Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Data collection and processing

J. P. Turkenburg[a]* and K.E. McAuley[b]*

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, UK, and [b]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot OX11 0DE, UK

Correspondence e-mail: johan.turkenburg@york.ac.uk, katherine.mcauley@diamond.ac.uk

In 2012 the title for the annual CCP4 Study Weekend was Data Collection and Processing, following the tradition of revisiting major subjects in macromolecular crystallography. The conference was held at the University of Warwick and attended by more than 400 scientists, mainly from the UK but also a significant number from further afield.

Data collection is the last experiment in an often arduous pursuit of determining the three-dimensional structure of a macromolecule. A very large amount of time, effort and hard-to-get funding has generally been spent on obtaining sufficient quantities of purified material, and crystals of sufficient quality. Exposing the crystal to X-rays, either using a home source or a synchrotron, and obtaining the best possible data should therefore not be taken lightly. Surprisingly, this final experiment is often carried out hastily and without due attention, simply taking an approach that has become habitual. The disappointment felt when a data set turns out to be incomplete, or of insufficient quality to solve or refine the structure, is enormous and is an excellent reason to study the proceedings in this issue with the attention they deserve.

In recent years there have been several major developments in data collection and processing. Third-generation synchrotron sources are now widely available, sample handling has been automated at most facilities, and pixel array detectors have made it possible to perform shutterless data collections, reducing errors and allowing data sets to be collected in mere minutes. At the same time data-processing software has been streamlined, and expert software packages have become available that in many cases are able to analyse the raw images and provide the user with data reduced in the correct space group. It is paramount that these developments are used properly, so as to optimize the use of large scale facilities such as synchrotrons and more recently free electron lasers. At the same time it gives the user more time to carefully consider how to utilize all these tools in order to obtain the best data.

The meeting was opened with introductory talks by Zbyszek Dauter and James Holton. Zbyszek covered the basics of data collection, including the consideration of symmetry, unit-cell alignment in the beam, crystal lifetime, how to avoid overlaps and overloads and many other fundamental concepts that need to become second nature. He stressed that how to collect data depends on what the data are needed for, *e.g.* native data or data for molecular replacement or phasing calculations. James re-iterated that the ratio of data sets collected to structures deposited remains surprisingly large and went on to warn about overlaps and radiation damage. The last part of his talk concentrated on 'signal to noise', and what can be done to optimize this.

After lunch there was a series of talks about data processing and data quality. Andrew Leslie covered the theory and practice of autoindexing, explaining the underlying principles and highlighting what parameter choices are important for a successful outcome. Jim Pflugrath described the program *d*TREK*. Phil Evans and Kay Diederichs then discussed aspects of data quality. Both highlighted the use of $CC_{1/2}$ as a far better measure of data quality, and a criterion for determining the resolution limit of the data. The acceptance of such better measures of data quality than the $R_{\rm merge}$ is not yet universal, but several recent publications highlighting their usefulness should improve this situation.

The tea break was followed by four talks on more technical aspects of data collection. Jan-Pieter Abrahams spoke about progress in the collection and processing of electron diffraction data. Thomas White described the sample handling when using the free electron laser as a radiation source where a suspension of crystals is injected into the beam path, ensuring a continuous supply of fresh crystals. A very large number of data images are collected, and the approach taken to processing these was described in some detail. Andrew McCarthy detailed the use of a mini-kappa goniometer head and the associated software as implemented at the ESRF. Using several examples he showed how

data quality can be improved by having the ability to optimize the orientation of the crystal. The most obvious applications are for crystals with (one) large unit-cell dimension, and optimizing anomalous signals by collecting Bijvoet pairs in close temporal proximity. Gwyndaf Evans concluded the first day with a talk on the difficulties of dealing with very small crystals and the advantages offered by microfocus beamlines. Visualization of such crystals using microtomography and microradiography is especially useful for membrane protein crystals grown in lipid cubic phase, where the crystals are optically invisible. The relative merits of these two approaches were discussed, as well as the potential for routine application on beamlines. The conference dinner afforded an opportunity to freely discuss the topics raised, and to relax in preparation for the next day.

The second day started with a talk by Alexander Popov on features and development of *BEST*, a program which aids the determination of optimal data collection strategies. Symmetry, orientation of the crystal and assumptions about the lifetime of crystals are all taken into consideration to suggest how best to collect data on a particular crystal. Graeme Winter followed this with a talk on automating the outcome of the diffraction experiment, *i.e.* the characterization and processing of a set of diffraction images. His suite *xia*2 employs existing programs for indexing, integration, scaling and merging of data to present the user with a data set ready for structural analysis. Nicholas Sauter presented methods for handling the large amount of data collected at very high rates with modern detectors, focusing on how he has used the Python language to create a suite for Bragg peak data processing.

After coffee the focus was on more technological aspects. Tadeusz Skarzynski gave on overview of the evolution of X-ray sources, optics and detectors for the home laboratory. For many scientists the use of such equipment has shifted away from data collection towards screening and characterization of crystals, either *in situ* or mounted. This has implications for the desired properties of in-house X-ray diffraction systems. Elspeth Gordon highlighted what industrial users expect from a synchrotron source, specifically the ESRF, and how these demands have led to data collection approaches and developments in sample handling automation and tracking that now also offer enormous benefits to academic users who want to work more efficiently. The morning session was concluded by Armin Wagner with a presentation on the use of laser tweezers to manipulate microcrystals. He demonstrated the successful mounting of such crystals using optical traps.

The final session started with a talk by Marcus Muller about single photon counting pixel detectors and their optimal use in data collection. The availability of these devices requires a new approach to diffraction experiments due to their near-instant read-out and essentially noiseless images. Recent discussions on the CCP4BB emphasize the need for careful consideration of how to make the best use these exciting detectors. The Structural Genomics Consortium in Oxford has considerable experience in dealing with many projects and associated data collection experiments and data sets. Tobias Krojer gave a lucid account of what works and doesn't work in terms of crystal handling, data collection and processing, providing many invaluable tips and warnings for pitfalls. David Stuart presented the results of *in situ* data collection experiments on virus crystals where the whole crystallization tray is mounted in the beam, scanned for crystals, and partial data sets are collected from each crystal. This circumvents the problems associated with crystal mounting and cryoprotection. The meeting was concluded by Wayne Hendrickson who showed how data that are very carefully collected from several crystals, can be combined to give SAD phasing from light atoms naturally present in proteins. Using long-wavelength beamlines that are becoming available, this approach will undoubtedly become more generally applicable.

This issue contains a collection of research papers based on the presentations at the CCP4 Study Weekend 2012. Not all speakers saw the need to publish their talks in written form. The excellent introductions given by Zbigniew Dauter and James Holton are already described extensively in their earlier papers (for example Dauter, 1999, 2010; Holton & Frankel, 2010; Holton *et al.*, 2012). Readers are strongly encouraged to familiarize themselves with the wealth of information in those seminal publications.

We would like to thank Shirley Miller and her excellent team, the CCP4 staff and Stuart Eyres for their organization of the practical aspects of the Study Weekend. We also thank all the speakers and the contributors to this issue.

References

Dauter, Z. (1999). *Acta Cryst.* D**55**, 1703–1717.
Dauter, Z. (2010). *Acta Cryst.* D**60**, 389–392.
Holton, J. M. & Frankel K. A. (2010). *Acta Cryst.* D**60**, 393–408.
Holton, J. M., Nielsen, C. & Frankel, K. A. (2012). *J. Synchrotron Rad.* **19**, 1006–1011.