Vol. **114** (2008) *ACTA PHYSICA POLONICA A* No. 2

Proceedings of the 7-th National Meeting of Synchrotron Radiation Users

Automation and Remote Synchrotron Data Collection

M. Gilski[*]

Department of Crystallography, Faculty of Chemistry
A. Mickiewicz University, Poznań, Poland
and
Center for Biocrystallographic Research, Institute of Bioorganic Chemistry
Polish Academy of Sciences, Poznań, Poland

X-ray crystallography is the natural choice for macromolecular structure determination by virtue of its accuracy, speed, and potential for further speed gains, while synchrotron radiation is indispensable because of its intensity and tuneability. Good X-ray crystallographic diffraction patterns are essential and frequently this is achievable through using the few large synchrotrons located worldwide. Beamline time on these facilities have long queues, and increasing the efficiency of utilization of these facilities will help in expediting the structure determination process. Automation and remote data collection are therefore essential steps in ensuring that macromolecular structure determination becomes a very high throughput process.

PACS numbers: 07.85.Qe, 61.05.cp, 61.05.cf

1. Introduction

The post-genomic research in biology and medicine depends critically on the rate of progress towards the understanding of complex biological processes at the molecular level, which in turn depends on our capability of rapidly determining the three-dimensional molecular structures of proteins. The goals of the several European projects are to develop, assemble and provide an integrated platform for high-throughput structure determination using X-ray crystallography with synchrotron radiation, which will boost such capabilities throughout Europe. Knowledge of the structure is critical to understanding the biological function of proteins, particularly if the genetic sequence is already determined.

Contemporary crystallography, in particular the protein crystallography, to a great degree owes its fast progress to the accessibility of synchrotron radiation.

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

Access to beamlines at synchrotrons is important to the elucidation of the structure of biological molecules. With the high-throughput sequencing effort of the Genome Projects, many genes and even entire genomes are already sequenced but their function is as yet unknown. One of the bottlenecks is the long drawn process of structure determination of protein crystals. Acquiring structural information on biological macromolecules on a genomic scale constitutes the main goal of Structural Genomics (SG). Nowadays, the only techniques for determining three--dimensional structures of biological macromolecules at atomic level and at a rate appropriate for SG are biological X-ray crystallography and NMR spectroscopy.

The cornerstone of any SG initiative is synchrotron radiation. While about 10 to 15 years ago most 3D structures of biological macromolecules were still solved using local X-ray laboratory sources, the use of synchrotron radiation is now paramount. The fraction of deposited structures in the Protein Data Bank (PDB) for which synchrotron radiation was used rose from 28% in 1992 to over 80% in 2000 [1]. Currently, when the number of deposited structures in the PDB is four-times bigger, this fraction is almost the same.

2. High-throughput gene-to-structure pipeline

The process of determining macromolecular crystal structures, which consists of a number of relatively simple steps, can be described as a pipeline [2]. In many ways, the process may be thought of as a simple sequence of operations: protein production, crystallization, diffraction, phasing, model building, refinement and deposition of coordinates. That this sequence is well understood leads one to hope that all the steps of the pipeline may be readily automated. The reality is of course much more complicated since the process of proceeding from crystal to structure deposition is more labyrinthine than the straight pipeline.

The number of EU-funded projects like Structural Proteomics In Europe (SPINE) [3] and Biocrystallography (X) on a Highly Integrated Technology Platform for European Structural Genomics (BIOXHIT) [4] are involved in the development of a data-collection pipeline. In collaboration with similar European initiatives e-Science Resource for High-Throughput Protein Crystallography (e-HTPX) [5] and AutomateD CollectioN of DatA (DNA) [6] researchers are developing hardware and software basis for automation of the whole process of data collection, remote access and control of the synchrotron experiment.

The developments span the whole range of components required to produce an efficient pipeline linking the crystallization of a protein to the delivery of its completed 3D-structure. This pipeline should operate with minimal user intervention and should be made fully accessible to the wider life sciences research community through remote access facilities.

