computer programs

XDSAPP: a graphical user interface for the convenient processing of diffraction data using XDS

Michael Krug,[a,b] Manfred S. Weiss,[a] Udo Heinemann[b,c] and Uwe Mueller[a]*

[a]Helmholtz-Zentrum Berlin für Materialien und Energie, Institute F-I2, Macromolecular Crystallography Group (HZB-MX), Albert-Einstein-Strasse 15, D-12489 Berlin, Germany, [b]Max-Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13125 Berlin, Germany, and [c]Institute for Chemistry and Biochemistry, Freie Universität, Takustrasse 6, D-14195 Berlin, Germany. Correspondence e-mail: uwe.mueller@helmholtz-berlin.de

XDSAPP is a Tcl/Tk-based graphical user interface for the easy and convenient processing of diffraction data sets using *XDS*. It provides easy access to all *XDS* functionalities, automates the data processing and generates graphical plots of various data set statistics provided by *XDS*. By incorporating additional software, further information on certain features of the data set, such as radiation decay during data collection or the presence of pseudo-translational symmetry and/or twinning, can be obtained. Intensity files suitable for *CCP4*, *CNS* and *SHELX* are generated.

1. Introduction

Today's third-generation synchrotron sources allow the collection of diffraction images of macromolecular crystals using typical exposure times of less than one second per image. Since most of the corresponding end stations for macromolecular crystallography are equipped with fast detector hardware, the time necessary to collect a complete diffraction data set has decreased from hours a few years ago to minutes. Consequently, the available time to process the collected data in 'real time' during a synchrotron shift has decreased considerably. This puts the experimenter under significant pressure and has resulted in the development of more and more software solutions for automated data processing. Nowadays, it is possible to process the collected data more or less automatically and with little user input, and at the same time to generate some visual feedback about the course of the processing and the resulting data quality for examination by the user. Available software suites for data integration are, for instance, *MOSFLM* and its graphical user interface (GUI)-based application *iMOSFLM* (Leslie, 2006), *HKL-2000* (Otwinowski & Minor, 1997), *XDS* (Kabsch, 1993, 2010a,b), and *D*TREK* (Pflugrath, 1999). The programs differ in their auto-indexing methods (fast Fourier transform or difference vector clustering), their integration methods (two- and three-dimensional profile fitting) and their level of automation. Efforts to automate the integration process vary from GUIs that provide help in editing, for example, the input file of *XDS* and converting the output to different formats as in *XDSi* (Kursula, 2004) to almost complete automation of the various steps from images to a fully processed data set like in *autoPROC* (Vonrhein *et al.*, 2011) or *xia2* (Winter, 2010).

The data processing program *XDS* is a sophisticated piece of software, which was originally developed for processing single-crystal monochromatic diffraction data from crystals recorded by the rotation method on a multiwire area detector (Kabsch, 1993). All the steps required to process a diffraction data set from spatial detector corrections, to indexing, integration and scaling are contained in one executable of the program. Since it does not read the headers of the diffraction images to be processed, all data relevant to the processing have to be supplied to the program in one input file, called `XDS.INP`. The main disadvantage of *XDS* is that it is a text-based software, which means that it is difficult, especially for inexperienced users and those who are not very computer proficient, to run the program and to evaluate the quality of the processed data sets.

In this paper, we present and describe the development and the use of *XDSAPP* (*XDS Automation and Plotting Protocols*), a GUI that combines the software packages *XDS*, *POINTLESS* (Evans, 2006) of the *CCP4* software suite (Collaborative Computational Project, Number 4, 1994), *XDSSTAT* (Diederichs, 2006), *SFCHECK* (Vaguine *et al.*, 1999) and *PHENIX.XTRIAGE* (Adams *et al.*, 2010), thus allowing the crystallographer to index and integrate one or multiple diffraction data sets and to evaluate the quality of the processed data sets with minimum effort and minimum user input. Besides indexing, integration and scaling, the most probable space group is automatically determined and assigned, additional statistics are generated, and tests for potential problems such as pseudo-symmetry and/or twinning are performed. A short summary of the indexing as well as the integration step is displayed for examination by the user. During integration, the most important statistics (*e.g.* scale factors, mosaicity, number of strong spots, beam divergence, unit-cell dimensions) are plotted synchronously. All important statistics of the different processing steps are presented within the GUI and are stored in a pdf file. Plots also include statistics generated by *XDSSTAT*, like the redundancy-independent merging R factor R_{meas} or $R_{r.i.m.}$ (Diederichs & Karplus, 1997; Weiss, 2001; Weiss & Hilgenfeld, 1997; Einspahr & Weiss, 2011) per frame. These can be used to identify images within the data set that are of much poorer quality than the rest of the data set. The decay R factor (R_D) plot (Diederichs, 2006) provides information on the extent of radiation damage within the crystal during the process of data collection.

