# A Data Warehouse-based Gene Expression Analysis Platform

Toralf Kirsten, Hong-Hai Do, Erhard Rahm

Microarrays allow to measure gene activities in cells at a whole-genome scale, i.e. for thousands of genes at the same time. Each experiment generates large amounts of expression data, which should be managed together with all relevant annotations to support different kinds of comparative analyses. To better support such large-scale studies, we have designed and implemented a comprehensive 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 which is currently operational in a first version. In particular, we present the experiment annotation module which allows the user to define controlled vocabularies and so-called annotation templates. Such templates comprise domain-specific categories to whose annotation values are associated by the user in the annotation process. In joint work with local biologists different templates, e.g. human biopsy, human cell lines etc., have been established which are necessary for local user groups. Such annotated experiments can be subsumed to experiment groups by using specified annotations filters.

Furthermore, we have integrated several attributes, e.g. gene symbol, unigene, locuslink etc., from the external source *Netaffx*. These gene annotations are associated with each probeset available on Affymetrix microarrays and can be used for both, to extend report results for a better interpretation and to define gene groups by specifying various filter conditions using these annotation attributes and values. All stored experiment and gene groups can be reused in several statistical reports and built-in analyses.

In the third part, we discuss how the system, in particular the database, handles different types of gene expression matrices which are derived from available expression values. Such matrices can simply be exported for analysis using external tools, such as *GeneSpring* and *GenMapp*. On the other side, the gene expression matrices are input for advanced built-in analyses like the *Westphal/Young univariate test using resampling strategy*. The analysis results are stored in a so-called *data mart* as part of the data warehouse. The underlying multidimensional data model of the data mart provides flexible analysis capabilities e.g. filtering and comparing genes by value, e.g. the p-value, which is calculated in the analysis, and comparing different analysis results like the top ten genes of each specified analysis. In addition, the data model enables to report analysis results by filtering experiment and gene annotations.