A Gene Expression Data Warehouse Platform

Toralf Kirsten, Hong-Hai Do, Erhard Rahm
University of Leipzig

Microarray experiments result in a large amount of expression data. This data should be managed together with all relevant annotations to support different kinds of comparative analysis. To better support such large-scale studies, we have designed and implemented a comprehensive analysis platform based on a central data warehouse called Gene Expression Warehouse (*GeWare*). *GeWare* centrally integrates and stores all relevant data, i.e. expression data and annotations.

In the talk we give an overview of the system. In particular, we present the experiment annotation module which allows the definition of controlled vocabularies and so-called annotation templates. Such templates comprise domain-specific categories which can be specified by the user for an experiment. Annotated experiments can be subsumed to experiment groups by using specified annotations filters. Furthermore, we describe the integration of gene annotations from publicly available data sources, such as LocusLink, Ensembl, GeneOntology, NetAffx and UniGene. These gene annotations can be used to extend expression analysis and to analyze gene annotations. Gene groups can be derived from the results of both, expression and annotation analysis.

GeWare supports various methods for pre-processing, statistical analysis, canned queries, and visualization. The pre-processing transforms Affymetrix raw data into gene expression values by proven methods, such as MAS5, RMA, Li/Wong etc. These expression values are input for built-in statistical analysis, e.g. different Westfall/Young tests, whose analysis results are also stored in the database. The underlying multidimensional data model of the data warehouse provides flexible analysis capabilities e.g. filtering and comparing genes by value, e.g. the p-value, which is calculated in the analysis. In addition, the defined experiment and gene groups can be used in queries and built-in analyses to focus on expression values for experiments and genes of interest.