Proceedings of the 9th National Symposium of Synchrotron Radiation Users, Warsaw, September 26–27, 2011

Data Processing Programs for Analysis of Diffraction Images of Macromolecular Crystals Recorded using Synchrotron Radiation

M. Gilski*

Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University

Grunwaldzka 6, Poznań 60-780, Poland

and Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, PAS

Noskowskiego 12/14, Poznań 61-704, Poland

Besides the crystal preparation, the first and crucial step in the process of protein structure determination is proper processing of the collected diffraction images, as they provide the experimental observations used throughout the entire process of structure solution and refinement. In the last two decades several computer programs have been developed. Among the most used and popular are: HKL2000, MOSFLM, d*TREK and XDS package. To find out the advantages and disadvantages of the data processing programs, several very different data sets, including diffraction data from DNA/RNA and protein crystals were tested. It has been found that all the major programs for processing and analysis of diffraction data give excellent and comparable results with good quality, medium resolution data sets, but their treatment of very high resolution or imperfect data differs in terms of indexing, spot integration, scaling and the treatment of errors. If the diffraction data are of good quality and the problem is relatively straightforward, the automated approach to data processing may be most appropriate. On the other hand, if one is trying to squeeze out as much information from the experimental data as possible, then only expert manual processing can be successful, regardless of the data quality.

PACS: 61.10.–i, 07.85.Qe, 61.05.cp

1. Introduction

X-ray crystallography is a dominant technique used in determination of the three-dimensional structures of macromolecules. It is the most successful when applied on the third-generation synchrotron sources that allow rapid collection of X-ray data from macromolecular single crystals.

At present there are over 100 synchrotron radiation facilities all over the world and most of them have, usually more than one, dedicated line suitable for X-ray macromolecular diffraction experiments. Recently often they provide full automatic and remote access mode [1–3]. In rough estimate, synchrotron stations are capable of producing more than 500,000 data sets per year. Comparing that with the number of structures deposited every year in PDB database ($\sim$ 5000 last ten years average, $\sim$ 8000 last year) we can estimate that about 60 data sets are needed per one successful PDB deposit [4–6].

There are many reasons of such situation, including but not limited to, poor or unsatisfactory data quality, problems with automatic/semi-automatic processing and indexing of raw diffraction images or difficulty with structure solving, model building and/or refinement.

The quality of the diffraction data is determined by set of factors interrelated with instrumentation, experiment parameters and many biochemical and physical features of the crystal. Some of them are connected with quality of the crystal itself: sample dimensions, twinning, mosaicity and macromolecules internal disorder, others e.g. completeness, partiality, detector pixel saturation depend on the experiment settings and some like spot separation, recorded resolution, radiation damage are correlated with all mentioned above factors. The certain crystal 'imperfections' (some types of twinning) or undesirable effects (radiation damage) can be corrected or reduced by choosing optimal data acquisition strategy and/or by appropriate data processing procedure but some of them like crystal size and poor diffraction are suggesting changing the crystal sample.

When crystal, especially small one and with large unit cell, is exposed on strong X-ray radiation, it could readily show progressive radiation damage, despite cryoprotection. This problem become specially acute with the third-generation synchrotron X-ray sources which are capable to put in short time a huge amount of X-ray photons into the crystal. During the experiment, depending on the crystal atom composition the decay of the recorded Bragg intensities is observed and accumulated radiation damage decreases the signal-to-noise ratio of collected diffraction images, induces specific chemical modifications in the macromolecules and changes of the unit cell volume, crystal mosacity and increases the R(free)-value during structure refinement. Degradation of sample's scattering power can be substantially reduced by choosing a proper strategy using specialized software BEST [7], RADDOSE [8] or dedicated module of data processing packages.

