Data Analysis: Keeping Pace with Extraordinary Change

Gwyndaf Evans

To cite this article: Gwyndaf Evans (2015) Data Analysis: Keeping Pace with Extraordinary Change, Synchrotron Radiation News, 28:6, 31-36, DOI: 10.1080/08940886.2015.1101325

To link to this article: https://doi.org/10.1080/08940886.2015.1101325
Introduction

It is a very exciting time in the field of macromolecular crystallography for those of us who are fortunate enough to be involved in the development of instrumentation and software methods. The driver for much of this change has been the remarkable developments in synchrotron light sources and beamline instrumentation over the last two decades. In the 1990s, the ESRF, APS, and SPring-8 set the tone for many of these developments and the 2000s capitalized on them by seeding a host of medium-energy national light sources around the world.

Macromolecular crystallography (MX) has been, and still is, a synchrotron success story. The number of macromolecular structures deposited in the Protein Data Bank (PDB) (www.rcsb.org) [1] using X-ray methods since the 1990s has grown exponentially, with a doubling in the total number of PDB entries roughly every five years (Figure 1a). The impact of the arrival of third-generation sources is evident when one looks at the increase in the percentage of X-ray structures solved using synchrotron sources (as opposed to in-house X-ray sources). From less than 20% in 1995, around 90% of all X-ray structures deposited in the PDB now arise from synchrotron data (Figure 1b).

Underpinning this exponential growth, and its dominance by synchrotron MX, have been revolutions in protein production and crystallization methods; a series of step changes in detector technology: image plates to phosphor-coupled CCDs to hybrid pixel array detectors (PAD) [2]; developments in automatic sample exchange systems [3, 4]; advances in optics and instrumentation that deliver higher flux densities and smaller X-ray beams at the sample [5, 6]; and significant developments in data analysis and automated processing [7–13].

The time taken to measure a single complete diffraction data set from a macromolecular crystal sample has reduced over the course of 25 years from around a few hours at a second-generation source using an image plate detector to tens of minutes using CCD detectors [14] to a few seconds at a modern third-generation source using the latest 100 fps hybrid PAD detectors [15]. A major consequence of this radical reduction in data collection time is how users have approached the challenge of structural biology using crystallography. It is now common place that hundreds, even thousands, of crystals are tested and characterized at synchrotron beamlines with the aim of finding just a few that diffract to a resolution required to answer the biological question posed by the user. Synchrotron MX has enabled structural biologists to address more and more challenging problems and has, in many respects, played a key part in several Nobel prizes in the last two decades by yielding structures of the F1-ATPase molecular machines [16], RNA polymerase [17], the 70 S ribosome structure and its subunits [18, 19], and many important G-protein coupled receptor (GPCR) structures [20].

The analysis of the copious quantities of X-ray diffraction data that were measured in order to deliver the ~100,000 PDB entries is its own
success story. It has, however, been a constant challenge to meet the high expectations of the community of macromolecular crystallographers. Nowadays, this is highlighted by their desire to screen hundreds or thousands of crystals, measure complete data, analyze that data in as close to real-time as possible, and arrive at a final structure—or at least an electron density map—before they have completed their shift of synchrotron beamtime. Such real-time analysis gives users confidence that the data meet the requirements so they can proceed with studies of further crystals.

MX Data Analysis

The staple method of recording diffraction data in MX is, and has been for 30 years, the rotation method [21]. The rotation method involves the rotation of a crystalline sample through a small angle around a fixed axis and the simultaneous measurement of the excited Bragg diffraction intensities on an area detector. The determination of atomic structure by X-ray crystallography requires the precise measurement of these X-ray intensities, from which crystallographic structure factors amplitudes are calculated. The structure factor phases are typically deduced by one of two methods: (1) experimentally by measuring additional data sets from similar crystals where a known and specific perturbation to the structure has been introduced (e.g., addition of a heavy element atom or perturbation of the resonant scattering of a heavy element atom), the so-called isomorphous replacement [22] and/or anomalous scattering methods [23–25]; or (2) computationally by using a prior knowledge of the atomic structure of a similar homologous protein molecule [26, 27]. Given structure factor amplitudes and phases, an electron density map representing the contents of a crystallography asymmetric unit can be calculated and must be interpreted by building a protein model into the 3D density map. At this stage, it is common that the number, identity,
and sequence of the protein amino acid chain is known and this can assist in map interpretation and building protein models.

Broadly speaking, this process (Figure 2) has not changed in decades, but the rate at which data analysis and structure determination can happen has changed radically. Significantly, the user base of MX has also changed; once the domain of expert structural biologist crystallographers, MX is now used as a technique by molecular biologists, biochemists, the pharmaceutical industry, and many structural biologists who are not specifically trained as crystallographers. This change has placed new demands on software and algorithm developers and increased the need for automated structure solution software and other expert-systems [9, 11, 28, 29] that can partly play the role of crystallographer.