The high-throughput approach of automation, miniaturization and parallelization is increasingly being applied to the techniques of protein sample production. Laboratory methods are changing as these technologies become standard, but information tracking software has not managed to keep pace.

Contemporary macromolecular crystallography has reached a level of automation, where complete computer-assisted robotic crystallization pipelines are capable of cocktail preparation, crystallization plate setup, inspection and interpretation of results. At this stage researcher can use a software system called Protein Information Management System (PIMS) [7], to keep track of individual samples using a unique sample-holder identifier [8]. PIMS is a version of the industrial Laboratory Information Management System (LIMS) which is being developed to support data management for all stages of protein production, from target selection to crystallization and is suitable for tracking the complex and rapidly evolving laboratory practices associated with protein production in the context of structural biology.

All information stored in PIMS database are then passed into the e-HTPX which is managing all data concerning obtained crystals, freezing, cryoprotection and soaking protocols. It includes also details of the planned diffraction experiment, shipment organization to the synchrotron site, together with information about shipping agents and current location of the sample.

After the safety approval is obtained, the samples are dispatched to the synchrotron. At the same time e-HTPX is sending the sample information records pertinent to the actual experiment to the Information System for Protein Crystallography Beamlines (ISPyB). This database, as currently deployed at the European Synchrotron Radiation Facility (ESRF), allows the logging of information concerning sample data [9], sample shipping and data collection and allows the harvesting of data-reduction statistics from DNA system [6].

3. Automation of data collection

Automation of a synchrotron macromolecular crystallography experiment may be divided into two major parts: provision of the X-ray beam automation and data collection with data processing automation.

Beamline alignment automation include advanced technologies developed towards standard, automated and easy-to-operate beamlines providing high quality stable X-rays automatically delivered to the sample. It should ensure that the users have access to a beam that is stable in both intensity and position during their experiments, without having to carry out manual alignment procedures [10–12].

Data collection automation should include processes for sample mounting/dismounting, crystal aligning and screening, while data processing module is responsible for image indexing and integration, data reduction and scaling. All of the aspects described above have been addressed in the automation program currently being developed at the synchrotron centers in Europe and they are incorporated into a DNA system, a part of the e-HTPX and BIOXHIT high throughput protein crystallography projects.

DNA is an expert system capable of analyzing X-ray diffraction data and making sensible decisions about how best to collect data. This makes use of both

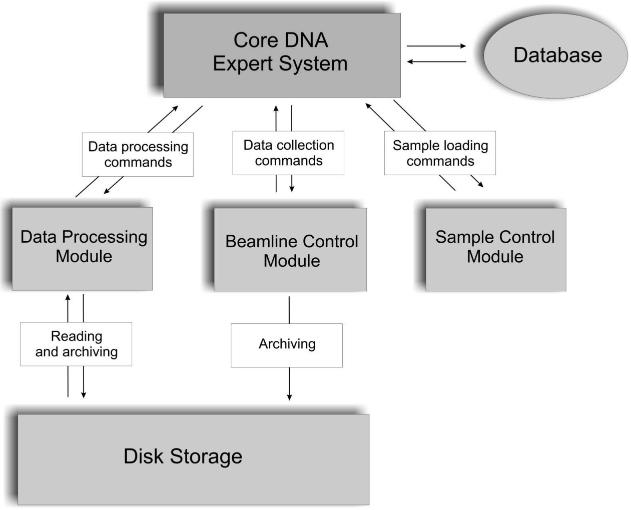
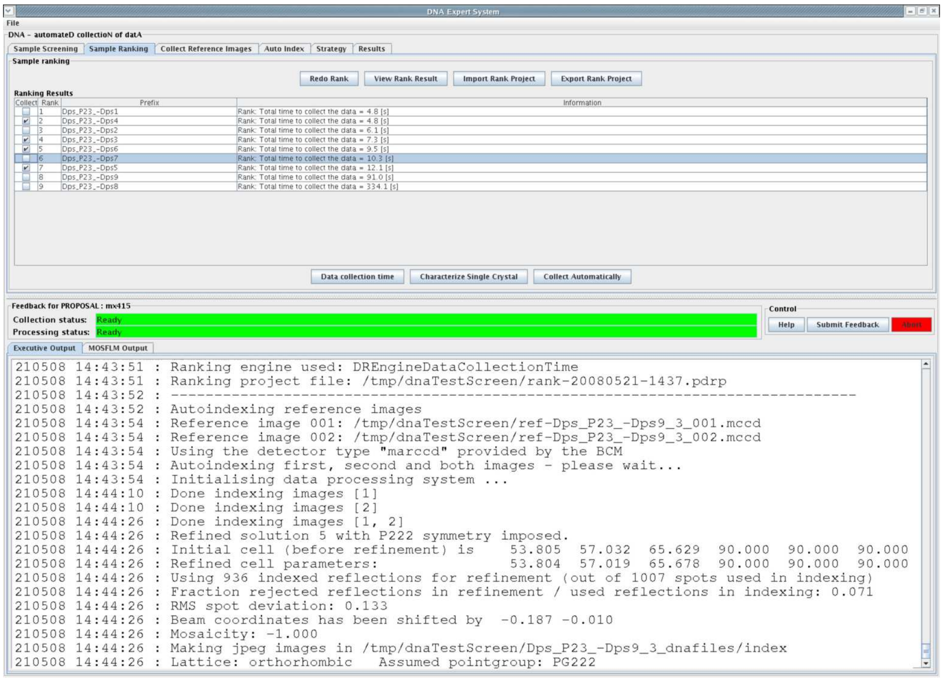