* e-mail: mirek@amu.edu.pl

Twinning, other phenomenon influencing diffraction quality is one of the most common crystal defects in macromolecular crystallography. There is no general algorithm to index a diffraction image from multiple crystals. According to the twinning type usually it can be recognized during inspection of diffraction patterns (non-merohedral twinning) or when the X-ray diffraction data are analyzed using intensity statistics (merohedral/pseudo-merohedral twinning). Except the crystal diffraction quality, there are many factors and parameters which must be considered both during data collection and processing data. It is always recommended to spend some time before starting recording the full data set to determine optimal parameters for the data collection. Experimenters should try to avoid an approach (used currently very often) termed as the 'American method' - shoot first and ask questions later [9] and don't use a routine procedure for data collection. The information from the first few images allows to assess the crystal symmetry, potential twinning and resolution and get some directions to find out a proper mode of full data acquisition. Except a fully automated approach, careful visual inspection of the initially exposed images should be the primary means of ensuring quality and setting the strategy of the experiment. The first and easiest features to check are the reflection profiles, spot separation and resolution of the diffraction patterns. Reflection profiles should be regular with a single peak and spots and lunes (rings of spots from one reciprocal plane) should be well resolved [10, 11]. If they are overlapping (even after setting smaller oscillation range or increasing crystal to detector distance) usually there is no point in collecting such data. It may be worth considering the use of one of a standalone strategy simulation and data collection experts systems like EDNA [12] or choose e.g. a multi-dataset data-collection strategy [13] which produces slightly better and more accurate data by acquiring diffraction data in multiple passes keeping fixed radiation dose. Unfortunately the latest method is rather suitable for the low resolution anomalous data collected with home-laboratory X-ray sources.

Processing macromolecular diffraction data in modern crystallography is a set of well-defined and validated procedures. It consists of several steps like indexing of the diffraction pattern, refinement of the crystal and detector parameters, integration, scaling and statistical analysis of the measurements. Each of this steps can introduce some errors, dependent on data quality and also on algorithm used during the particular task.

After evaluation and refinement of sample's unit cell and crystal orientation, the intensities for the Bragg spots can be determined. Algorithmically it is a very complex task but the most important during whole diffraction data processing. The ultimate result of the processing is a list of reflections which appear on images with their Miller indices (*hkl*), estimated intensities, and standard deviations.

The aim of this work is to give a short description of the diffraction data processing programs supported by conclusions and problems occurred during processing a couple of macromolecular datasets by different programs. Each of the selected software packages were developed for more than twenty years ago and they are very well tested and for the most typical cases produces similar results and the differences can be observed only during processing much more difficult datasets with extreme high resolution, the twinning or other various crystal and processing pathologies.

2. Diffraction data processing software

To analyze single-crystal diffraction data, several computer programs have been developed. To the group of the most popular and used programs belongs: HKL2000 [14], MOSFLM [15], d*TREK [16] and XDS package [17].

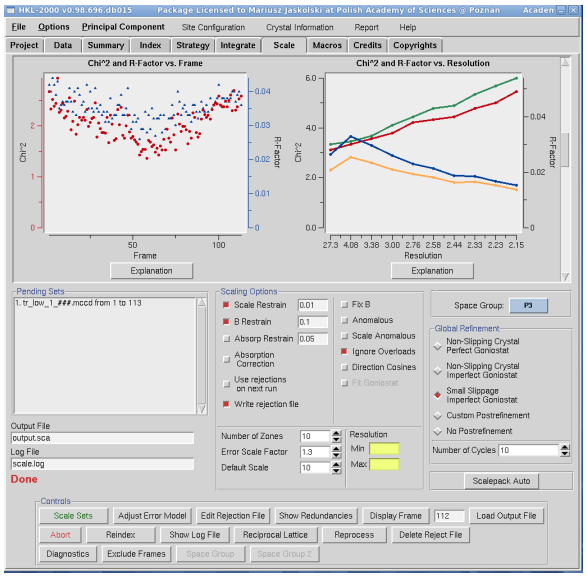

Fig. 1. View of graphical user interface of HKL2000 showing the Scale tab. Visible plots illustrating statistical analysis of different processing parameters.

HKL2000 is the most popular diffraction data processing program package based on the extended versions of Denzo, Xdisplayf and Scalepack. Graphical user interface is well designed (Fig. 1) and consists of several tabs to start consecutive tasks like indexing, parameter refinement, strategy, integration and scaling. It improves very much the convenience of working with various data sets and provides very good and reasonable set of default input parameters. The three-dimensional fast Fourier transform (FFT) autoindexing routine implemented in Denzo is very powerful and efficient [14]. Absorption correction, using spherical harmonics, dramatically improves anomalous signal. Due to its new processing strategy, HKL2000 can handle data from crystals with high mosaicity.

The diffraction data-integration program MOSFLM with its graphical user interface iMOSFLM [18] was de-

signed to provide an intuitive path to data processing even for inexperienced and advanced users, though the full functionality is available from the command-line. MOSFLM includes the one-dimensional FFT autoindexing based on Rossmann's DPS algorithm [19] and integrate data using two-dimensional profile fitting. It can accumulate spot profiles over several, adjacent images and its graphical user interface (GUI) enable easy definition of shaded areas.