Traditional Synchrotron Data Analysis Approaches

Data Integration, Scaling, and Merging

In the last 30 years, approaches to evaluating the intensities of Bragg diffracted spots on 2D detectors have evolved steadily for MX in the guise of programs like MOSFLM [30], XDS [31], HKL2000/Denzo [32], and D*TREK [33]. Each of these programs has adopted a linear approach to the problem and encompasses diffraction image I/O, algorithms for diffraction spot finding, indexing (assignment of miller indices), refinement of experiment and crystal cell parameters, integration of diffraction intensities, placing intensities on a common scale and merging symmetry equivalent measurements to form a unique set of Miller indices, associated intensities and their errors. Up to about a decade ago, the users would commonly perform the integration of raw diffraction data manually while at the beamline. Diffraction images would be visually inspected first, indexed to determine a data collection strategy, and then integrated after or during data collection. This initial round of data analysis was aimed at giving the user the peace of mind that the data was complete, of high enough quality and therefore sufficient for structure determination. In favorable cases, the structure would be determined at the beamline, but more commonly the data was backed up to a portable hard drive and/or transferred to the user home laboratory for post-analysis.

Back at the home laboratory, the data would be reanalyzed in the cold light of day and a final assessment made of the experiment outcome. Data might be combined with other data sets from prior synchrotron visits to attempt structure determination.

Data Volumes and Rates

Data collection strategies, as well as being determined by sample properties, are also a function of the detector technology. The small but finite readout noise, and readout time of around a second, associated with CCD detectors encouraged the use of wide-sliced data collection where each image corresponded to around a degree of crystal rotation. The significant readout time, during which the detector was not measuring, necessitated the use of X-ray shutters to prevent needless exposure of the sample to X-rays and hence radiation damage. Consequently, exposure times of less than 0.5 s were rarely used as they increased the risk of introducing systematic error from imperfect shutter performance or goniometer synchronization.

The introduction of hybrid PAD detectors with readout times of a few milliseconds or even less revolutionized synchrotron MX in the mid-2000s by enabling shutterless data collections [36, 37] whereby the crystal was rotated continuously throughout the full data collection and diffraction images were streamed onto disk storage. This new data collection style had the effect of decreasing the time for a full data set by more than an order of magnitude (Table 1) and made the use of fine rotation widths possible producing data sets tens of Gb in size.

In a field driven by throughput and the need to characterize and record data from many hundreds of samples, the speed benefits were significant but enabling the measurement of so-called fine-sliced data [38] brought an added benefit in terms of data quality [37].

Typical peak data rates from MX area detectors had increased from around 15 Mb/s (typical CCD) to 600 Mb/s (Pilatus3 6M), creating new challenges for the administrators of scientific computing resources at synchrotron sites. Data had to be moved from the detectors (in the case of Diamond Light Source, five Pilatus 6M detectors) to local or centralized storage and eventually backed up on a long-term storage medium and/or copied to a portable hard drive for users to take to their home laboratories.

It is now relatively common for user groups to measure 500 Gb of data or more in a single visit of less than a day. Not only is this problematic for synchrotron facilities in terms of storage and data rates, but also becomes a massive burden for the users in terms of data analysis, both during the synchrotron beam time and afterwards at their home laboratory. At Diamond Light Source and other sites, this problem has been addressed by significant investment in the automatic analysis of data. It

---

Table 1: Typical data collection parameters and data set sizes for representative CCD and hybrid PAD detectors used routinely at Diamond Light Source since 2007

<table>
<thead>
<tr>
<th>Detector</th>
<th>Readout time</th>
<th>File size (Mb)</th>
<th>Angle (degs)</th>
<th>Time (s)</th>
<th>Time* (s)</th>
<th>Volume (Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADSC Q315r</td>
<td>0.9 s</td>
<td>18.9</td>
<td>1.0</td>
<td>0.9</td>
<td>252</td>
<td>3.4</td>
</tr>
<tr>
<td>Pilatus-3 6M</td>
<td>&lt;1 ms</td>
<td>6 Mb</td>
<td>0.10</td>
<td>0.01</td>
<td>18</td>
<td>10.8</td>
</tr>
</tbody>
</table>

---
must, however, be accompanied by continued investment to deal with
the moving target that is large-scale data management [39].

Automated Data Analysis Pipelines

The high speed at which samples can be exchanged and data col-
clected using modern detectors and automated hardware has made it
close to impossible for users at beamlines to manually analyze every
data set as it is collected or soon thereafter. Diamond Light Source was
an early adopter of Pilatus PAD detectors for MX beamlines and, as
such, faced this problem early on. The solution was to automate the task
of initial data processing and image analysis to provide quick feedback
for the users about crystal and data quality while also attempting to give
users a carefully analyzed, high-quality data set to take back to their
home institute (Figure 3).

At Diamond, the result was a script called FastDP [8, 40] that would
typically produce data merging statistics for inspection by the user
within 1–3 minutes of the data collection ending. A more comprehen-
sive pass through the data was simultaneously launched using an ex-
pert system called Xia2 [28] and would take about 10–20 minutes. On
these time scales, users were able to respond rapidly at the beamline to
various scenarios dictated to them by sample quality and behavior and
thereby optimize the use of their valuable beamtime.