Fig. 1. A schematic representation of the DNA architecture.

existing software for data processing and software developed specifically for the task. The modular architecture (Fig. 1) enables rapid development of functionality, and assists the distributed development across a number of sites.



Ranking Results

Collect	Rank	Prefix	Information
☐	1	Dps_P23_-Dps1	Rank: Total time to collect the data = 4.8 [s]
☑	2	Dps_P23_-Dps4	Rank: Total time to collect the data = 4.8 [s]
☐	3	Dps_P23_-Dps2	Rank: Total time to collect the data = 6.1 [s]
☑	4	Dps_P23_-Dps3	Rank: Total time to collect the data = 7.3 [s]
☑	5	Dps_P23_-Dps6	Rank: Total time to collect the data = 9.5 [s]
☐	6	Dps_P23_-Dps7	Rank: Total time to collect the data = 10.3 [s]
☑	7	Dps_P23_-Dps5	Rank: Total time to collect the data = 12.1 [s]
☐	8	Dps_P23_-Dps9	Rank: Total time to collect the data = 91.0 [s]
☐	9	Dps_P23_-Dps8	Rank: Total time to collect the data = 334.1 [s]

Feedback for PROPOSAL : mx415

Collection status: Ready

Processing status: Ready

```
210508 14:43:51 : Ranking engine used: DREngineDataCollectionTime
210508 14:43:51 : Ranking project file: /tmp/dnaTestScreen/rank-20080521-1437.pdrp
210508 14:43:52 : ----------------------------------------------------------------------------
210508 14:43:52 : Autoindexing reference images
210508 14:43:54 : Reference image 001: /tmp/dnaTestScreen/ref-Dps_P23_-Dps9_3_001.mccd
210508 14:43:54 : Reference image 002: /tmp/dnaTestScreen/ref-Dps_P23_-Dps9_3_002.mccd
210508 14:43:54 : Using the detector type "marccd" provided by the BCM
210508 14:43:54 : Autoindexing first, second and both images - please wait...
210508 14:43:54 : Initialising data processing system ...
210508 14:44:10 : Done indexing images [1]
210508 14:44:10 : Done indexing images [2]
210508 14:44:26 : Done indexing images [1, 2]
210508 14:44:26 : Refined solution 5 with P222 symmetry imposed.
210508 14:44:26 : Initial cell (before refinement) is    53.805  57.032  65.629  90.000  90.000  90.000
210508 14:44:26 : Refined cell parameters:               53.804  57.019  65.678  90.000  90.000  90.000
210508 14:44:26 : Using 936 indexed reflections for refinement (out of 1007 spots used in indexing)
210508 14:44:26 : Fraction rejected reflections in refinement / used reflections in indexing: 0.071
210508 14:44:26 : RMS spot deviation: 0.133
210508 14:44:26 : Beam coordinates has been shifted by  -0.187 -0.010
210508 14:44:26 : Mosaicity: -1.000
210508 14:44:26 : Making jpeg images in /tmp/dnaTestScreen/Dps_P23_-Dps9_3_dnafiles/index
210508 14:44:26 : Lattice: orthorhombic   Assumed pointgroup: PG222
```

