Proceedings of the Symposium on Synchrotron Crystallography, Krynica, Poland 2001

Automatic System for Crystallographic Data Collection and Analysis

W. Minor[a], M. Cymborowski[a] and Z. Otwinowski[b]

[a]Department of Molecular Physiology and Biological Physics
University of Virginia, Charlottesville, Virginia 22908, USA
[b]Department of Biochemistry, UT Southwestern Medical Center at Dallas
Dallas, TX 75235, USA

During the last 10 years the rate of new protein structures determined by X-ray crystallography has risen about tenfold. The use of high flux sources was instrumental in this growth. There are numerous advantages of using synchrotron radiation for protein crystallography: rapid data collection, use of micro-crystals and the ability to conduct measurements at wide range of wavelengths. The rate-limiting step is often the ability to analyze and back up a fast stream of data produced by a multi-module CCD detector. The goal of the newly developed HKL-2000 package is to integrate all computational activities that have to be performed during the data collection experiment. The Graphical Command Center of HKL-2000 organizes and forwards the data collection parameters to the display, indexing, strategy, simulation, refinement, integration, scaling, and merging tasks. Data acquisition can become a part of data processing (or vice versa), which includes indexing, integration, scaling, and even phasing. The increase in internet band width will provide an opportunity to remotely interact with the experimental setup and perform the synchrotron experiment from the home laboratory.

PACS numbers: 61.10.Nz, 87.15.Aa

1. Introduction

During the last 10 years, the rate of new protein structures determined by X-ray crystallography has risen about tenfold [1]. The use of high flux sources was instrumental in this growth [2]. The fraction of protein structures reported in *Science, Nature* and *Cell*, obtained with the use of synchrotron radiation increased from 35% in 1993 to over 90% in 2001 (Fig. 1) [3]. There are numerous advantages of using synchrotron radiation for protein crystallography which include:

a) rapid data collection that allows for fast crystal screening, collection of high resolution data and high throughput of projects,

b) the ability to use micro-crystals,

c) the ability to conduct measurements at a wide range of wavelengths in order to maximize the anomalous signal.

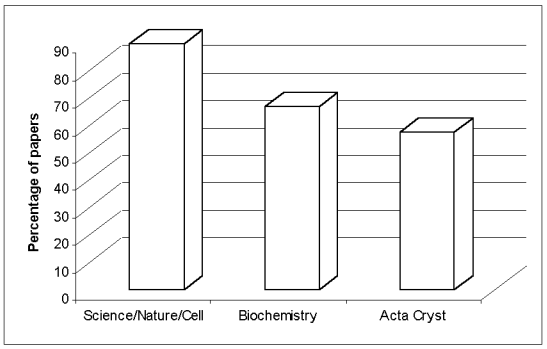

Fig. 1. The usage of synchrotron facilities as reported in selected scientific journals in the year 2000.

The rate-limiting step is often the ability to analyze a fast stream of data produced by a multi-module CCD detector [4]. Efficient data analysis changes the protein crystallography experiment as the data collection is a part of a process that includes indexing, integration, scaling, and phasing. In this approach, the final result of experiment is a high quality electron density map that provides assurance that the experiment was successful. The expected result of efficient data collection/processing is not only high-throughput crystallography but also reduction of the effort needed to produce proteins and grow crystals of sufficient quality for structure solution. This approach will become an essential part of the structural genomics effort.

2. Experiment

A typical protein crystallography experiment sounds very simple: a single crystal is placed in the beam and an X-ray diffraction pattern is recorded as a series of images while the crystal is rotated through the angular range, sufficient to record desired completeness. In practice, experimenters use complicated experimental protocols with a goal of collecting optimal data for phasing and/or refinement [5, 6]. These protocols are often designed on the experimenter's previous experience with the particular beamline and take into account beamline hardware limitations. For instance the lack of collision maps quite often limit the use of the kappa goniostat to a single-axis operation and additional axes are used only for

crystal recovery. For that reason most of the 63 synchrotron beamlines (HKL Research, private communication) work with a single-axis goniostat despite the clear advantage of a 3-circle system. The single-axis goniostats simplify the experimental protocols but at the expense of the overall data quality and completeness. In addition, the separation of data acquisition from data backup and analysis makes optimal data collection much more difficult, as the editing of the script files takes more time than the collection of a full data set. In practice, the experiment tends to be designed around the detector, beamline hardware, and computing resource limitations.

3. HKL-2000

The goal of the newly developed HKL-2000 package [4] is to integrate all computational activities that have to be performed during the data collection experiment. The Graphical Command Center (GCC) of HKL-2000 organizes and forwards the data collection parameters to the display, indexing, strategy, simulation, refinement, integration, scaling, and merging tasks. Data acquisition can become a part of data processing (or vice versa), which includes indexing, integration, scaling, and even phasing.

The Command Center consists of three components: a database, a transition state engine (a set of rules that define possible atomic changes of the database), and a Graphical User Interface (GUI). It is based on the idea of a single database that stores all the information about data processing and data collection. The database is a dynamic one; it can describe not only the data already collected, but also those being collected and even those planned or considered to be collected. Each data entry step or program execution step, including the data collection program, induces a change in the database. One of the main functions of the GUI is to provide user input and editing of the database. The complexity of the database requires the creation of a hierarchical access to the information.