XDSAPP is currently installed at the macromolecular crystallography beamlines BL14.1-3 at Helmholtz-Zentrum Berlin (HZB) (Mueller *et al.*, 2012).

2. Methods

2.1. Architecture

XDSAPP is a Tcl/Tk-based GUI. The main window consists of eight tabs used for the different tasks of data processing such as data input, indexing, integration, output or plotting (Fig. 1). Initially, the

data set to be processed has to be selected and *IPMOSFLM* (Leslie, 1999) is used to read the image headers in order to create an appropriate `XDS.INP` file. Then, *XDS* is launched to run the tasks INIT, COLSPOT and IDXREF, and a summary of the indexing process is displayed. Next, *XDSAPP* calls *XDS* with the tasks DEFPIX, INTEGRATE and CORRECT, all with space group option '0' for processing of the data in space group $P1$. After postrefinement of the orientation and the cell parameters within the CORRECT step, one reintegration step using the refined parameters (from `GXPARM.XDS`) and the suggested values for BEAM DIVERGENCE, REFLECTING RANGE and their s.u. values (from `INTEGRATE.LP`) is performed by default. Preliminary results are displayed after each integration step. Further reintegration steps can be performed until no further improvements of resolution and/or quality in the previous integration step are detected. After the (re)integration has finished, the program *POINTLESS* (Evans, 2006) is used to determine the most probable space group of the data set and the CORRECT step is rerun in that space group. If the space group and unit-cell parameters of the data set are known beforehand, this step can be omitted by providing *XDSAPP* with these values *via* the 'Settings' tab. A resolution cut-off is applied, and intensity checks for twinning and/or pseudo-translation are done by using *SFCHECK* (Vaguine *et al.*, 1999) and *PHENIX.XTRIAGE* (Adams *et al.*, 2010). Several plots of the quality indicators provided by the CORRECT step are generated, as well as plots of R_{meas} or $R_{r.i.m.}$ per frame, and the decay R factor R_D provided by *XDSSTAT* (Diederichs, 2006). All plots are also stored as a pdf file in the data processing directory. A summary of the processing results is presented for a quick results overview. Using the program *XDSCONV*, intensity files suitable for *CCP4* (Collaborative Computational Project, Number 4, 1994), *CNS* (Brunger, 2007) and *SHELX* (Sheldrick, 2008) are produced.

2.2. Hardware provisions

XDS is the only data processing program known to us that is able to utilize multiprocessor architectures during the course of the data processing. In particular, the most time-consuming steps, such as the search for strong spots for auto-indexing or the integration of the images, are highly parallelized. Within the infrastructure at HZB, we

are using an HP-DL580 G7 40-CPU-core server, which is attached to a large storage area network array *via* several fibre channel uplinks.

2.3. Software environment

XDSAPP runs under Linux and needs the installation of Tcl/Tk version 8.4 or newer. In addition to the above-mentioned crystallographic software, an installation of *gnuplot* (at least version 4.0) and *ImageMagick* (6.2.5 or newer) is needed. In order to have the full functionality of *XDSAPP*, the *XDS* viewer and the old *XDS* viewing program *VIEW* also have to be installed.

2.4. Usage

The main window of *XDSAPP* is separated into eight different tabs representing the different possible tasks (Fig. 1). After having started *XDSAPP*, the user first has to select the detector used for data collection within the 'Input' tab. Here, one can also select the directory in which *XDSAPP* is looking for diffraction images (the default is the directory from which *XDSAPP* has been started). By pushing the appropriate button within the 'SingleDataset' tab, a dialog window (Fig. 1 inset) allowing the selection of a data set opens. The selected data set can then be indexed or processed by clicking the appropriate buttons within the 'SingleDataset' tab. Results of the indexing step are displayed within the tab. During integration, preliminary results are given within the 'SingleDataset' tab. The most important statistics of the integration step are presented as plots synchronously for each integration cycle within the 'Statistics' tab (Fig. 2). After finishing, a summary of the processing results of the data set is given within the 'SingleDataset' tab (Fig. 3), and plots of statistics generated by the CORRECT step and *XDSSTAT* can be viewed as plots in the 'Statistics' tab (Fig. 4).