XDS package includes a set of three programs: XDS, XSCALE and XDSCONV. The main program XDS consist of eight major routines which are called in succession exchanging information between the steps by files. It does not have a dedicated graphical user interface but visual feedback is available through external diffraction image viewers [20]. Provided as part of the documentation detector specific input file templates greatly simplify the use of the package. XDS can use a whole dataset for indexing and constructs 3D profiles already in the indexing step. Implemented OpenMP technology enable execution on multi-processor clusters (up to 32 CPU's).

The next processing program d*TREK is a flexible, customizable, device-independent software suite for the visualization and processing of single crystal diffraction images. It consists of modules for all data processing steps accessed through a graphical user interface developed with X Window and OSF/Motif toolkit. It has three-dimensional Fourier autoindexing routine and uses full 3D profile integration. Simple intuitive and well organized graphical user interface is very helpful both for beginners and advanced users. d*TREK package has implemented method to evaluate the quality of crystals and diffraction images and can assign a rank per sample.

All software packages are based on similar algorithms but diverse implementation leads to different efficiency, performance and accuracy of the particular tasks.

A very important issue is the way of processing partially recorded reflections - reflections recorded on two or more consecutive images [21]. In order to reduce the amount of background recorded on the image the modern data collection protocols use relatively small oscillation angle (fine-sliced phi method) where most spots on the image are recorded as partials [16, 22]. This is especially relevant when diffraction data are recorded using a new generation synchrotron radiation sources with high speed detectors and shutters or shutterless systems. Only XDS and d*TREK have routines for full three-dimensional profile analysis where 3D profile is used to evaluate the total intensity. The HKL2000 and MOSFLM packages use a 2D method where partially recorded reflections are evaluated independently by two-dimensional profile fitting and only summed to give the total intensity.

Except one all programs are controlled through a dedicated GUI. Only the XDS package intentionally does not have its own GUI and most often is used as a command line application. It seems that the latest feature of XDS is one of the strong sides of this package - the clear and well defined as well as organized input file with very good default parameters gives the user a full control over the program. On the other side, MOSFLM has the most expanded GUI but generally all interfaces are to similar extend intuitive and provide reasonable starting values for the essential parameters of the data processing.

3. Results

To find out the advantages and disadvantages of data processing programs, four very different data sets (Table), including diffraction data from DNA/RNA (Fig. 2) and protein crystals were tested.

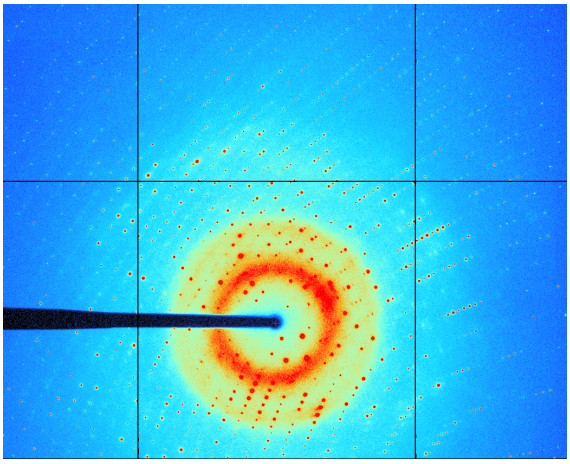

Fig. 2. A macromolecular diffraction pattern for a strongly diffracting crystal of Z-DNA collected at Advanced Photon Source (APS), Argonne – beamline 24-ID-C, detector ADSC Q315. The maximum resolution at the edge of the detector is 0.53Å..

Three of the data sets Z-DNA [23], BPTI [24] and NRAD [25] derive from macromolecular crystals associated with structures which have been previously solved and refined. Originally, diffraction data of Z-DNA and BPTI which demonstrated extremely high resolution 0.55Å and 0.75Å respectively, were processed by HKL2000. Re-processing using the XDS package gave, in both cases, higher resolution (0.53Å and 0.74Å), similar statistics and significantly larger number of reflections. The MOSFLM and d*TREK gave slightly worse results, rather similar to HKL2000. Visual inspection of the diffraction images of the NRAD data set indicated an evident non-merohedral twinning of the crystal employed during data collection (Fig. 3). Although all software packages are able to process diffraction images of twinned crystals [11] after several trials of data reprocessing, it seems that the manual separation of not indexed and indexed spots in XDS is fastest, easiest and the most efficient way to treat this type of very common twinning. Using this method, after processing the diffraction data of the NRAD crystal, two sets of reflections from two different lattices existing in the crystal were obtained. Lattice parameters were different (about 5%) and both

data sets have completeness above 96% and good statistics. It means that in one collected data set complete diffraction data deriving from two different structures of two forms of protein are included.

