Design of a Scientific Workflow for the Analysis of Microarray experiments with Taverna and R

Marcus Ertelt

Proposal for a diploma thesis
December 2006 - May 2007
referees: Prof. Dr. Ulf Leser, PD Dr. Wolfgang Kemmner

1 Background

Microarrays are state of the art for the measurement of expression of thousands of genes in a single experiment. The quality of the final results of a microarray experiment is determined largely by the quality of the chips used (e.g. chips with a bad signal to noise ratio or scratches on the surface that influence only parts of the chip), the quality of the examined tissue samples and the appropriate choice of statistical tools and methods on the usually huge amount of data.

Taverna[1] is a program that allows to develop and use complex workflows over distributed systems. It is developed in Java and therefore runs on most common operating systems (Windows, Unix, Linux, MacOSX). Taverna allows users to create their own workflows using common script languages like Java Script or Beanshell. However, most processes of a workflow will probably be made available through webservices from all around the world so the majority of a workflow's computation can be performed with remote resources assuming the required webservices are available. The design process of a workflow takes place in the Taverna workbench, a graphical editor that allows to place and connect processes (called processors), inputs, and outputs, without forcing the developer to learn and use Taverna's own simple conceptual unified flow language (Scufl).

The R language[2] includes a comprehensive collection of statistical and analytic tools and function. It is free and available for all common operating systems. One of R's major

advantages in this context is the possibility to analyse microarray data with advanced methods thanks to the ongoing development of the Bioconductor packages[3].

There are many different kind of workflows and definitions for them. A workflow may be just a simple sketch of a flowchart. A business workflow may describe steps within a company for developing and marketing of a new product. Usually business workflows are more control-flow centric. Scientific workflows on the other hand are often data-flow oriented. Bowers and Ludäscher[4] for example describe scientific workflows in the following way: 'Scientific workflows are [...] unifying mechanism to combine scientific data management, analysis, simulation, and visualisation tasks.' For them a scientific workflow management system like Taverna '[...] is a problem-solving environment, that aims at simplifying the task of "gluing" these [workflow] steps together to form executable data management and analysis pipelines'. In the paper the authors give a detailed introduction to the concepts of workflows and describe a formal model of scientific workflows based on actor-oriented modeling and design. Actors are the principal components of a workflow similar to the processors of Taverna.

Most biologist involved in gene expression data analysis are not proficient with R, the Bioconductor packages or programming in general. Therefore an easy to use interface that completely hides the programing aspect of R and the details involved with using Bioconductor is necessary. With the help of Taverna we can create one or multiple workflows that do the complete analysis of a microarray experiment. Basically a scientist only needs to provide the data as input, execute the workflow, and then get the results after the calculations are finished. The workflow could be optimized to use the best and most commonly used methods, yet still allow for customization if wanted by the user.

Raw microarray data can be quiet huge in size. Over hundred megabytes of data are not uncommon. The analysis of these large data sets can take a lot of time depending on the computing power available and the implementation of the used function. The very same algorithm can require a lot more time in R than in C (for instance, a loop over all rows or columns of a large dataframe in R can take hours, but is much faster in native C). Therefore it may be a good approach to look into Taverna's ability to access remote resources. This means that one strong server runs R and does the computation while the usually old and slow personal computers in most laboratories can access the results via Taverna without having to do any major computing on their part. Small laboratories and institutions can run the webserver locally on the strongest computer available while all researchers there can then do their computations on it through Taverna.

In order to write and deploy R scripts as webservices you need to invest more work aside from creating a workflow with Taverna. One widely used tool to turn command-line applications into webservices is Soaplab which is supported by Taverna. 'Soaplab is a set of Web Services providing programmatic access to many applications on remote computers.'[5]

2 Objective

The goal of this thesis is to research the possibilities to create and deploy an executable scientific workflow for the analysis of microarray data with Taverna. A webserver setup will be developed and the Bioconductor functions will be added to it with the aid of Soaplab. In order to keep the system open and enhanceable a virtual environment will be used.

3 Related Work

This topic is of great interest to the e-science community. Due to its actuality there are no publications available yet. The EP-EMAAS-R-Taverna Project[6] for instance proposes a similar project using the Expression Profiler[7] and EMAAS - "Extensible MicroArray Analysis System" which is currently in development at the Imperial College. In this proposal workflows and future functionality will be accessible via the EP-EMAAS Portal. The question arises whether scientists should be forced to use an off site web-portal in order to run their analysis and how scientists outside of this project can change the implemented methods or add new ones them self. Still today new or improved methods for the analysis of microarray data are developed and researchers will probably want to know exactly what is done with their data or even modify the methods used or develop and test new methods them self. With the proposed approach of this diploma thesis scientists can eventually download the virtual server and run and enhance it them self at will. Ludäscher et al give a similar introduction to the disadvantages of form- and browser-based interfaces to computational tools in scientific research: '[...] the serious limitations of the "copy-paste-click" mode of end-users interacting directly with different web sites have now become apparent, and this error-prone and inefficient manual approach to data and process integration is not adequate for the goal of automated, high throughput scientific workflows.'[8]

