Contribution ID: **105** Type: **not specified**

Workflow and Data Management for Nuclear Magnetic Resonance.

Wednesday, 28 March 2012 14:20 (20 minutes)

Conclusions

At the start of the project, a survey of potential users showed enthusiasm for a number of key objectives of this work, notably improved access to remote services and a more formal data audit trail. We are currently in the process of evaluating to what extent the web-based interface we have developed fulfils the user demands.

At a technical level, we have had to address the difficulties of building workflows from processes with highly complex input and output parameter types, and handling tricky format interconversions. Both of these would not have been possible without the CCPN API [MEMOPS: Data modelling and automatic code generation; Fogh et al.; J. Integrative Bioinformatics, 7(3):123, 2010], which in turn derives much of its power from the way in which the library code is generated
automatically from a UML data model: this makes the maintenance of code representing a highly complex underlying data model tractable even within a relatively niche area such as NMR.

Description of the Work

Nuclear Magnetic Resonance (NMR) is an important method for the analysis of macromolecular structure and dynamics. There are a host of different computational process involved in the processing and interpretation of NMR data, and traditionally these have used a wide range of formats. Interconversion between these formats is not trivial - the underlying information is extremely complex - and this has limited the ability to chain these processes into generic workflows.

As part of the WeNMR project, a number of the NMR analysis processes have been made available as web portals, many of them making use of GRID services on the server side. We are developing a set of wrappers around these portals that use a single format (CCPN) as input and output and expose the methods as WSDL-defined web services. Alongside this, we are developing a web interface that provides access to the individual services, and allows them to be linked together within workflows. The basic architecture is a GWT client with a java/Hibernate server deployed under Tomcat. The workflow management itself is delegated to a Taverna server.

A useful spin-off from tracking the processes is the management of the data itself. NMR data analysis is rarely a linear process. Typically, many different processes are set up with subtly different input data resulting in many different versions of output data, some which are then used as input for further rounds of refinement. Our interface keeps track of all these various versions and helps scientists keep track of their work - something which a survey at the start of the project identified as a significant problem.

As a case study we are developing a workflow that takes a single set of input data, sends it to four different analysis services (CYANA ARIA, UNIO, AutoStructure) and integrates the results.

Impact

This is work under development / just completed and so it is too early to judge the impact fully. The main objectives are to (i) improve accessibility to the WeNMR services, (ii) to provide a framework for the development of novel workflows that can address scientific problems. With respect to both objectives, the major impact is likely to come from increasing the usability for users who are not specialists in computational structural biology. As structural biology becomes more established and the focus increasingly moves from the methods themselves to the biological relevance of the results, there will be ever more demand from non-specialist users who need a simple method to obtain reliable results. The establishment of standardized interfaces shared between different programs, and seamless data transfer that does not require users to intervene, or to understand the details of program specific formats, should be very helpful in this regard.

In addition, the interface has the potential to be extremely useful in terms of laboratory data management. There is a clear need to gather and track the large amounts of data generated in NMR spectroscopy. Historically there has been a reluctance to use such LIMS tools within scientific communities, but there is reason to hope that this effort may be more successful. Unlike biochemical laboratory work, NMR data processing is already heavily computerised, so there is no need for cumbersome data capture steps. The direct integration with the programs that are already being used in the field, and the automatic capture of data in a standard form, should further minimise the need for changing established work practices, thus minimising the demands on the user and increasing the ultimate take-up.

Overview (For the conference guide)

We are developing a web-based workflow and data management system targeted at scientists within the field of nuclear magnetic resonance (NMR). NMR is a key method for investigating biological macromolecules, and analysis of NMR data currently involves both manual and computational steps. For the computational stages there is a strong tradition of using distributed computing methods, as the popularity of the WeNMR VO testifies. However, presenting these methods so that they can be easily built into workflows presents significant technical challenges, particularly in terms of format interconversion. The prize at the end is a framework for the development of novel workflows that can address important scientific problems - particularly in the area of improving automated NMR data analysis - as well as improved accessibility to WeNMR services and the potential for improved lab data management.

Primary author: Dr IONIDES, Jonn (University of Cambridge)

Co-author: Dr FOGH, Rasmus (University of Cambridge)

Presenter: Dr FOGH, Rasmus (University of Cambridge)

Session Classification: Workflows: solutions currently in use