TABLE

Result of processing by different software packages of selected four diffraction datasets from DNA and proteins.

		HKL2000	MOSFLM	D*TREK	XDS	
Z-DNA[1]	Max. resolution (Å)	0.55	0.54	0.55	0.53	
	$I/\sigma(I)$	2.7	2.7	2.3	2.2	
	R_{merge} (%)	5.7	5.0	5.2	4.4	
	No. reflections	78206	78273	77845	79935	
	Completeness (%)	96.6	97.1	97.6	90.7	
BPTI[2]	Max. resolution (Å)	0.75	0.74	0.75	0.74	
	$I/\sigma(I)$	2.4	2.3	2.9	2.5	
	R_{merge} (%)	6.1 75275	8.9	7.7	6.4	
	No. reflections	75275	73964	74482	77447	
	Completeness (%)	99.9	95.5	98.1	100	
NRAD[3]	Max. resolution (Å)	1.58	1.57	[5]	1.45[6]	1.48[7]
	$I/\sigma(I)$	2.4	3.9		2.3	2.2
	R_{merge} (%)	8.6	10.6		7.5	11.5
	No. reflections	27745	28269		34225	33132
	Completeness (%)	98.9	99.6		95.8	97.0
MPMV[4]	Max. resolution (Å)	1.65	1.66	1.69	1.63	
	$I/\sigma(I)$	2.9	2.4	1.9	1.9	
	R_{merge} (%)	5.1	6.2	7.6	6.8	
	No. reflections	18643	18037	17371	21583	
	Completeness (%)	89.8	91.2	90.1	99.0	

[1]Z-DNA hexamer duplex d(CGCGCG) [23], [2] Mutant of bovine pancreatic trypsin inhibitor [24], [3] DNA repair and recombination protein [25], [4] Mason-Pfizer Monkey Virus protease [26], [27], [5] not processed due to the license problem, [6,7] Two forms of NRAD extracted from one dataset.

Processing the different X-ray diffraction data sets reveals that XDS package with manually edited input file result in very good integrated intensities with the highest resolution and good statistical parameters. The XDS and MOSFLM are much more sensitive than other to the precise values of the direct-beam position (the x and y convention sometimes is swapped between different programs and detectors). The HKL2000 and XDS have somewhat more powerful autoindexing procedure and with default input parameters, they seem to give better merging statistics. The strength of XDS lies in its ability to process data using all resources of the computer. The parallel XDS version (xds_par) uses OpenMP for execution by a team of up to 32 threads and relies on a shared memory multiprocessor platform. All packages can be run from a script, which makes them more suited for automation.

4. Conclusions

It has been found that all the major programs for processing and analysis of diffraction data give excellent and comparable results with good quality, mid-range resolution data sets, but their treatment of very high resolution or imperfect data differs in terms of indexing, spot integration, scaling and the treatment of errors. The easy cases can be processed with any program, using default parameters, but for the difficult ones the best resolution and statistics can be achieved by experienced user, manual setting and monitoring processing parameters. All program have reasonable initial input parameters, however the XDS defaults seems to be most universal end effective. Many of the parameters have assigned default values that work fine in most cases and rarely need to be changed. The task of running XDS can be limited to just editing a few parameter values in the selected input file template appropriate for detector type used during data collection and renaming the edited file into XDS.INP.

If diffraction data are of good quality and the problem is relatively straightforward the automation approach to data processing may be appropriate. On the other hand, if one is trying to squeeze out as much information from the experimental data as possible only the expert man-

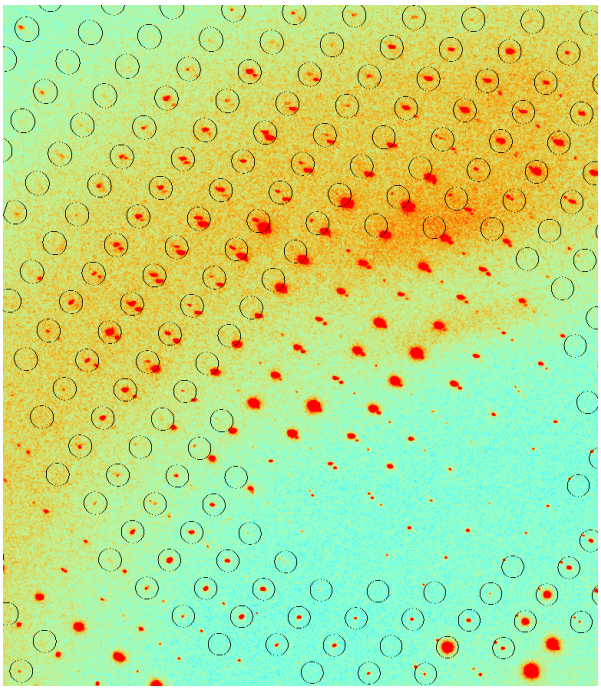

Fig. 3. Fragment of the diffraction image of NRAD with predicted reflections positions superimposed (circles). Not all reflections are predicted, visible spot splitting, associated with the non-merohedral twinning.

ual processing can be successful, regardless of the data quality.