Fig. 2. Graphical User Interface of DNA used to drive the system.

The beamline control system knows about the capabilities of the beamline, and the data processing module stores all parameters of data processing step. The communication is mediated through an executive system, which simply knows how to ask for jobs to be done, but knows nothing of how they are done. Although the DNA system includes a large number of components, a single user interface (Fig. 2) is used to drive the system. This provides a simple interface to the experiment, given in terms that the scientist is familiar with, and sensible set of default values. With this system it is possible to collect a good quality data set with a single button press.

4. Remote data collection

Automation of data processing is critical in case of remote data collection. In DNA and e-HTPX projects, a key aim is to allow the remote user to process and analyze the collected data. Traditionally this would involve providing a remote graphical interface to the data processing packages. Automating data processing can allow the same functionality but with far fewer transactions, making remote access more reliable.

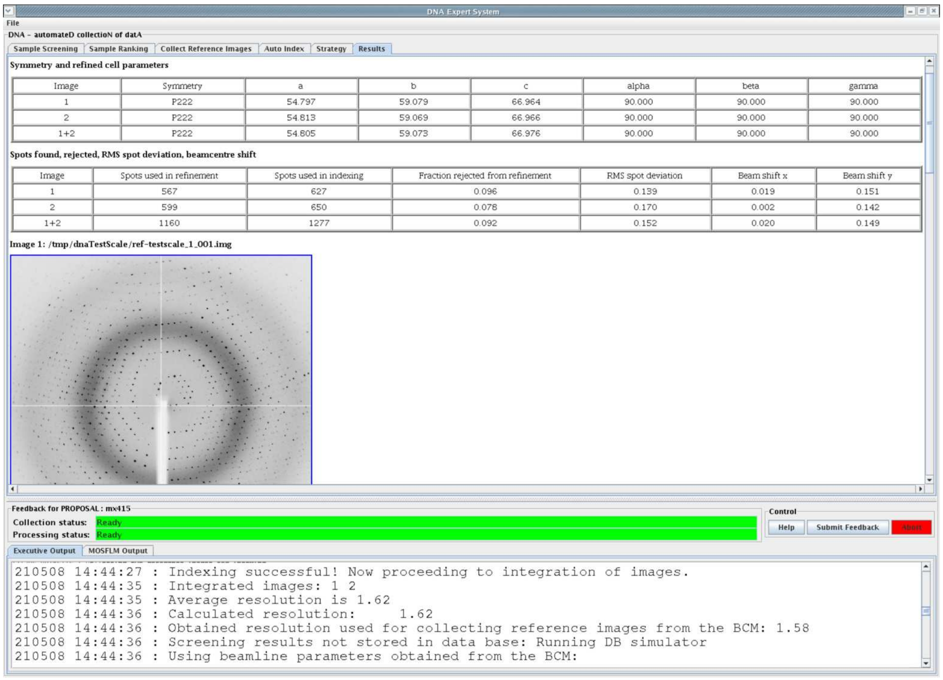
The European project BIOXHIT together with ESRF plays the major role in that field, providing technologically advanced macromolecular beamlines.

The users of such beamlines have the option to collect data remotely and, during last years, more and more experiments have been scheduled for remote access, thereby saving travel time and expenses. While remaining at their home institutions, remote users conduct experiments by means of advanced software tools, like DNA and ISPyB, enabling network-based systems monitoring and control. Remote users have the capability to mount, center, and screen samples as well as to collect, analyze, and backup diffraction data. Automated sample handling like mounting and dismounting are accomplished with the robotic system called Sample Changer [13, 14].