Interestingly, the initial response from many in the synchrotron MX
community was one of scepticism; there was a strong sense that the
automated data analysis pipelines would not be able to do as well as
a capable crystallographer (or indeed any capable human), but over a
period of months, this mood of scepticism soon turned to one of cau-
tious optimism and, finally, relief. In the intervening years since the late
2000s, this mood has evolved into a dependency on automated analysis.
The availability of the results of raw data analysis within minutes
of completing a data collection has since become an integral part of the
synchrotron MX experience and is seen to be as important to an MX
beamline’s operation as the presence of a detector or an X-ray mirror.

At Diamond, it is estimated that over 50% (conservatively) of all
protein structures submitted to the PDB were solved using data inte-
grated using the automated data analysis pipeline. This illustrates both
that users now have a good level of confidence in the quality of auto-
matic data analysis, with respect to their own abilities, and/or that user
groups are swamped with data to analyze and do not have the resource
or time to reanalyze their many data sets. Needless to say, the auto-
mated analysis is successful mainly on the most straightforward cases
where diffraction is of good quality and samples are relatively well
behaved; e.g., unambiguous symmetry determination, high resolution
data, etc. The automated pipelines are not yet sophisticated enough to
manage the most challenging cases where, for example, samples contain multiple lattices or diffract to very low resolution with significant anisotropy in the diffraction. This forms part of the challenge for future software developments.

MX has pioneered automated data analysis approaches at synchrotrons and these ideas are now being adopted by other techniques, such as small-angle X-ray scattering (SAXS) and tomographic imaging.

Data Analysis for New Data Collection Approaches and New Sources

The emergence of microfocus beamlines in the last decade has spawned interest in methods to compile complete data sets from many incomplete sets from multiple crystals. The radiation damage suffered by protein microcrystals (<1000 μm³) usually means that complete data cannot be recorded from a single crystal. The inhomogeneity of the diffraction properties of crystals samples (termed non-isomorphism) often gives rise to problems when data from many crystals are merged together in an attempt to form a complete set. A number of groups have developed clustering methods to address this [41, 42], and software packages such as BLEND [41] are geared towards dealing with the extreme end of this problem where individual crystals yield only very small amounts of data (a few degrees or less). The analysis and merging of multi-crystal synchrotron data has addressed a number of challenging protein structures, including several GPCR structures [43] and a number of viral protein structures where crystal sizes range between 1 and 10 μm on each side [44, 45].

A recent development in MX has been the introduction of serial femtosecond crystallography (SFX) [46] using X-ray Free Electron Lasers (XFELs), most notably at the Linear Coherent Light Source (LCLS) in California. From a synchrotron perspective, XFELs have raised the bar in the field of MX in three ways. First, they have encouraged a fresh look at methods of sample preparation and sample delivery that are an alternative to standard mounting methods for the rotation method. Second, they have pushed the X-ray optics and instrumentation demands to a new level, both in terms of beam delivery and in terms of sample environment. Finally, SFX, producing single diffraction shots per static crystal, has demanded a fresh look at the data analysis due to the relative uncertainty in some XFEL beam parameters. As a result, new software has been written, most notably CrystFEL [47] and cctbx.xfel [48], and new methods to manage or model the inherent partial measurement of Bragg reflections in SFX data [49–51]. The new data collection approaches have necessitated a ground-up re-evaluation of diffraction data processing. The modelling of crystal size, X-ray wavelength distribution and mean, crystal mosaic domain size, and other parameters have become essential components of obtaining meaningful measurement of diffraction intensities and their errors to allow the determination of atomic structure. New software packages like DIALS (dials.diamond.ac.uk) have emerged in light of these new sources and data analysis challenges and are aimed at bridging the gap between synchrotron and XFEL data analysis.

In particular, multicrystal data collection and analysis has emerged at the forefront of methodologies in recent years and DIALS is able to respond by offering the ability to perform parameter refinement and integration simultaneously for multiple crystals and, more importantly, acting as a toolbox for new developments in a rapidly shifting field [52].

Conclusion

Diffraction data analysis and structure determination in macromolecular crystallography is, in many cases, a routine procedure that can be automated. As with other automated procedures in MX, such as crystallization and sample exchange, automated data analysis serves the user by removing a task that must be repeated many times and freeing up researchers to address more challenging questions. In the case of structural biologists, automation enables them to readily address challenging crystallographic problems: larger protein complexes, membrane proteins, viral targets, etc. Automated data analysis pipelines are still far from fully replacing the experienced crystallographer, but a number of groups worldwide are now working to build expert systems that learn from the now significant library of data sets and structures that have been accumulated in the last two decades. In turn, this raises questions of data/structure validation and the availability of raw data against which to develop such algorithms and software. Fortunately, significant effort is already ongoing in this area (e.g., IUCr Diffraction Data Deposition Working Group chaired by Professor John R. Helliwell: http://www.iucr.org/resources/data/dddg), and developers of software will surely benefit from a strong tradition of openness and collaboration in macromolecular crystallography.