References

[1] M. Gilski, *Acta Phys. Pol. A* **114**, 331 (2008).

[2] M. Jaskolski, M. Gilski, *Academia* **2**, 8 (2007).

[3] M. Gilski, *Synchrotron Rad. in Nat. Science*, Vol. 6, No. 1–2 (2007).

[4] M. Grabowski, M. Chruszcz, M.D. Zimmerman, O. Kirillova, W. Minor, *Infect. Disord. Drug Targets* **9**, 459 (2009).

[5] Z. Dauter, M. Jaskolski, A. Wlodawer, *J. Synchrotron Rad.* **17**, 433 (2010).

[6] M. Cymborowski, M. Klimecka, M. Chruszcz, M.D. Zimmerman, I.A. Shumilin, D. Borek, K. Lazarski, A. Joachimiak, Z. Otwinowski, W. Anderson, W. Minor, *J. Struct. Funct. Genomics* **11**, 211 (2010).

[7] G.P. Bourenkov, A.N. Popov, *Acta Cryst. D* **62**, 58 (2006).

[8] K.S. Paithankar, R.L. Owen, E.F. Garman, *J. Synch. Rad.* **16**, 152 (2009).

[9] M. G. Rossmann, J. W. Erickson, *J. Appl. Crystallogr.* **16**, 629 (1983).

[10] Z. Dauter, *Acta Cryst. D* **55**, 1703 (1999).

[11] C. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, G. Bricogne, *Acta Cryst. D* **67**, 293 (2011).

[12] M.-F. Incardona, G.P. Bourenkov, K. Levik, R.A. Pieritz, A.N. Popov, O. Svensson, *J. Synchrotron Rad.* **16**, 872 (2009).

[13] Z.-J. Liu, L. Chen, D. Wu, W. Ding, H. Zhang, W. Zhou, Z.-Q. Fu, B.-C. Wang, *Acta Cryst. A* **67**, 544 (2011).

[14] Z. Otwinowski, W. Minor, *Methods Enzymol.* **276**, 307 (1997).

[15] A.G.W. Leslie, *Jnt CCP4/ESF-EACBM Newsl. Protein Crystallogr.* **26** (1992).

[16] J. W. Pflugrath, *Acta Cryst. D* **55**, 1718 (1999).

[17] W. Kabsch, *Acta Cryst. D* **66**, 125 (2010).

[18] T.G.G. Battye, L. Kontogiannis, O. Johnson, H. R.Powell, A.G.W. Leslie, *Acta Cryst. D* **67**, 271 (2011).

[19] I. Steller, R. Bolotovsky, M.G. Rossmann,*J. Appl. Cryst.* **30**, 1036 (1997).

[20] A. Arvai, ADXV – a program to display X-ray diffraction images, http://www.scripps.edu/~arvai/adxv.html (2009) .

[21] W. Minor, D. R. Tomchick, Z. Otwinowski, *Structure* **8**, R105 (2000).

[22] Z. Dauter, *Acta Cryst. D* **66**, 389 (2010).

[23] K.Brzezinski, A. Brzuszkiewicz, M. Dauter, M. Kubicki, M. Jaskolski, Z. Dauter, *Nucleic Acids Res.* **39**, 6238 (2011).

[24] R. Thaimattam, E. Tykarska, A. Bierzynski, G.M. Sheldrick, M. Jaskolski, *Acta Cryst. D* **58**, 1448 (2002).

[25] A. Wlodawer, private communication.

[26] M. Gilski, M. Kazmierczyk, S. Krzywda, H. Zábranská, S. Cooper, Z. Popovic, F. Khatib, F. DiMaio, J. Thompson, D. Baker, I. Pichová, M. Jaskolski, *Acta Cryst. D* **67**, 907 (2011).

[27] F.Khatib, F. DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski, D. Baker, *Nature Struct. Mol. Biol.* **18**, 1175 (2011).