As the result of a joint collaborative effort between the ESRF, the EMBL Grenoble outstation and BM14 (MRC France), eight Sample Changers are installed on the seven ESRF MX beamlines.

Beamline and experimental control is carried out using DNA, and additional remote monitoring of the experiment and data backup is supported with several Web-based applications, like ISPyB database.

Program BEST [15] is used for optimal choice of the data collection parameters. It is a novel software that can accurately predict some of the characteristics of the data yet to be collected based on a few initial images. Then MOSFLM [16], a leading European software for processing of X-ray data, is performing an automated data processing. For rapid crystal ranking and data analysis, the diffraction images collected during screening are automatically analyzed, and the results, which include the number of spots, Bravais lattice, unit cell, estimated mosaicity, and maximum diffraction resolution, are visible through DNA user interface



Symmetry and refined cell parameters

Image	Symmetry	a	b	c	alpha	beta	gamma
1	P222	54.797	59.079	66.964	90.000	90.000	90.000
2	P222	54.813	59.069	66.966	90.000	90.000	90.000
1+2	P222	54.805	59.073	66.976	90.000	90.000	90.000

Spots found, rejected, RMS spot deviation, beamcentre shift

Image	Spots used in refinement	Spots used in indexing	Fraction rejected from refinement	RMS spot deviation	Beam shift x	Beam shift y
1	567	627	0.096	0.139	0.019	0.151
2	599	650	0.078	0.170	0.002	0.142
1+2	1160	1277	0.092	0.152	0.020	0.149

Image 1: /tmp/dnaTestScale/ref-testscale_1_001.img

Feedback for PROPOSAL : mx415

Collection status: Ready

Processing status: Ready

```
210508 14:44:27 : Indexing successful! Now proceeding to integration of images.
210508 14:44:35 : Integrated images: 1 2
210508 14:44:35 : Average resolution is 1.62
210508 14:44:36 : Calculated resolution:     1.62
210508 14:44:36 : Obtained resolution used for collecting reference images from the BCM: 1.58
210508 14:44:36 : Screening results not stored in data base: Running DB simulator
210508 14:44:36 : Using beamline parameters obtained from the BCM:
```

Fig. 3. Results screen of the DNA showing the preliminary crystal checking data and diffraction images.

(Fig. 3). In this way, users can screen crystals 2–3 times more rapidly with less human error.

Remote experimenters have access to the same tools as local users and have the capability to mount, center, and screen crystalline samples and to collect, analyze, and backup diffraction data.

5. Conclusions

The appropriate level of automation will depend of course on the problem in hand. If your data are good and the problem relatively straightforward, then the complete automation approach may be appropriate. If, on the other hand, one is trying to push the boundaries of macromolecular crystallography, then one might need to do most of the work by hand.

Applying automation techniques to data processing offers a couple of key advantages. Firstly the results will be reproducible, since exactly the same procedures will be followed every time. Secondly, the decision about the best program to use will not be limited by the user's experience, for instance preferring one program over another due to familiarity alone. Finally, for novice users and those who care more about the biology than the technique, the processing and analysis can be performed with little or no effort.

During last two years, in the Poznań Training, Implementation and Dissemination (TID) Center, together with ESRF in Grenoble and University of Oulu, Finland, a remote and automated data collection had been tested and the pipelining system is now being implemented into laboratory practices [17, 18]. The Poznań TID Center, located at the Center for Biocrystallographic Research in Poznań, which is affiliated with the Institute of Bioorganic Chemistry, Polish Academy of Sciences, was established in October 2004 by the BIOXHIT TID Committee as one of two such training centers in the Central and Eastern Europe. Its mission is to train young crystallographers and implement and disseminate the results obtained within the BIOXHIT integrated project.

References

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

[2] R. Stevens, I. Wilson, *Science* **293**, 519 (2001).