The Taverna project itself is also working on ways to run R scripts directly from within the Taverna workbench. At this moment however 'the R processor [is] not considered stable, and is disabled by default'[9]. Also in order to run the R Processor you need yet another tool, the Rserv TCP/IP Interface[10]. When the development of the R Processor eventually reaches a stable release it may be very helpful for the integration of R-based webservices into Taverna workflows, but the distributed approach in general seems to be better suited for the analysis of large microarray data sets. This approach for instance would require the installation and administration of Rserve, R and Bioconductor on every single computer that is supposed to run a workflow with Taverna. Once again older computers will be at a serious disadvantage in terms of required time for computations. Additionally the ability to freely share workflows and scripts used within them is dimin-

ished because the developer will have a hard time creating R scripts that run on every computer due to different version of R and different or not the required libraries for it installed. Running those scripts on different operating systems (especially with different file systems) can also be problematic.

4 Procedure

The work on this thesis will start taking a close look into Taverna and scientific workflows in general and into ways to use R scripts as processors of Taverna (e.g. with Soaplab). Following that there will be a basic test to proof the concepts of accessing R through Taverna to ensure that the principal idea is valid. Then a set of core Bioconductor functions for microarray analysis will be implemented to explore the possibilities of Taverna to correctly chain them in a data and control flow (e.g. scheduling, branching, loops, user-interaction, access to results etc.). The workflow will then be subject to enhancement and refinement with the help of feedback and feature requests by the biologists of Dr. Kemmner's group at the Max-Delbrück-Center.

The initial focus is on Affymetrix GeneChips© and cDNA microarrays.There are many ways to transform and analyse microarray data. For instance there are at least four or five methods implemented in Bioconductor for Affymetrix microrays to normalize the data beside fundamental differences like choosing between EDA (Explorative Data Analysis) and CDA (Confirmatory Data Analysis). There is a lot of literature available in this regard which will be used to decide on default methods[11]. Another important aspect for workflows and Taverna is composability. It is the aim that processors of the workflow (e.g. the webservices) can be run outside of the developed workflow and that the workflow can be altered and will still execute as long as it still makes sense. Therefore the workflow needs to be structured in a way that it is possible to replace sub-workflows without destroying the functionality of the whole. More so when implementing the webservices there will be a focus on independence and composability so they can be run out of order and used by other scientists in possibly completely different applications. The most basic workflow will roughly contain the following sup-workflows:

- 1. setup/upload of data
- 2. generation of gene expression values (background correction, normalization, summarization...)
- 3. Quality Control (user interaction, eventually return to step 2)
- 4. finding differential expressed genes (statistical tests, heatmaps, cluster-analysis etc. as requested by the user)

After finding genes of interest a researcher can also access the KEGG (Kyoto Encyclo-

pedia of Genes and Genomes) database for instance to look for pathways in order to find proteins that are regulated by those genes (too high or too low amounts of a protein may be the cause of a disease). Finding and testing marker genes for cancer prognosis is also of importance so there are many possible additions to the basic workflow. Those will be added, refined and enhanced by user requests.

References

[1] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li. Taverna: A tool for the composition and enactment of bioinformatics workflows. Bioinformatics journal, 17(20):3045–3054, 2004.

[2] The R Project for Statistical Computing, http://www.r-project.org/, August 2006

[3] Mark Reimers and Vincent J. Carey, [8] Bioconductor: An Open Source Framework for Bioinformatics and Computational Biology. In: Alan Kimmel and Brian Oliver, Editor(s), Methods in Enzymology, Academic Press, 2006, Volume 411, DNA Microarrays, Part B: Databases and Statistics, Pages 119-134.

[4] S. Bowers and B. Ludäscher. Actor-Oriented Design of Scientific Workflows. In Proc. of the Intl. Conf. on Conceptual Modeling (ER), 2005.

[5] M. Senger, P. Rice, T. Oinn. Soaplab - a unified Sesame door to analysis tools. Proceedings of the UK e-Science All Hands Meeting 2003. http://industry.ebi.ac.uk/soaplab

[6] T. Oinn. http://twiki.mygrid.info/twiki/pub/Mygrid/TavernaRIntegration/EP-EMAAS-R-Taverna-Requirements.doc. 24th July 2006

[7] EBI, http://www.ebi.ac.uk/expressionprofiler/, August 2006

[8] B. Ludascher, I. Altintas, A. Gupta. Compiling Abstract Scientific Workflows into Web Service Workflows. Conference on Scientific and Statistical Database Management, 2003. 15th International p. 251- 254.

[9] The R Processor RC1, http://www.mygrid.org.uk/wiki/Mygrid/UsingRProcessor, August 2006

[10] Rserve a TCP/IP interface to R , http://stats.math.uni-augsburg.de/Rserve/, August 2006

[11] R. Irizarry, F. C. B. Hobbs, Y. Beaxer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4:249–264, 2003.