In the 'MultipleDatasets' tab the user can provide *XDSAPP* with a directory containing multiple data sets. All the data sets within this directory and its subdirectories will then be processed subsequently. In the 'Logfiles and pictures' tab all the log files of the different processing steps and the images generated by *XDS* and *XDSSTAT* can be viewed. In the 'Plotting Utilities' tab plots of the statistics of a processed data set can be generated. This can also be applied to a

Figure 1
Start-up window of *XDSAPP*. The eight tabs are (from left to right) 'Input', 'SingleDataset', 'MultipleDatasets', 'Statistics', 'Plotting Utilities', 'Logfiles and pictures', 'Settings' and 'Running output'.
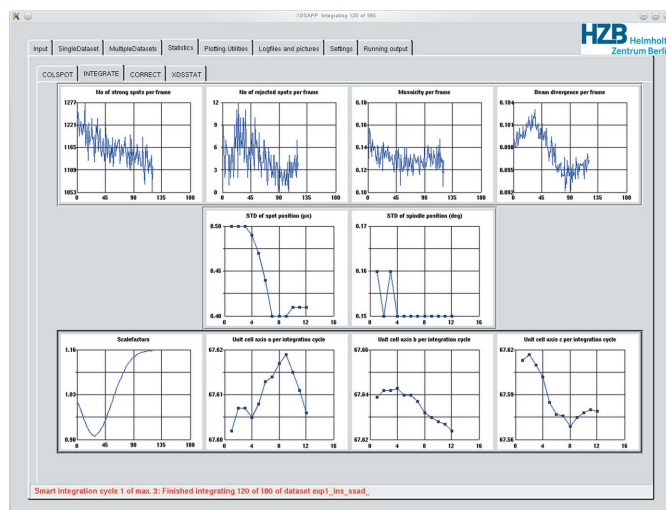

Figure 2
The 'Statistics' tab displays plots of the most relevant statistics of the different *XDS* processing steps (COLSPOT, INTEGRATE, CORRECT) and *XDSSTAT*. The statistics of the INTEGRATE step can be followed concurrently during the integration. This example (case study 1) shows the statistics after 120 out of the 180 frames of the data set have been integrated.

data set that has been processed by *XDS* in the conventional manual way. In the 'Settings' tab the settings of, for example, parameters for *XDS* like the spot range, direct beam position or resolution range can be changed. If the space group and unit-cell parameters of a data set are known beforehand they can also be given here and *XDSAPP* will process the data set in the given space group.

The 'Settings' tab also provides the possibility to define the resolution limit of a data set by setting threshold values for completeness, $I/\sigma(I)$ and R_{meas} (or $R_{r.i.m.}$). *XDSAPP* will then run *XDS* to the maximum resolution possible during the first integration run and use the threshold values to define the resolution cut-off of the data set for the subsequent integration run.

3. Results and discussion

In order to demonstrate the versatility of *XDSAPP*, two data sets collected at BL14.1 at BESSY II were processed with *XDSAPP* using the default settings.

3.1. Case study 1

The first case study is a data set of a cubic insulin crystal collected at a wavelength of 1.77 Å (Faust *et al.*, 2008). The image range(s) for collecting strong spots during the COLSPOT step are set dynamically by *XDSAPP* depending on the overall number of images a data set consists of. In this case, frames 1–12 and 84–96 were used to collect strong spots from. As this data set is of rather high quality, the resulting indexing was very accurate, as shown by a standard deviation of spot position of 0.34 pixels before and 0.29 pixels after refinement. The mean absolute difference between observed and fitted indices was also quite low (DH = 0.03, DK = 0.02, DL = 0.03). Consequently, the first integration step already produced good results so that the second, automatically launched, integration step only led to a minor improvement of data quality. Since it is not possible to distinguish between space group $I23$ and $I2_13$ from diffraction data alone, *XDSAPP* processed the data set in space group $I23$. A comparison with the manual *XDS* processing of the data set is given in Table 1. For the manual processing of this data set the value range for trusted pixels on the detector was adjusted, because otherwise *XDS* did not exclude all of the intruding hardware shadows from the trusted region of the detector. This was done by inspecting the file

