Data integration for analyzing gene expression data

Toralf Kirsten, Hong-Hai Do, Erhard Rahm

Various user groups at the University of Leipzig use microarray technology (mainly from Affymetrix Inc.) for gene expression analysis in their research and are producing several hundreds of experiment a year. To manage the large amounts of data resulting from these experiments, a comprehensive database solution is necessary, in particular to store raw microarray data as well as derived data and to support various analysis forms. Furthermore, annotations from different sources should be integrated with the microarray data to help the users in interpreting detected gene expression patterns. These integration and analysis tasks can be well served by a data warehouse approach to data management which we follow.

We first analysed the end users' requirements for the management and analysis of microarray data. We then comprehensively evaluated previously developed gene expression database approaches and assessed whether they are useful for our needs. To optimally serve the local requirements, we designed a data warehouse architecture, which is based on a specific multidimensional data model allowing the representation of gene expression data in different ways, in particular raw form and according to several normalization and aggregation methods. We designed and implemented a load management concept to load and transform the data from different sources, such as Affymetrix data files and public data sources, into the data warehouse. In addition, in close cooperation with the user groups of Prof. Horn (Medical Dept.) and Dr. Krohn (IZKF) we implemented several analysis reports for descriptive statistics. Our data warehouse implementation, *GeWare*, is completed in a first version for the commercial DBMS DB2 and is being migrated to the production platform on a large Unix server. All data access and analysis routines are accessible through a uniform web-based user interface.

Currently, our work focuses on how to incorporate and use annotation data in the process of gene expression analysis. Sample and experiment annotations (documenting the specific biological focus of an experiment and the technical process of conducting it) may be manually specified by the user or imported from existing databases, e.g. for patient and clinical data. We are evaluating the emerging MIAME standard and will adapt it to our specific requirements to capture necessary sample and experiment information. On the other hand, gene annotations (describing all known aspects about the sequences on the microarray chips) are to be exploited from the corresponding public sources. In cooperation with the user groups of Svante Pääbo (MPI-EVA) we have implemented a database-based tool called *GeneMapper* to identify the associations between accession numbers from different public sources. We will integrate GeneMapper with GeWare so that the associations between the identified genes of interest and annotations from any sources can be easily exploited.

Contact: **Toralf Kirsten, Hong-Hai Do**
University of Leipzig
Interdisciplinary Centre for Bioinformatics (IZBI)
{kirsten, do}@izbi.uni-leipzig.de, http://www.izbi.de
Erhard Rahm
University of Leipzig
Department of Computer Science
rahm@informatik.uni-leipzig.de, http://dbs.uni-leipzig.de