The RapidMiner Plugin for Taverna: bringing Data Mining Tools to Bioinformatics Workflows

Simon Jupp[1], James Eales[1], Simon Fischer[2], Sebastian Land[2],
Rishi Ramgolam[1], Alan Williams[1] and Robert Stevens[1]
{simon.jupp, james.eales, rishi.ramgolam, alan.r.williams, robert.stevens}@manchester.ac.uk,
{fischer, land}@rapid-i.com

[1]School of Computer Science, University of Manchester, Oxford Road. Manchester M13 9PL, UK
[2]Rapid-I GmbH, Stockumer Str. 475, 44227 Dortmund, Germany
Project Site: http://www.e-lico.eu/
Source code: http://taverna.googlecode.com/svn/unsorted/taverna-elico/

Knowledge discovery through pattern finding in data is central to modern molecular biology, which now has thousands of databases and similar numbers of tools for processing those data. Any data analysis in molecular biology involves gathering and processing data from many sources, even before the analysis for the central biological question takes place. Taverna (http://www.taverna.org.uk) is a workflow workbench that allows bioinformaticians to create data pipelines involving distributed Web services and other forms of tool; these workflows gather and manage data in order to perform analyses that ask biological questions.

RapidMiner (RM) is an open source, cross-platform application, released under the AGPLv3, that brings a large suite of data processing, visualization and data mining tools to bear upon tables of data, such as those that can be gathered by Taverna. A typical task for RM is to apply a series of operators to a table of gene products, their functions and locations to perform simple correlations of location and function. More sophisticated tasks will involve training classifiers over a set of features, selecting the best features and then applying the classifier to test data. The RapidAnalytics enterprise server from Rapid-I (http://rapid-i.com/) provides a platform for interacting with these data mining operators from RM via the RapidAnalytics execution service.

Through the RM plugin for Taverna we have combined the ability to gather and process data from many molecular biological sources with RM's data mining capabilities to provide a powerful tool for scientific analysis.

The RapidAnalytics execution service is a single polymorphic WSDL service. It takes a reference to the input file locations for the operator as input, along with a set of parameters including the name of the operator to execute. The polymorphic functionality, however, can make it difficult to work with in an environment like Taverna. For this reason the RM-specific plugin was developed. Using the plugin, the available operators within RM are exposed within Taverna. In addition we provide dialog-based interactions for setting input file locations and operator invocation parameters.

RapidAnalytics requires the data being processed to sit within the RapidAnalytics repository. This reduces the need to pass large amounts of data between services and improves the execution time on file transfer overheads associated with running distributed workflows. So, before data can be analyzed in Taverna, these data must be first uploaded onto the RapidAnalytics server. When any RM operator in Taverna is used, the user is given a configuration dialog. From this dialog a user can access the RapidAnalytics server in order to browse or upload new data to the repository.

The first release of the plugin is currently limited to a subset of RM operators. Simple operators that work on input files and generate a set of output files are easily handled. However, RM also contains some specialized operators, that we call dominating operators that control the execution of one or more sub data mining processes. This control currently requires some logic that cannot be expressed in Taverna via web services alone. These tools are important in data mining, so work is currently underway to extend the RM plugin.

The RM plugin already makes available a large number of data processing, visualization and data mining tools for bioinformatics analyses implemented as workflows within Taverna. In order to illustrate the benefits the RM plugin brings, we have developed several workflows to demonstrate its functionality in some typical bioinformatics tasks. These workflows are available for download from myExperiment in http://www.myexperiment.org/groups/402.html

This work is funded by the EU/FP7/ICT-2007.4.4 e-LICO project.