An Interface to Search Time-Oriented Functional Genomics Databases

Raveena Pentyam, Shahrzad Eslamian, Swati Gupta, Christopher Theisen, Guenter Tusch (PhD)
Medical and Bioinformatics Graduate Program, School of Computing and Information Systems,
Grand Valley State University, Allendale, MI, USA

Summary

We created a web based temporal database mediator using open-source platforms that supports the R statistical package, PHP, Bioconductor, and Web 2.0 knowledge representation standards using the open source Semantic Web tool Protégé-OWL.

Introduction

Development in microarray technologies have led to intricate datasets often to multiple biological questions. There are quite a few interfaces to search functional genomics repositories, but none supports pattern search across studies in time oriented databases. We extended our tool (SPOT) to help researchers find temporal patterns in large sets of microarray data across platforms. This can be the first step in a meta analysis by identifying appropriate data sets.

Exploratory analysis of microarray data often includes exploration of temporal data. For example, researchers might look for temporal changes in gene expression to determine responses to an injury. Different methodological approaches to model temporal data were explored to implement this approach by creating an open-source platform. It supports the R statistical package and knowledge representation standards (OWL, SWRL) using the open source Semantic Web tool Protégé-OWL.

Methodology

Our program (SPOT) helps the researcher find these patterns in large sets of microarray data. A researcher proceeds through three subsequent steps: first, selection of microarray data of interesting experiments from NCBI GEO, second, translating the temporal measurements into time intervals, and third, defining temporal concepts like “peaks” based on those intervals. Then he/she can search for genes that exhibit that particular pattern within the previously selected data pool.

We report here on the web interface that connects to programs based on R and Bioconductor.

The SPOT project is implemented mainly in free open source technology. The project is hosted on an Apache server (Fedora Linux). The project’s database is built on MySQL. This database is accessed by the interface to display the information that the users selects, also by the back end code which contains the logic of the application developed in R. The interface is developed in PHP, a scripting language, which makes it very easy executing R commands and queries on the MySQL database. The R-scripts performs many different tasks depending on the step of the project that is being executed (fig. 1). At the beginning the user selects the criteria of the experiments through a PHP form (fig. 2). PHP calls an R script that connects with the GEOMetaDB [4] SQLite database.

Then the user selects different algorithms or parameters in the system (fig. 3), that allow the system to train patterns to recognize, for instance, a peak in fold changes of the temporal expression data [2]. The program generates R code and OWL/SWRL rules (fig. 4). The researcher (user) can define different gene expression profile peaks, e.g., “Early” or “Late” in the time course, and search for similar profiles in the database of interesting studies. We developed tools to support this process using the Protégé-OWL ontology development toolkit (compare [3]). SWRL allows users to write rules that can be expressed in terms of OWL concepts and that can reason about OWL individuals. The Protégé OWL plug-in allows to easily build ontologies that are backed by OWL code.

User Interface and Evaluation

The R/Protégé interface is implemented as a Java program, which interfaces with R and Protégé through respective API’s. The user selects the different algorithms or parameters in the system via pull-down menus. The program then generates corresponding R macros and OWL/SWRL code.

The user interface was evaluated regarding functionality, robustness, user friendliness, and the number of clicks necessary to achieve one’s goal. The Spot interface has been in use for over a few years in the research group and has proven to be robust and easy to use. The core search results are obtained after only a few minutes. The evaluation was mainly performed by running sample gene data obtained from the Melterlab GEO site [6]. Multiple platforms were tested and corresponding GDS numbers were verified for all these platforms. In addition, a time series sample was tested, and the expected GDS number was confirmed. We also tested the search by organisms with Drosophila melanogaster. The query obtained all expected datasets.

As a next step, we need to establish standard test and validation criteria, and have more biologists to evaluate this application.

References:

Conclusion

The web interface allows the researcher to identify several platforms and for these platforms some or all NCBI GEO datasets based on that platform.

Finally, a set of genes from the intersection of all genes in the selected platforms can be identified for the search for pattern in all selected datasets.

We report experiences and challenges of using public open source databases. This tool shows promise to help researchers in the analysis of temporal gene expression data and presents a novel opportunity to identify new drug targets. It may be one potential step to evaluate drugs for their overall effects.

Acknowledgements:

We would like to thank the following individuals without whose support this study would not have been possible:

Dr. Amar Das (Bartmouth), Martin O'Connor (Stanford U)
Dr. Craig Webb, Dr. Jeremy Miller, (VA)
Dr. Timothy Redmond, Dr. Mark Musen (Stanford U)
Ramya Gunda and Jayashree Gaggapally (GVUS)

Figure 1: SPOT – Overview

Figure 2: SPOT – Architecture

Figure 3: The concept ‘Early Peak’