BKGPIX.cbf, adjusting the values for the trusted pixel range and repeating the DEFPIX step until all of the intruding hardware shadows were visually removed from the trusted region. Presently, *XDSAPP* is not able to perform such an adjustment automatically. As a consequence, the number of total reflections is smaller for the manually processed data set. Also, the R values as well as the completeness for the highest-resolution shell are lower in the case of the manual processing. Nevertheless, the quality of the two processing runs is very much comparable. As can be seen from Fig. 4(*a*), data collection at this long wavelength led to a clearly detectable anomalous signal of the data set. The anomalous correlation coefficient CCanom is 0.65 for the data set processed by *XDSAPP* and 0.59 for the data set processed manually. *XDSAPP* also ran *SFCHECK* and *PHENIX.XTRIAGE* on the output of *XDS* to test for twinning or pseudo-translation. In this case *SFCHECK* identified a twinning fraction of 5.8%, which is a number too small to be significant.
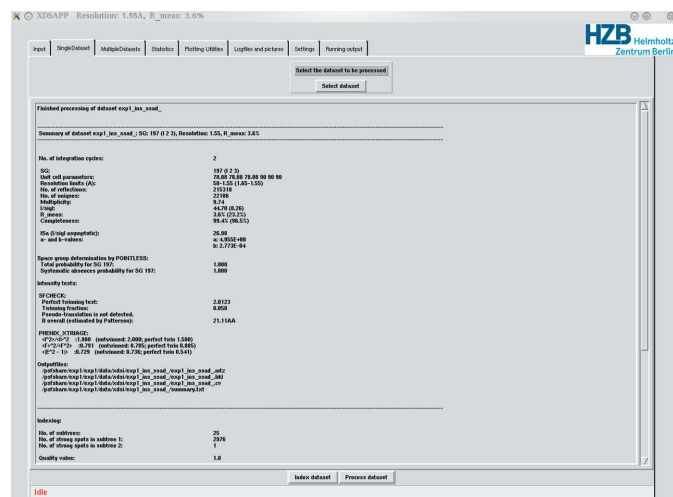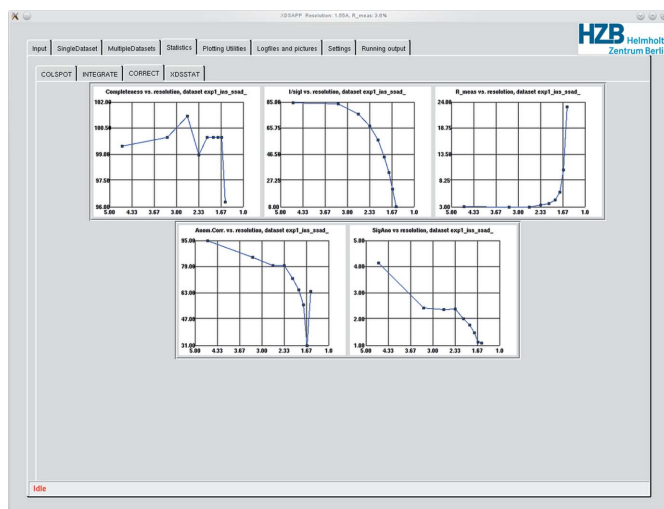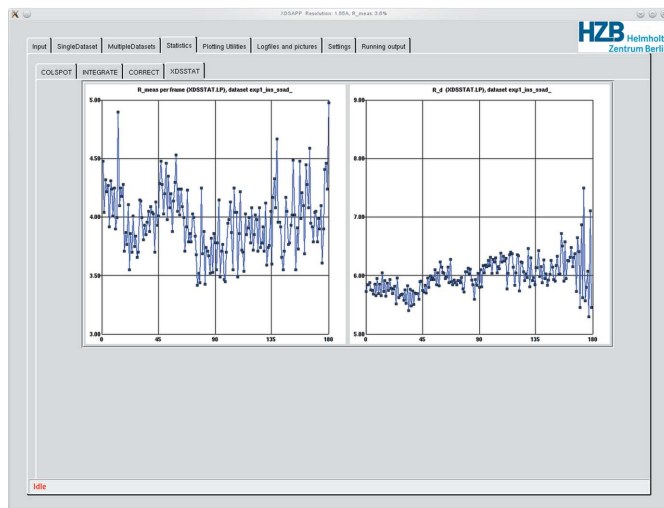

(*a*)

(*b*)

Figure 4
(*a*) Case study 1. Plots of statistics generated by the CORRECT step. The two lower plots of the anomalous correlation and SigAno = $[|F(+) - F(-)|]/\sigma[|F(+) - F(-)|]$ *versus* the resolution show that this data set exhibits a clearly detectable anomalous signal. (*b*) Case study 1. Plots of statistics generated by *XDSSTAT*. R_{meas} (or $R_{r.i.m.}$) per frame (left) helps to identify parts of high or low quality of a data set. A steady increase of R_D with increasing frame number difference is an indicator of radiation damage (right).