The GCC is working in multi-group mode where uniform series of diffraction images form one 3D group. There is no limit on the number of 3D groups and in the case of non-uniformity in the series (e.g., found during data analysis), the 3D group can be split into two or more smaller 3D groups. The smallest 3D group can consist of one image. At the moment the only limitation is the assumption that a set of 3D groups have an a priori known relative orientation. In practice, it means that data were collected from one sample at one site with potentially different settings of goniostat, data collection axis, crystal translation, detector position, detector mode (e.g., binned/unbinned), or exposure level.

The coordination of all phases of the experiment allows for automatic update of all statistics and reports when new data are collected. The utilization of interactive experiments in which data analysis is done online make it possible to adjust the data collection strategy to guarantee the desired result, particularly with regard to data completeness and detection of anomalous signal (Fig. 2).

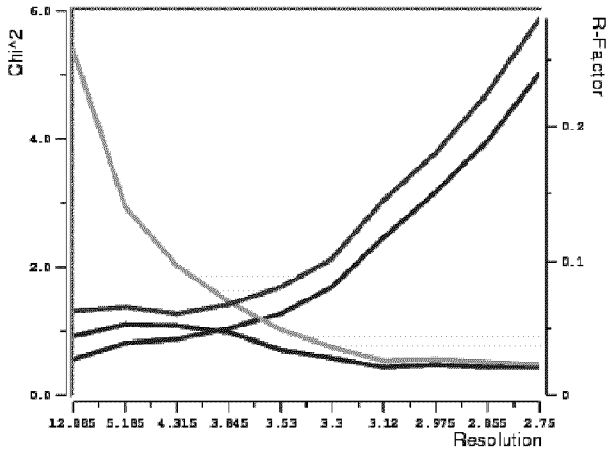

Fig. 2. The χ^2 vs. resolution plot of merged (orange) and unmerged (blue) anomalous pairs for typical Se-Met experiment performed on ID-19 beamline at Structural Biology Center at Argonne. Black and red lines represent R-factors for merged and unmerged anomalous pairs, respectively.

Even for frozen crystals, radiation damage on high brilliance beamlines severely limits the quality and quantity of information that can be measured from each crystal [7]. The radiation damage can be determined by evaluating real time changes in scale and B-factors and may allow for fast derivation of the optimal strategy to minimize the total dose of X-rays. The problem of efficient experimental design is solved by calculating the completeness as a function of start and range of the phi angles to be collected. From such a function the user is able to identify optimal start angles for the desired coverage. In the future, strategy program will have an option to include a previously collected data set to identify optimally small sector needed to complete measurements of all unique reflections.

HKL-2000 diagnostic tools can monitor the performance of experimental system. During the integration stage, there are several tools that may monitor crystal slippage. Even more tools are provided for the data scaling stage. By observing scale factors one can detect poor crystal alignment. The other tools provide information about X-ray shutter malfunction, spindle axis alignment, and internal detector alignment. The final inspection of outliers may again provide valuable information about the detector quality. The clustering of outliers in one area of the detector may indicate a damaged surface. If most outliers are partials, it may indicate a problem with spindle backlash or shutter control. The zoom mode may be used to display the area around the outliers to identify the source of a problem (the existence of a satellite crystal or single pixel spikes due to electronic failure, for example).

4. Experimental protocols

The Graphical Command Center of HKL-2000 provides control over all phases of the experiment:

a) crystal evaluation and alignment,
b) setup of experiment,
c) running the data collection,
d) experiment monitoring,
e) data processing,
f) data archiving.

The experiment can be simplified by the design of pre-defined experimental protocols for different classes of projects. The GCC provides default values and ranges for reasonable input values and statistical output. The user interface allows for overriding these default values and even allows for input of unreasonable values, in the latter case with warning and confirmation by the user. The defaults should be good enough for most projects and close to optimal for difficult projects. This system is based on heuristics derived from authors experience and when fully implemented it will become the expert system that leads the experimenter through the whole process. The expert system will also allow for experimental protocol templates. This may allow the automation of projects that involve large numbers of similar crystals (in drug design and discovery, for example). The expert systems should have a large impact on the reliability of results, as they will catch problems that can be missed in current routine analysis.

5. Automation

The fully automated data collection facility for protein crystallography requires the system that will automate, control and coordinate the following activities:

a) crystal changing and alignment,
b) beamline optimization,
c) absorption edge measurement,
d) sample evaluation and data integration, merging, and scaling,
e) phasing.

There are several groups, including commercial companies, which are working on automatic crystal changing and alignment hardware and software [8]. Only a slight increase in temperature is observed during the transfer process, so the sample always remains at a low temperature.

Several synchrotron beamlines are working on automatic beamline optimization and automation of the absorption edge measurements. However, to the knowledge of the authors of this paper, there is no working system that would allow for interpreting the absorption scan and automatically selecting energies required for multiwavelength anomalous diffraction (MAD) experiments.

