

Enable Taverna workflows in a Shared Genomics causality workbench

1. What research is being done by the researcher(s)

The Manchester Asthma and Allergy Study (MAAS) research team propose that the development of allergies and asthma is likely to be a consequence of gene-environment interactions. MAAS has charted 1000 children through a birth cohort study lasting 10 years. Through analysis of these children's Single Nucleotide Polymorphisms, SNPs (i.e. a DNA sequence variation), and the measurements of environmental exposure in early life, researchers hope to identify the genetic predisposition and environmental toxins that influence the development and severity of asthma.

John New (Diabetologist) is combining diabetic patient data, routinely collected over a number of years in Salford, with corresponding SNP data from patients' DNA samples. Analysis of this combined data will determine which diabetic patients are most prone to developing long-term complications of their disease. The impact of any medical intervention given to the patients will also be assessed in the context of the SNPs for that patient.

The NIBHI Shared Genomics workbench software offers a clinician the capability to analyse large-scale complex genetic data using a High Performance Computing platform. The SNPs identified in this analysis as having a potential causal association to a particular disease, are ordered by statistical significance. High quality annotation of these results enables the clinical researchers to make a better assessment of the biological plausibility of the SNP having a causal role in the disease. In this the annotation helps to weed out any false positives that are returned from the statistical analysis.

The most efficient approach to generate this annotation is to re-use existing bioinformatics resources that have been linked together in workflows. A number of such workflows have already been developed using the Taverna platform, e.g. workflows to take a SNP ID and retrieve KEGG pathways (i.e. molecular interaction networks), information about relevant publications and gene summary information.

NIBHI bioinformaticians are interested in developing methods to mine this annotation data to identify significant annotation patterns within single data sets and correlations between multiple data sets. As well as using workflows to retrieve annotation data, identification of causal mechanisms may involve accessing computational services that quantify the physical effect of the genetic variation, e.g. estimating the change in thermodynamic stability of any protein whose coding sequence has been affected, or calculating the reduction in receptor domain accessibility. Web services already exist to provide these calculations and NIBHI researchers are interested in assessing the impact of novel workflows that combine both third-party annotation retrieval services and third-party computational services.

2. What are the issues that ENGAGE could address?

The Taverna platform is an excellent choice for the Shared Genomics annotation process providing it can be made to scale and perform well as a web service. Our aim is to retrieve relevant bio-health information in an XML format through the invocation of a high volume of Taverna workflows. After each workflow completes, the XML output would be parsed and displayed in real time to the user as a row annotation within their analytical results.

The current version of Taverna is presently spawned as a new process for each call received. This adds a several-second overhead to each call and uses a large amount of memory which

does not scale well for the aforementioned usage pattern. Support from ENGAGE would enable better usage of the Taverna workflow.

We believe the following improvements to Taverna are essential for the Shared Genomics use case:

- Install Taverna on a web server behind IIS (could use Tomcat and the JK connector).
- Handle requests from multiple clients concurrently who may be requesting different workflows.
- Decrease start-up times to enable quick service of each request.
- Ensure each workflow has sufficiently low memory consumption so as not to take the hosting web server out of memory.

3. Future benefit to user community

NIBHI will publish the new Taverna workflows that they develop through the myExperiment portal (as well as their own). Comments will be made by their bio-informaticians on the usage pattern of these workflows. This will provide an exemplar to the wider user community on how to utilise the capability to execute a high volume of workflows concurrently.

Taverna is primarily operated by users who have received some formal training. This limits Taverna's appeal with some researchers who could have otherwise made use of the bio-informatic workflows that have been developed on it. An immediate impact of this ENGAGE project would be to enable end users not familiar with Taverna to call its workflows from other information systems to annotate their data in real time.

The short-term benefits will be assessed by performance and scalability benchmarking before and after the ENGAGE work and by a qualitative evaluation exercise. The feedback received from this exercise and from the online myExperiment community will be used to make further revisions to the Taverna workflow engine.

One longer term benefit would be that new workflows authored as part of this research could be fed back and shared to the wider user community through the myExperiment (<http://www.myexperiment.org>) collaborative environment.