioesterase

	SAD
High	0.58
Peak	0.61
Inflection	0.53

Correlation coefficients from HySS obtained using SAD data.

Table 2: 2 wavelength results for Acyl-CoA Thioesterase II

	Δdispersive	F_A
High,Peak	-	0.65
High,Inflection	0.61	0.73
Peak,Inflection	0.50	0.67

Correlation coefficients as obtained from *HySS* using only dispersive differences or estimated F_A values. Note that the dispersive difference between peak and high energy remote did not contain enough dispersive signal to locate the substructure.

pairs of datasets. The correlation coefficient of the best solution is shown in Table 1 for SAD and in table 2, for the 2 wavelength combinations, using only dispersive differences as well as F_A estimates. Although when using SAD, one is able to locate the substructure without any problem, substructure solution with F_A values gave significantly larger correlation coefficients.

Example: SIR

The results of *HySS* using ΔF's obtained from *Fest* using native data and a mercury derivative of GroEl are shown in table 3. Note that the mercury derivative data is incomplete, resulting in a rather incomplete difference data set. Although the completeness is low, *HySS* is readily able to locate the full substructure, and finds 21 out of 21 sites with an *rmsd* of 0.66 Å to the true heavy atom model.

Table 3: GroEl SIR *Fest* and *HySS* details.

Spacegroup	C2221	
Nominal resolution	30-3.8 Å	
	Native	**Derivative**
Completeness	Full: 94% 30-5: 98%	Full: 35% 30-5: 53%
Absolute scaling ln(scale)	-0.67	-0.68
Trace of absolute scaling $\mathbf{B}_{cart}$	51, 47, 66	62, 55, 70
Relative scaling ln(scale)	-0.004	
Trace of relative scaling $\mathbf{B}_{cart}$	-13, -6, 25	
ΔF completeness	34%	
$CC_{consensus}$	46%	
Sites found (correct)	21 (out of 21)	

Discussion

The command line utility *Xtriage* has shown to be a practical tool that can be used quickly after data processing to detect possible twinning, as well as point out incorrect assignment of the space group.

The relative scaling and ΔF estimation techniques implemented in *Fest* seem to be able to handle a variety of data originating from various sources. The F_A estimation for the 2 wavelength MAD case did improve substructure solution as compared to SAD, although the SAD data alone was sufficient to locate the substructure. Currently the value of f' and f'' are required on input for the 2 wavelength MAD case. However, approaches are being developed that do not require manual specification of these parameters.
Fest and *HySS* have shown to be able to deal with RIP and SIR data as well. The full customizability of *Fest* might be of use in unusual cases where the default scaling protocol are suboptimal or when downscaling of a native(-like) data set turns out to be critical.

Acknowledgements

PHZ thanks Tom Terwilliger for stimulating discussion regarding local scaling issues and outlier rejection details as implemented in *SOLVE*. The authors thank Dr. Z. Dauter for making the Acyl-CoA Thioesterase II MAD data set available via the Autostruct (http://www.ccp4.ac.uk/autostruct/) web pages.

References

Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N. K., Storoni, L. C. & Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 53-55.

Dudewicz, E. J. & Mishra, S. N. (1988). *Modern Mathematical Statistics.* New York: Wiley.

Flack, H.D. (1987). Acta Cryst. A43, 564-568.

Garman, E.F. & Schneider, T.R. (1997). *J. App. Cryst.* **30**, 211-237.

Grosse-Kunstleve, R.W., Sauter, N.K. & P.D. Adams. (2004). *IUCr Computing Commission Newsletter* **3**.

Grosse-Kunstleve, R.W., Afonine, P.A., Sauter, N.K. & P.D. Adams. (2005). *IUCr Computing Commission Newsletter* **5**.

Grosse-Kunstleve, R.W. & Adams, P.D. (2003). *Acta Cryst.* **D59**, 1966-1973.

Grosse-Kunstleve, R.W. & Brunger, A.T. (1999). *Acta Cryst.* **D55**, 1568-1577

Kantardjieff, K. & Rupp, B. (2003). *Prot. Sci.* **12**, 1865-1871.

Kingston, R.L. (2001). *Acta Cryst.* **D57**, 101-107.

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A. & Jone, T.A. (2004). *Acta Cryst.* **D60**, 2240-2249.

Koch, E. & Fischer, W. (1983). *International Tables for Crystallography*, Vol. *A*, edited by T. Hahn, ch. 15. Dordrecht: Kluwer Academic Publishers.

Li, J., Derewenda, U., Dauter, Z., Smith, S., Derewenda, Z. S. (2000). *Nat.Struct.Biol.* **7** 555-559.

Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* **A31**, 480-487

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56-59.

Nanao, H.M., Sheldrick, G.M. & Ravelli, R.B.G. (2005). *Acta Cryst.* **D61**, 1227-1237.

Nikonov, S. (1983). *Acta Cryst.* A**39**, 693-697.

Padilla, J.E. & Yeates, T.O. (2003). *Acta Cryst.* **D59***,* 1124-1130.

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* **D59,** 1145-1153.

Read, R.J. (1999). *Acta Cryst.* **D55**, 1759-1764.

Rudolph, M. G., Wingren, C., Crowley, M. P., Chien, Y. H. & Wilson, I. A. (2004). *Acta Cryst.* **D60**, 656-664.

Singh, A.K. & Ramaseshan, S. (1968). *Acta Cryst.* **B24**, 35-39.

Zwart, P.H., Grosse-Kunstleve, R.W. & Adams, P.D. (2005). *CCP4 Newsletter* **42**.

Zwart, P.H. (2005). Acta Cryst **D61**, 1437-1448.