Figure 3
Case study 1. A summary of the processing results of the data set is given within the 'SingleDataset' tab.

Table 1
Data collection and processing statistics of case study 1.

Data set	Insulin *XDSAPP*	Insulin manual
Wavelength (Å)	1.77	
Crystal–detector distance (mm)	50	
Rotation range per image (°)	1.0	
Total rotation range (°)	180	
Exposure time per image (s)	2.4	
Space group	*I*23 or *I*213	
Resolution range (outer range) (Å)	50.0–1.55 (1.65–1.55)	50.0–1.55 (1.65–1.55)
Unit-cell parameters	78.08, 78.08, 78.08,	78.07, 78.07, 78.07,
$a, b, c, \alpha, \beta, \gamma$ (Å, °)	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Mosaicity (°)	0.098	0.108
Total No. of reflections	215 318	179 837
Unique reflections	22 106	22 057
Multiplicity	9.74	8.15
$\langle I/\sigma(I)\rangle$	44.7 (8.26)	42.2 (9.1)
ISa	26.98	27.36
Completeness (%)	99.4 (96.5)	99.1 (94.6)
R_{merge}† (%)	3.9 (24.1)	3.2 (14.9)
$R_{meas}/R_{r.i.m.}$† (%)	4.0 (25.6)	3.4 (16.8)
$R_{p.i.m.}$† (%)	0.9 (8.2)	1.0 (5.5)
Overall *B* factor from Wilson plot (Å²)	22.13	21.52

† $R_{merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the integrated intensity of a given reflection and $\langle I(hkl)\rangle$ is the mean intensity of multiple corresponding symmetry-related reflections. $R_{meas} = R_{r.i.m.} = \sum_{hkl}[N/(N-1)]^{1/2}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$, where N is the multiplicity of a given reflection. $R_{p.i.m.} = \sum_{hkl}[1/(N-1)]^{1/2}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$. ISa $= I/\sigma(I)_{asymptotic}$ represents the highest possible signal-to-noise ratio of a data set (Diederichs, 2010). Data sets with ISa values of 25 or greater are considered to be very good and should allow for straightforward structure determination. The calculation of ISa has been implemented in *XDS* since December 2009.

Table 2
Comparison of overall statistics after the first and the second integration step with refined parameters for the Tt-IPMDH data set (case study 2).

No resolution cut-off was applied at that stage of processing.

	Integration 1	Integration 2
$\langle I/\sigma(I)\rangle$	6.12	9.93
Completeness (%)	88.1	92.7
R_{meas} or $R_{r.i.m.}$ (%)	20.4	13.7
ISa	13.89	26.12

XDSAPP displays the overall numbers of the twinning tests by *SFCHECK* and *PHENIX.XTRIAGE* and stores the complete output of these programs in log files within the processing directory.

3.2. Case study 2

The second case study is a data set of a Tt-IPMDH (isopropyl-malate dehydrogenase from *Thermus thermophilus*) crystal. This data set consists of 720 diffraction images. In this case, frames 1–20 and 350–370 were used to collect strong spots from. Owing to the comparatively poor quality of the diffraction images, the indexing turned out to be less accurate than in case study 1, which is reflected in a rather large standard deviation of the spot position of 5.66 pixels before and 1.22 pixels after refinement and quite big mean absolute differences between observed and fitted indices (DH = 0.03, DK = 0.09, DL = 0.08). Consequently, the results of the first integration step were also unsatisfactory. Nevertheless, it turned out to be possible to improve the integration significantly by the second, automatically launched, integration step using the post-refined parameters (Table 2). For example, the value of ISa (Diederichs, 2010) almost doubled from integration step 1 to integration step 2. As can be seen from the numbers presented in Table 3, *XDSAPP* processed the data set to slightly lower resolution than in the manual processing. This is a consequence of the default parameters for resolution cut-off in the

Table 3
Data collection and processing statistics of case study 2.