[3] A. Betova, F. Cipriani, S. Cusack, S. Delageniere, J. Gabadinho, E.J. Gordon, M. Guijarro, D.R. Hall, S. Larsen, L. Launer, C.B. Lavault, G.A. Leonard, T. Mairs, A. McCarthy, J. McCarthy, J. Meyer, E. Mitchell, S. Monaco, D. Nurizzo, P. Pernot, R. Pieritz, R.G.B. Ravelli, V. Rey, W. Shepard, D. Spruce, D.I. Stuart, O. Svensson, P. Theveneau, X. Thibault, J. Turkenburg, M. Walsh, S.M. McSweeney, *Acta Crystallogr. D* **62**, 1162 (2006).

[4] http://www.bioxhit.org.

[5] http://www.e-htpx.ac.uk.

[6] A.G.W. Leslie, H.R. Powell, G. Winter, O. Svensson, D. Spruce, S. McSweeney, D. Love, S. Kinder, E. Duke, C. Nave, *Acta Crystallogr. D* **58**, 1924 (2002).

[7] A. Pajon, J. Ionides, J. Diprose, J. Fillon, R. Fogh, A.W. Ashton, H. Berman, W. Boucher, M. Cygler, E. Deleury, R. Esnouf, J. Janin, R. Kim, I. Krimm, K.L. Lawson, E. Oeuillet, A. Poupon, S. Raymond, T. Stevens, H. van Tilbeurgh, J. Westbrook, P. Wood, E. Ulrich, W. Vranken , L. Xueli, E. Laue, D.I. Stuart, K. Henrick, *Proteins* **58**, 278 (2005).

[8] F. Cipriani, F. Felisaz, L. Launer, J.-S. Aksoy, H. Caserotto, S. Cusack, M. Dallery, F. di-Chiaro, M. Guijarro, J. Huet, S. Larsen, M. Lentini, J. McCarthy, S. McSweeney, R. Ravelli, M. Renier, C. Taffut, A. Thompson, G.A. Leonard, M.A. Walsh, *Acta Crystallogr. D* **62**, 1251 (2006).

[9] G. Ueno, R. Hirose, K. Ida, T. Kumasaka, M. Yamamoto, *J. Appl. Crystallogr.* **37**, 867 (2004).

[10] S. Arzt, A. Beteva, F. Cipriani, S. Delageniere, F. Felisaz, G. Forstner, E. Gordon, L. Launer, B. Lavault, G. Leonard, T. Mairs, A. McCarthy, J. McCarthy, S. McSweeney, J. Meyer, E. Mitchell, S. Monaco, D. Nurizzo, R.B.G. Ravelli, V. Rey, W. Shepard, D. Spruce, O. Svensson, P. Theveneau, *Prog. Biophys. Mol. Biol.* **89**, 124 (2005).

[11] E. Pohl, U. Ristau, T. Gehrmann, D. Jahn, B. Robrahn, D. Malthan, H. Dobler, C. Hermes, *J. Synchrotron Rad.* **11**, 372 (2004).

[12] Y. Gaponov, N. Igarishi, M. Hiraki, K. Sasajima, N. Matsugaki, M. Suzuki, T. Kosuge, S. Wakatsuki, *J. Synchrotron Rad.* **11**, 17 (2004).

[13] G. Snell, C. Cork, R. Nordmeyer, E. Cornell, G. Meigs, D. Yegian, J. Jaklevic, J. Jin, R. Stevens, T. Earnest, *Structure* **12**, 537 (2004).

[14] A.E. Cohen, P.J. Ellis, M.D. Miller, A.M. Deacon, R.P. Phizackerley, *J. Appl. Crystallogr.* **35**, 720 (2002).

[15] A.N. Popov, G.P. Bourenkov, *Acta Crystallogr. D* **59**, 1145 (2003).

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

[17] M. Jaskolski, M. Gilski, *Academia* **3**, 8 (2007).

[18] M. Gilski, *Synchr. Radiat. Nat. Sci.* **6**, 95 (2007).