Data Management for PX Structure Determination: CCP4 and BioXHIT

Peter Briggs, CCLRC Daresbury Laboratory, Warrington WA4 4AD, UK

BioXHIT (*Biocrystallography (X) on an Highly Integrated Technology Platform for Structural Genomics*) [1] is a four-year integrated project funded within the 6th Framework Programme of the European Commission. BioXHIT is coordinating scientists at all European synchrotrons along with leading software developers with the aim of consolidating the whole process of macromolecular structure determination using X-ray protein crystallography (PX). The project aims to deliver an integrated platform with a completely automated approach that encompasses crystallisation, data collection and structure determination.

Automation & the need for data management

Several projects are developing automated structure determination software pipelines [2,3,4], joining together one or more computational units (programs or other applications which perform a single part of the process) to cover some or all of the stages from the data processing and reduction through to model building, refinement and model validation. These projects are vital if Structural Genomics efforts are to succeed.

In all cases it is essential to accurately record, organise and track the data used in and generated by the procedures. Components used in the pipeline need to access the required data on demand, and be able to store their outputs for use by other components downstream in the process. An accurate record of the process is also required when depositing the resulting structures in public databases such as the wwPDB [5].

Different applications may have very different needs and the situation is further complicated by the possibility that data will be stored in a number of different systems at geographically diverse locations (for example a LIMS, facility database or local data store).

CCP4 and BioXHIT

The Collaborative Computational Project No4 (CCP4) [6] is a UK initiative based at CCLRC Daresbury Laboratory and which provides a software suite for macromolecular structure determination by X-ray crystallography. Currently CCP4 offers basic data management within its graphical user interface system CCP4i [7] (which records information such as date, status, input parameters and files associated with the run of a particular task) and through technologies such as Data Harvesting [8].

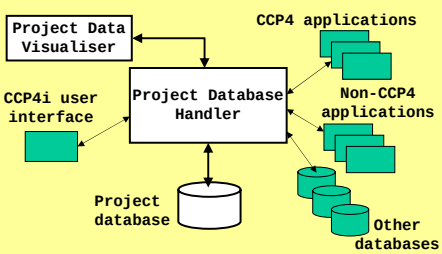
CCP4's role within the BioXHIT project is to implement data management and project tracking within the structure solution software pipeline, by building on the existing data management functionality within CCP4i, and by exploiting links with projects such as e-Science initiatives *e.g.* [9] which face similar challenges.

The tools being developed by CCP4 are described below and will use open standards developed within BioXHIT in order to exchange data efficiently and accurately between applications and databases. The tools will be made freely available as part of the CCP4 software suite.

The CCP4 contribution will fill the need for project tracking within the BioXHIT structure solution software pipeline, and consists of three components:

Project Database Handler

The Project Database Handler is a brokering application which will mediate interactions between the project database and other applications and external databases (local or remote). It will act as a single point of access to the data for the applications that talk to it.



Above: schematic representation showing the interactions with the Project Database Handler

Database for Project & Data Tracking

A database will be designed and implemented which will be capable of storing project history information (the links between the steps performed travelling through a pipeline) and data history (the provenance and evolution of information as the project progresses).

Visualisation Tools

These tools will be Interfaces to the database that provide display the project data in selective views, to focus on particular aspects of data-flow or logical-flow – for example as work-flow diagrams.

Project Status

Work is ongoing to provide a prototype of the Project Database Handler, after which work will begin on redesigning the database. A programmer funded by BioXHIT is currently being recruited to work on the project full-time. In addition to the BioXHIT Partners the system will be used within CCP4i and the CCP4 Automation Project, and parts of the eHTPX [10] and DNA [11] projects which are being carried out at the SRS Daresbury in collaboration with other partners.

More information can be found at **`http://www.ccp4.ac.uk/projects/bioxhit.html`** or by contacting Peter Briggs (*p.j.briggs@dl.ac.uk*).

Acknowledgements

CCP4 is funded by BBSRC. PJB is funded by CCLRC from CCP4 industrial income. The BioXHIT Project is funded by the European Commission within its FP6 Programme, under the thematic area "Life Sciences, genomics and biotechnology for health", contract number LHSG-CT-2003-503420

References

[1] BioXHIT: http://www.bioxhit.org
[2] CCP4 Automation Project: *contact Charles Ballard c.c.ballard@dl.ac.uk*
[3] Elves: Holton & Alber, *PNAS* **101**, 1537-1542 (2004)
[4] ANTS: Brunzelle *et al*, *Acta Cryst* **D59,** 1138-1144
[5] Worldwide Protein Data Bank: http://www.wwpdb.org
[6] CCP4: http://www.ccp4.ac.uk
[7] CCP4i: Potterton *et al*, *Acta Cryst* **D59** 1131-1137 (2003)
[8] Data harvesting: Winn, *CCP4 Newsletter* **37** (October 1999)
[9] Data Portal: http://www.e-science.clrc.ac.uk/web/projects/dataportal
[10] e-HTPX: http://www.e-htpx.ac.uk
[11] DNA Project: http://www.dna.ac.uk