Grid-based expression QTL analysis

Ann-Kristin Grimm[a]*, Jan Kolbaum[a], Saleh Ibrahim[b], Steffen Möller[a]*
[a] University of Lübeck, Institute for Neuro- und Bioinformatics, [b] University Medical Center Schleswig-Holstein, Department of Dermatology, Ratzeburger Allee 160, 23538 Lübeck, Germany

Project URL: `http://eqtl.berlios.de/`
Code URL: `http://developer.berlios.de/git/?group_id=10090`
License: Affero GPL

Introduction

For finding new ways to treat diseases, an understanding of the molecular pathophysiology is extremely beneficial. For the analysis of complex diseases, the past years showed an increasing number of whole genome SNP association studies (GWAs). In animal models, microsatellites serve as genetic markers and combined with directed breeding of susceptible and resistant strains, disease-associated chromosomal loci (QTL) are determined.

For SNPs or QTLs, expression data of the genotyped samples may yield a functional characterisation of the genetic variation, allowing for the determination of expression QTL (eQTL). For QTL analyses, an increased density of genetic markers in regions with observed cross-overs and varying phenotypes may contribute to narrow down the disease causing regions. Later stages of a GWA may stimulate the sequencing of regions of interest. Data added demands a recomputation of previous analyses.

The computational effort for these computations and the amount of data are enormous. It needs to be organised and prepared for a computationally assisted interactive analysis with the researchers. The analyses demand many CPU years, which is not available to most research groups and is needed only between rounds of further wet-lab analyses, which should be as short as possible. Since the problem is data parallel when every gene is treated independently, there is barely a limit on the number of CPUs that can be employed for the computations.

Implementation

We present an approach that adopted the grid technology for the computations and established a workflow for the processing and presentation of the data. It uses the Advanced Resource Connector (ARC) grid middleware (M. Ellert *et al.*, 2007) to execute R scripts with the R/qtl library. Perl and MySQL organise the data, dynamic web pages with Apache/PHP allow for the presentation.

The SQL database is responsible for both, the representation of the computed data and the organisation of the jobs and data. The web interface to this data allows for browsing the computational results in an intuitive way and displaying additional biological information. This comprises local data on yet unpublished disease-associated regions and public data that is retrieved from Ensembl (T.J.P. Hubbard *et al.*, 2007) to faciliate the interpretation of the results.

Dependent on the computational results, new data is generated in the wet-lab and addtional computations are performed. Every result in the database has a reference to the compute task that yielded the finding. This way, refinements of previous results can be achieved incrementally: when there is demand, the internal task status table is indicating a recomputation that will substitute the earlier data.

The job submitted to the grid is a shell script, which starts an R script, which in turn request from a dynamic web page at a static URL a second R script. That second script directly corresponds to an internal compute task. It knows what data to retrieve and how to name the output files. When a time slot is used up or there are no more tasks to compute, the job ends. The results are continously downloaded from the grid and the database updated. New results lead to new information and trigger the production of more wet-lab data, a process to possibly be iterated multiple times.

Challenges

The grid technology helped establishing a direct interaction between wet-lab and computational analysis, i.e. the computed data can be supplied shortly after obtaining updates on the wet-lab data, and thus allows for refinements of the analyses. One remaining challenge is the complete automation of that process.

The current web interface performs analyses that are close to the data. The challenge here is to support the biological user more with machine learning techniques that attempt to model the disease, i.e. to provide access via pathways or functional classifications. One could use such generated hypothese to formulate more advanced analyses in statistical genetics and run these automatically.

*to whome correspondence should be adressed: {grimm, moeller}@inb.uni-luebeck.de

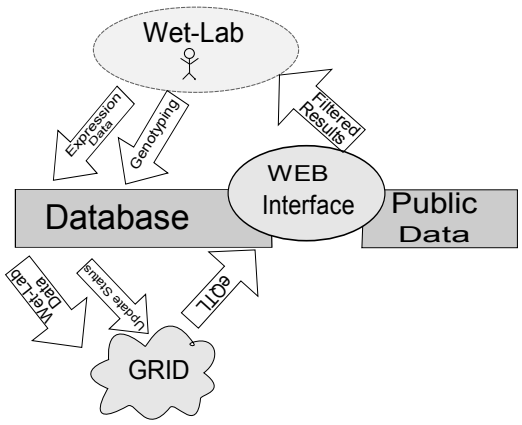

Figure 1: Graphical illustration of the data flow between wet-lab and computational analysis.

Acknowledgements

The authors thank Olli Tourunen and the NDGF BioGrid initiative for providing the ressources. Mélanie Thessén Hedreul and Maja Jagodic are thanked for testing and comments.

References

M. Ellert *et al.* (2007). Advanced resource connector moddleware for lightweight computational grids. *Future Gener. Comput. Syst.*, **23**, 219–617.

T.J.P. Hubbard *et al.* (2007). Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–617.