Data Integration for Analyzing Gene Expression Data

Toralf Kirsten, Hong-Hai Do, Erhard Rahm

Gene expression data is of essential importance for comparative investigations aiming at discovery of new genes, functional classification of genes, discovery of relationships between genes and their products, discovery of relationships between cells and environments etc. Among the various methods, such as EST Clustering, SAGE, developed for detecting and measuring gene expression, microarrays have become the predominant approach because they allow performing expression analysis on a very large scale, i.e. to measure and study the expression of thousands of genes simultaneously. To manage the large amounts of data constantly produced by the local users, a comprehensive database solution is necessary, in particular to store expression data together with all relevant annotations, and to support various analysis forms. These data integration and analysis tasks can be well served by a data warehouse approach to data management.

To assess the state of the art of current database solutions for gene expression analysis, we reviewed the available microarray databases described in recent scientific publications. To address the drawbacks of current solutions and to optimally serve the requirements of local molecular-biological research projects, we designed a data warehouse architecture and conceptual data model. The warehouse 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 different normalization and transformation methods. Data from different sources, i.e. Affymetrix data files and annotation data coming from local and public databases, are loaded into the data warehouse. In close cooperation with local user groups we implemented several analysis reports for descriptive statistics, which are accessible through a uniform web-based user interface. Currently, our work in this project focuses on flexible mechanisms for management of different kinds of annotation data, and advanced strategies for coupling analysis algorithms and tools with our data warehouse.