dbCCP4i: Tracking data in PX Structure Determination Software Pipelines

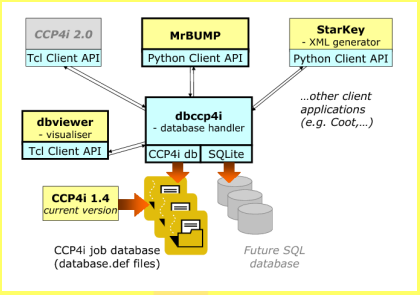
***Peter Briggs**, Wanjuan Yang and Ronan Keegan, STFC Daresbury Laboratory, Warrington WA4 4AD, UK*

Introduction and Background

The European Integrated Project "BIOXHIT" [1] aims to deliver an integrated technology platform for protein structure determination by X-ray crystallography (PX). A key part of this platform is the development and integration of automated structure solution software pipelines which cover various parts of the post-data collection stages of the macromolecular structure determination process. Such pipelines are useful in high-throughput contexts such as Structural Genomics, but are equally well suited to use by novice crystallographers.

Typically software pipelines generate large amounts of data in their operation, and accurate recording and tracking of the data is important both for the automated pipeline and for the end user who will need to interpret the final result. This poster describes the dbCCP4i project tracking system [2] that aims to address some of these issues, and how this system has been incorporated within the automated "MrBUMP" system [3] to track the progress of runs and present its output in a more easily understandable format.

Project Tracking System for the Structure Solution Software Pipeline



The project tracking system consists of a number of key components show in the figure (left):

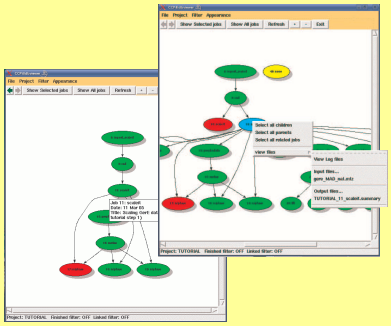
- Project database handler
- Database for storing job data
- Client APIs
- Visualisation tools

Each of these is described briefly in the following panels. More information can be found at www.ccp4.ac.uk/projects/bioxhit_public where version 0.2 of dbCCP4i is also available for download

Data Visualiser: dbviewer

The dbviewer client application aims to present the tracking data in an intuitive manner to the end user, by showing the jobs with links between them indicating the transfer of data. The user can interact with the view to see the files and other data associated with each job.

The viewer uses the Graphviz package[5] as an engine for generating the layouts of the history graphs. The viewer is written in the Tcl/Tk scripting language and so is able to reuse many of the tools already developed in CCP4i.



Project Database Handler

The project database handler dbCCP4i is a small server program that handles interactions between the job database and other programs.

The handler is written in Python but communications are via sockets, so the clients can be written in any language. Tcl and Python "client API" libraries are provided to hide the details of the socket communications and simplify access to the database. The system runs on both UNIX/Linux and MS Windows platforms.

Project Tracking Database

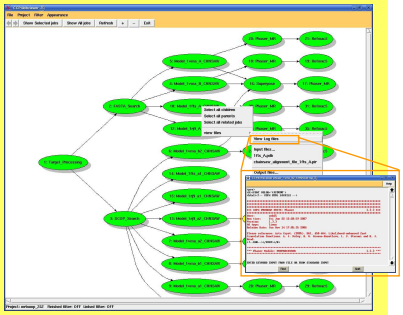
The dbCCP4i system is built on the existing database used within the graphical user interface CCP4i [4], and records information about each program run, including the date, input parameters, associated input and output files, and the outcome of the run.

At present dbCCP4i uses the same "database.def" file infrastructure as CCP4i in order to store the job data, to ensure compatibility with existing data. More sophisticated data storage mechanisms (for example, SQLite) are also being investigated for future use.

Using dbCCP4i in the MrBUMP Molecular Replacement Pipeline

MrBUMP is a framework that automates the determination of macromolecular structures via the Molecular Replacement method (MR). In MR, a trial model based on a known structure is used as a template for extracting the structural information of the target from the experimental reflection data. As the success of the MR procedure is critically dependent upon the choice of trial model, the approach used in MrBUMP emphasises the generation of a variety of possible search models; it has already been used successfully to solve several MR problems [6-8].

In good cases MrBUMP will find a single clear solution to an MR problem; in other cases it will suggest a number of likely search models that can be investigated manually by the scientist – typically around 10-15 search models will be generated (in atypical cases it could be ten times that number). Each model also undergoes rounds of processing by various programs. As a result the end user faces a challenge when reviewing and evaluating the results of the MrBUMP run, and in earlier versions end users needed to filter through many different output files and directories.



***Above:** Example of the dbviewer showing the output from a MrBUMP job*

The most recent release of MrBUMP (0.4.0) uses dbCCP4i to record information on the MR process in order to address this problem and present the data in a more intuitive graphical form via the dbviewer (see right). The graphical view makes it much easier for the end user to monitor the progress of the MrBUMP run, and to more easily find and examine a particular file associated with the processing of each search model. Thus the use of dbCCP4i substantially improves the usefulness of MrBUMP for the end user. MrBUMP version 0.4.0 incorporating dbCCP4i is available for download from www.ccp4.ac.uk/MrBUMP

The viewer is available for general CCP4 usage and could easily be used by other automation projects. The use of dbCCP4i allows for much more complexity in these automated systems whilst making the outcome much clearer and accessible to the user, allowing them to arrive at a solution for their structure more quickly.

Authors

Peter Briggs
Senior CCP4 software developer
p.j.briggs@dl.ac.uk

Wanjuan Yang
Principal BIOXHIT software developer
w.yang@dl.ac.uk

Ronan Keegan
Lead MrBUMP software developer
r.m.keegan@dl.ac.uk

References

[1] The BIOXHIT Project: www.bioxhit.org
[2] P.J.Briggs and W.Yang *CCP4 Newsletter on Protein Crystallography* **45** (Winter 2006/7) www.ccp4.ac.uk/newsletters/newsletter45/articles/ccp4_bioxhit.html
[3] R.M.Keegan and M.D.Winn *Acta Cryst.* **D63** (2007) 447-457
[4] E Potterton et al. *Acta Cryst.* **D59** (2003) 1131-1137
[5] The Graphviz package: www.graphviz.org
[6] Obiero et al. *Acta Cryst.* **F62** (2006) 757-760
[7] Karbat et al. *J. Mol. Biol.* **366** (2007) 586-601
[8] El Omari et al. *Acta Cryst.* **F62** (2006) 949-953

Acknowledgements

The work on dbCCP4i has largely been undertaken within the BIOXHIT project, which is funded by the European Commission within its 6th Framework Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2003-503420. Both the dbCCP4i and MrBUMP projects have also been generously supported by Collaborative Computational Project No.4 (CCP4), which is part of the UK's Science and Technology Facilities Council (STFC), and which is also supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC).

CCP4 The BIOXHIT Project bbsrc biotechnology and biological sciences research council Science & Technology Facilities Council