The BIOXHIT Project

www.bioxhit.org

WP 5.2: CCP4i Data Tracking in the PX Structure Determination Software Pipeline

Peter Briggs and Wanjuan (Wendy) Yang, CCLRC Daresbury Laboratory, Warrington WA4 4AD, UK

Introduction

A key part of the integrated technology platform being delivered by the BIOXHIT project is the development of automated structure determination software pipelines that cover the post-data collection stages of structure solution by protein X-ray crystallography (PX). These pipelines need to accurately record and track the data that they produce, both for their own operation and for final deposition of the determined structures.

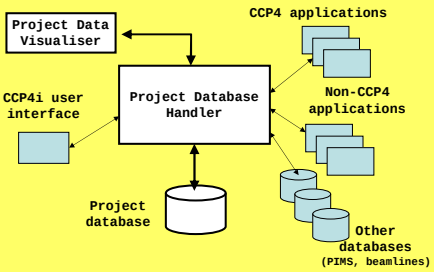
BIOXHIT Workpackage 5.2 is principally concerned with developing a system for performing this data management in order to address the needs of software pipelines. This poster reports on the progress that has been made by CCP4 (BIOXHIT Partner 10) towards this end in the last year.

Project tracking system for the structure solution software pipeline

Components of the system

- Project database handler
- Database for Project Data & Tracking
- Visualisation tools

These components and their relationships are shown schematically in the figure (right), and are described in more detail in the sections below.

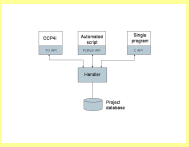


Key considerations

- Implement a system for both manual and automated structure determination
- Allow multiple database back-ends
- Gather as much information from client programs as possible automatically
- Open architecture accommodating heterogenous software components

Project Database Handler

The Project Database Handler **dbCCP4i** is a brokering application that mediates interactions between the project database and an external applications. It is a lightweight Python server program that acts as a single point of access to the data for external applications and hides the implementation of the database from them.



The handler and clients communicate using XML-encoded messages passed via sockets; however these details are hidden within "client API" libraries. (So far Tcl and Python libraries have been implemented).

The tracking database backend is based on the CCP4i "def file" format. An SQLite backend is also available which will be used for the proposed "knowledge database". Work is also ongoing to integrate the handler into CCP4i (the CCP4 graphical user interface).

Database for Project Data & Tracking

A database is being designed and implemented which will be capable of storing both project data (the information used by each step in a pipeline) and project history (the steps taken and the provenance and evolution of information as the project progresses).

The current "schema" is based on the CCP4i project and job database model, which supports simple tracking of project history by recording "jobs" (runs of software components). This is implemented using a CCP4i "def file" backend.

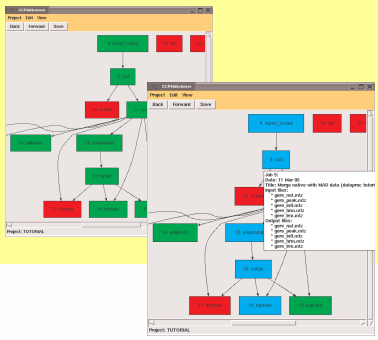
In the next year of the project we aim to expand the tracking database and associated functionality, and supplement it with the implementation of an SQLite "knowledge database" component that will be able to store the crystallographic data items that are common between different applications in the software pipeline.

The knowledge database content will be grown from a small "kernel" of data items and will be informed by a SQL schema developed in the previous year of the project.

Visualisation Tools

We have developed the **dbviewer** program as the first version of a visualiser for the project history data stored in the current version of the tracking database.

This is a Tcl client application of the handler which uses the Graphviz package (www.graphviz.org) to show the history as a directed graph. Additional functions allow the user to explore various aspects of the history data to better understand the progress of their structure determination project.



XML Tagging

We have worked with the EBI to develop an XML schema that describes the final project history for deposition. We have also developed a client application **starkey** that can generate the project history marked up using the BIOXHIT XML tags.

Next steps

Over the next 12 months the aims are to:

- Make release 0.1 of dbCCP4i and the visualiser to users and developers
- Finish integrating dbCCP4i into CCP4i
- Work with users to identify improvements to the visualiser
- Work with developers to integrate dbCCP4i into their applications and pipelines
- Expand the tracking database, and develop and implement the "knowledge database"

Availability

The initial version of the handler, database and visualiser are due for public release by the end of February 2007, and will be available at www.ccp4.ac.uk/projects/bioxhit_public/ along with other CCP4 BIOXHIT deliverables.

Acknowledgements

Workpackage 5.2 is co-ordinated by the **Collaborative Computational Project No4** (CCP4), which provides a software suite for macromolecular structure determination by X-ray crystallography. Funding for the work described here has been provided by the European Commission as part of the BIOXHIT project, and by CCLRC. Valuable contributions to the work have also been made by the developers of the CCP4 Automation (HAPPy), XIA/e-HTPX and CRANK Projects, and by the EMBL Hinxton and EMBL Hamburg Partners within the BIOXHIT project.

CCP4

See http://www.ccp4.ac.uk/projects/bioxhit.html for more information

CCLRC

3rd BIOXHIT Annual Meeting
19th - 21st February 2007
Diamond, Didcot, UK