

The ^{my}Grid / Taverna toolkit for workflow based bioinformatics

Tom Oinn¹, Robert Stevens², Phillip Lord²

¹European Bioinformatics Institute mo@ebi.ac.uk ²University of Manchester, UK

Life science researchers traditionally chain together database searches and analytical tools, using either complex scripts, or by manually copying between web pages. These "*in silico*" experiments are usually undertaken without support for the scientific process of managing, sharing and reusing the results, their provenance, and the methods used to generate them. The ^{my}Grid project [1] has developed a comprehensive loosely-coupled suite of middleware components specifically to support data intensive *in silico* experiments represented as workflows. 1000+ public domain resources are accessible via for example BioMOBY, SeqHound, BioMART and EMBOSS.

The project, which began in late 2001, is a UK EPSRC-funded e-Science pilot project made up of a consortium of five UK Universities, IT Innovation, and the European Bioinformatics Institute, supported by nine industrial partners, and it has now entered its second phase of funding and development. ^{my}Grid is available under the Lesser GNU Public License (LGPL) from <http://www.mygrid.org.uk>. The platform is written in Java.

^{my}Grid serves both expert bioinformaticians and typical biologists. ^{my}Grid has been used for building discovery workflows for investigations into, amongst others, Williams-Beuren Syndrome, Grave's Disease, Trypanosomiasis in cattle, small molecules, and analysis of protein families. A workflow can be built that represents complex analyses, such as chromosome walks, together with subsequent gene prediction and protein product characterisation involving 30+ services. Such workflows reduce the length of the task from days to hours. We will use this analysis, in two related demos, to show how ^{my}Grid is used by both groups of user.

Taverna is the popular ^{my}Grid workflow construction environment [2], and it is aimed at expert bioinformaticians. It allows the integration of disparate services into workflows and incorporates semantic web driven service discovery. Taverna and its associated components are inherently platform neutral. As well as being available as part of the ^{my}Grid releases, it is also available from <http://taverna.sf.net>

Demo Plan Part 1: Constructing an *in silico* experiment. The first part of the demonstration will show the construction and invocation of a bioinformatics analysis process using real world public services, showing the lifecycle of workflow construction. We will demonstrate the following aspects in detail:

(1) Installation and requirements of the Taverna system. As Taverna is a very lightweight client intended to give users access to existing distributed resources this section will be extremely short. (2) Location of available components through the Feta discovery tool. This tool allows users to search for services based on a description of the desired functionality, inputs and outputs. Without such a tool the user is faced with an overwhelming set of services. (3) Composition of the services into a workflow demonstrating Taverna's type reconciliation and implicit iteration mechanisms, showing a wide range of distributed services being composed into a complex workflow. (4) Invocation of the workflow thus constructed, showing status monitoring and management functions, reconnection to previously launched workflows. (5) Publishing of the workflow for subsequent use or adaptation by fellow bioinformaticians.

Those attending this demo will have a broad overview of how ^{my}Grid services are used to rapidly build and run workflows representing complex bioinformatics analyses.

Demo Plan Part 2: Running and managing *in silico* experiments. Taverna is just one part of the toolkit. ^{my}Grid has developed a portal for launching workflows, a metadata and a data repository. These collect and manage provenance records about the experiments and raw data identified at each point of the workflow. The Knowledge Annotation and Verification of Experiments component (KAVE) captures and stores provenance records of methods and purpose in RDF, and again semantically annotated by terms from an ontology. Life Science Identifiers are used throughout as a unifying mechanism for data generated by ^{my}Grid and data in the external services that ^{my}Grid accesses. This part of the demonstration will:

(1) Show the workflow constructed in part 1 executed through the ^{my}Grid Portal. (2) Show monitoring of the workflow enactment and notification of completion. (3) Inspection of result data and its associated provenance data, to verify and validate the outcomes of the experiment. (4) Follow the derivation path of each fragment of the results in this verification. (5) Finally, we will show how knowledge annotations on these data can aid a bioinformatician explore and use these data.

Those attending this demo will see how ^{my}Grid services can be used to change how bioinformaticians work by rapidly gathering results for subsequent evaluation based upon reliable and trustworthy provenance.

References

- [1] Carole Goble, Chris Wroe, Robert Stevens and the ^{my}Grid consortium, The ^{my}Grid project: services, architecture and demonstrator. In *Proc UK e-Science All Hands Meeting 2003*, pages 595-602, September 2003. <http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/128.pdf>
- [2] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat and Peter Li. *Taverna: A tool for the composition and enactment of bioinformatics workflows* *Bioinformatics* 20:(17) pp 3045-3054, 2004