The new version of HKL-2000 will be able to automatically score the crystal quality (mosaic spread, resolution, etc.) and hence will allow for rapid and automatic crystal screening. There will be two modes of operation:

a) Selection of the best crystal from the set in cryo-storage.

b) Full data collection on the first crystal that will meet pre-defined quality criteria.

After the crystal is selected, the system will start automatic data collection according to predefined experimental protocol (Fig. 3). The Graphical Command Center will allow for description of complex data collection protocols encountered in high-precision measurements.

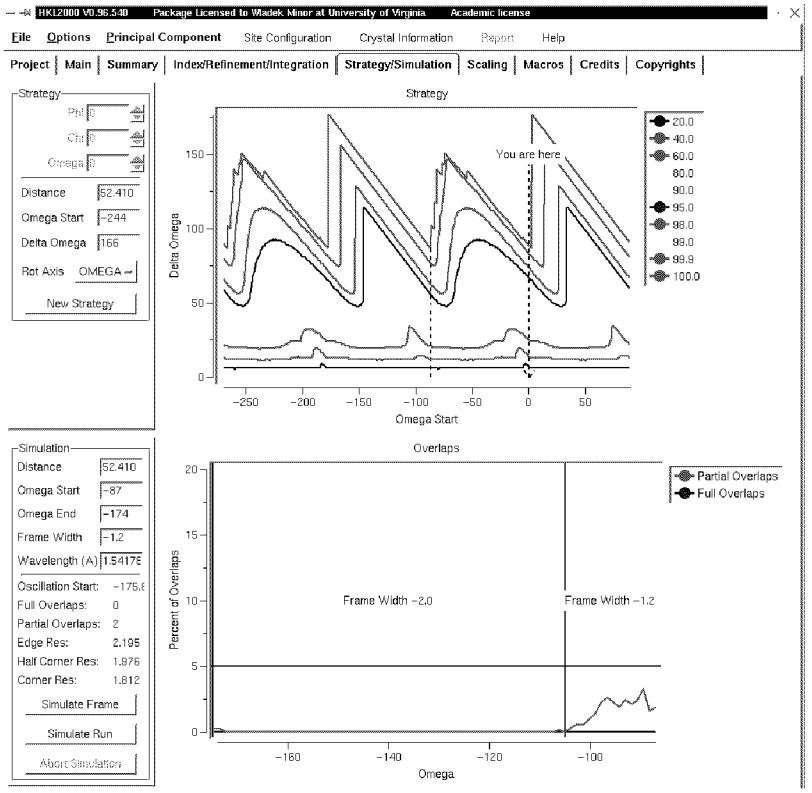

Fig. 3. The strategy/simulation window of HKL-2000 in full automatic mode. The experimental parameters are chosen from the set of data being the result of experiment simulation.

We assume that in the future 80% of the data will be collected in automatic mode. The remaining 20% will be collected over the internet with the use and help of the expert system.

6. Remote (over internet) data collection

The increase in internet band width and in particular the arrival of the internet-2 will provide an opportunity to remotely interact with the experimental setup and perform the synchrotron experiment from the home laboratory. The experimenter will use the same graphical control programs that will coordinate all crystallographic data collection and processing steps. The activities inside the experimental hutch could be supervised (again over the internet) by a camera placed close to the experimental system. The data will be automatically transferred to home computer. Remote data collection will require a fast network, but in fact, the current internet connection between the University of Virginia and the Argonne National Laboratory allows for real-time data backup over the internet. The remote data collection may be the preferable mode for the experimenter under the condition that automation of sample mounting and alignment will work impeccably.

The prototype of the remote data collection system is used routinely at the University of Virginia.

Acknowledgments

This work was supported by NIH grant GM-53163. The authors would like to thank Zbyszek Dauter and Andrzej Joachimiak for their comments and the SBC staff at Argonne's Advanced Photon Source for their help during synchrotron experiments.

References

[1] H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook, *Nature Structural Biology* **7,** Suppl, 957 (2000).

[2] W. Minor, D. Tomchick, Z. Otwinowski, *Structure Fold Des.* **8,** R105 (2000).

[3] A. Brysiak, Z. Otwinowski, W. Minor, in: *Int. Symp. on Synchrotron Crystallography — SYNCRYS 2001, Krynica (Poland) 2001*, SYNCRYS 2001 Book of Abstracts, p. 63.

[4] Z. Otwinowski, W. Minor, in: *International Tables for Crystallography*, Vol. F, Ed. M.G. Rossmann, E. Arnold, Kluwer Academic Publishers, Dordrecht 2000, p. 226.

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

[6] Z. Dauter, M. Dauter, *J. Mol. Biol.* **289,** 93 (1999).

[7] E. Garman, *Acta Crystallogr. D Biol. Crystallogr.* **55,** 1641 (1999).

[8] E. Abola, P. Kuhn, T. Earnest, R.C. Stevens, *Nat. Struct. Biol.* **7,** Suppl, 973 (2000).