Data set	Tt-IPMDH *XDSAPP*	Tt-IPMDH Manual
Wavelength (Å)	0.91841	
Crystal–detector distance (mm)	301.42	
Rotation range per image (°)	0.25	
Total rotation range (°)	150	
Exposure time per image (s)	4.0	
Space group	*C*222₁ or *C*222	
Resolution range (outer range) (Å)	50.0–2.94 (3.12–2.94)	50.0–2.8 (2.97–2.80)
Unit-cell parameters	51.33 153.69 176.48,	51.31 153.68 176.41,
$a, b, c, \alpha, \beta, \gamma$ (Å, °)	90, 90, 90	90, 90, 90
Mosaicity (°)	0.166	0.145
Total No. of reflections	92 765	128 591
Unique reflections	15 175	17 657
Multiplicity	6.11	7.28
$\langle I/\sigma(I)\rangle$	20.2 (3.2)	19.4 (2.0)
ISa	27.86	28.71
Completeness (%)	99.3 (98.3)	99.6 (98.5)
R_{merge}† (%)	6.6 (51.3)	7.6 (90.2)
$R_{meas}/R_{r.i.m.}$† (%)	7.2 (52.7)	8.2 (97.0)
$R_{p.i.m.}$† (%)	2.9 (21.8)	3.0 (36.1)
Overall *B* factor from Wilson plot (Å²)	70.55	74.58

† See Table 1 for definitions of the used metrics.

'Settings' tab. For this data set, no twinning or pseudo-translation was identified by *SFCHECK* and *PHENIX.XTRIAGE*.

The plot of $R_{meas}/R_{r.i.m.}$ per frame of Fig. 4(*b*) shows that the quality of this data set is almost constant over the whole data range, *i.e.* there are no parts of noticeably bad quality. The plot of R_D as a function of frame difference in Fig. 4(*b*) shows almost no increase with increasing frame difference. This indicates that there are no significant signs of radiation damage present within this data set.

4. Conclusion

XDSAPP is a GUI that significantly simplifies diffraction data processing with *XDS*. It can give novice users who are deterred by the command line input to *XDS* comprehensible access to this software. The visualization of tables of statistics as plots within the GUI simplifies the validation of the results.

The two case studies showed that *XDSAPP* is able to process diffraction data sets as well or almost as well as conventional manual processing. In particular, case study 2 showed that even a data set that could not be accurately indexed can be processed reliably. In this case, the automatically launched second integration step with post-refined parameters leads to a significant improvement of the results without any manual user interference.

The plots visualizing information generated by *XDSSTAT* can help to facilitate decision making of manual fine-tuning steps to further improve data quality.

The summarized results of the intensity evaluations carried out by *SFCHECK* and *PHENIX.XTRIAGE* can give a quick warning of special crystallographic cases and can help to identify, for example, the one untwinned crystal amongst many.

XDSAPP will be made available, free of charge for academic users, upon request and after registration from the web page http://www.helmholtz-berlin.de/bessy-mx.

computer programs

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Brunger, A. T. (2007). *Nat. Protoc.* **2**, 2728–2733.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Diederichs, K. (2006). *Acta Cryst.* D**62**, 96–101.
Diederichs, K. (2010). *Acta Cryst.* D**66**, 733–740.
Diederichs, K. & Karplus, P. A. (1997). *Nat. Struct. Biol.* **4**, 269–275.
Einspahr, H. M. & Weiss, M. S. (2012). *International Tables for Crystallography*, Vol. F, *Crystallography of Biological Macromolecules*, 2nd ed., ch. 2.2, pp. 64–74. Heidelberg: Springer.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Faust, A., Panjikar, S., Mueller, U., Parthasarathy, V., Schmidt, A., Lamzin, V. S. & Weiss, M. S. (2008). *J. Appl. Cryst.* **41**, 1161–1172.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Kabsch, W. (2010*a*). *Acta Cryst.* D**66**, 125–132.
Kabsch, W. (2010*b*). *Acta Cryst.* D**66**, 133–144.
Kursula, P. (2004). *J. Appl. Cryst.* **37**, 347–348.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696–1702.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.
Mueller, U., Darowski, N., Fuchs, M. R., Förster, R., Hellmig, M., Paithankar, K. S., Pühringer, S., Steffien, M., Zocher, G. & Weiss, M. S. (2012). *J. Synchrotron Rad.* **19**, 442–449.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Weiss, M. S. & Hilgenfeld, R. (1997). *J. Appl. Cryst.* **30**, 203–205.
Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.