A Data Warehouse for Multidimensional Gene Expression Analysis

Toralf Kirsten[†], Hong-Hai Do[†], Erhard Rahm[†‡]
[†]Interdisciplinary Centre for Bioinformatics, [‡]Department of Computer Science
University of Leipzig
www.izbi.de, dbs.uni-leipzig.de

Abstract: We introduce the *GeWare* data warehouse approach for microarray-based gene expression analysis. *GeWare* centrally stores raw and preprocessed expression data together with a variety of annotations to support different analysis forms. The flexibility of *GeWare* is made possible by a multidimensional data model where expression data is stored in several fact tables which are associated with multiple hierarchical dimensions holding describing metadata on genes, samples, experiments, and processing methods. Annotations are uniformly managed in a generic way thus supporting the high variability of annotations and easy extensibility. Gene annotations are imported from several public sources and integrated with each other based on commonly used accession numbers, e.g. UniGene clusters.

1 Introduction

Microarrays measure the expression of thousands of genes simultaneously so that huge amounts of data are produced with every experiment. To effectively support gene expression studies, a comprehensive database solution is necessary to manage the expression data of many experiments together with all relevant annotations. Previous database efforts for managing gene expression data which are evaluated in [2] still show several limitations. First, gene annotations are either ignored or only integrated using web links. Second, sample and experiment annotations are mostly captured as free texts. Third, expression analysis is mostly performed by standalone software tools outside the database system typically without involving all relevant annotations. On the other side, the solutions provided by microarray vendors such as Affymetrix are restricted to their specific algorithms, e.g. for normalization and analysis, which do not necessarily reflect the current state of research [5].

To overcome the limitations of previous approaches and to support gene expression analysis for research projects in Leipzig, we have designed and implemented an integrated database platform called Gene Expression Warehouse (*GeWare*). *GeWare* follows the data warehousing approach [4] to centrally integrate and store all relevant data, i.e. expression data and annotations. A data warehouse promises significant advantages because all relevant data is directly accessible for analysis, allowing for both good performance and extensive analysis capabilities. We designed and implemented a multidimensional data model compromising expression data in raw and preprocessed form and describing metadata on genes, samples, experiments, and processing methods. Furthermore, annotations are managed in a generic way thus supporting the high variability of annotations and easy extensibility. Gene annotations are imported from several public sources and integrated with each other based on commonly used accession numbers, e.g. UniGene clusters.

In the next two sections, we give an overview of the *GeWare* architecture and the multidimensional data model. Finally, we discuss some future work.

2 GeWare System Architecture

Data is imported from several sources, i.e. local and external, file-based and database sources. Like in business data warehouses imported data first has to be cleaned and transformed before it can be integrated. The intermediate results of these preprocessing steps are stored in a dedicated *staging area*. Several methods for normalization (e.g. global mean, normalization by a factor) and aggregation (e.g., Li/Wong [5]) are supported to process probe-level expression data since there are not yet generally accepted approaches for these tasks. By storing the preprocessing results for the same raw data we can use *GeWare* as a platform to evaluate and compare different normalization and aggregation methods.

The central component of *GeWare*, the data warehouse, is organized according to the multidimensional data model described in Section 3. For its implementation we use the relational database management system DB2 of IBM (on a high-end Unix server) supporting very high data volumes and a multitude of performance tuning options such as indexing, materialized views, data partitioning, etc. Specific portions of the warehouse can be redundantly stored in data marts to improve performance for special analysis tasks. Moreover, gene clusters identified during cluster analysis can be saved in data marts so that they can quickly be reused and visualized without recalculation.

All administration and analysis functions of *GeWare* are accessible via web interfaces which have been implemented using the Java Servlet technology. *GeWare* functionality is thus accessible through browsers, e.g. user / group administration, data logistics (data import and export, job management), data preprocessing (normalization and aggregation), and data analysis (predefined reports, queries, statistical analysis functions). Moreover, selected data can be extracted from the warehouse in file format required by independent analysis tools.

3 Data Warehouse Model

Figure 1 shows the *GeWare* data model following the multidimensional modeling methodology. The model is built of *dimension* and *fact* entity types which are implemented by dimension and fact tables, respectively. Facts are numeric and additive data, while dimensions provide information on the meaning of facts or how they have been determined.

Currently, our schema includes two main fact tables, *Probe Intensity* and *Gene Intensity,* representing expression intensity values at the oligo (probe) and gene (probe set) level, respectively. The probe fact table holds both raw expression intensities as well as the results after applying a normalization method. The gene fact table is used to store the probe set intensities imported from *Affymetrix MicroDB* and the results of other applied aggregation methods. Additional fact tables are kept for data marts, e.g. *Cluster Genes,* to store the intensities of those genes participating in clusters determined by specific cluster analyses.

The dimensions can be grouped into annotation- and processing-related dimensions.

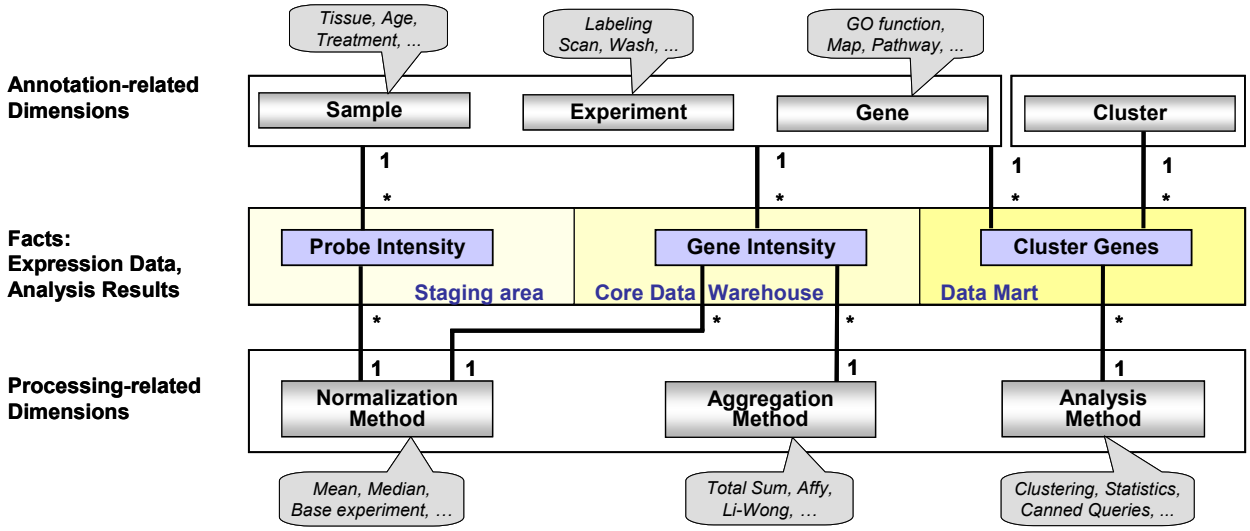

Figure 1. High-level data warehouse schema

Processing-related dimensions describe the computational methods and their parameters used to compute probe intensities, gene intensities and gene clusters, respectively. Annotation-related dimensions provide valuable describing metadata for the intensity values of the fact tables. Sample and experiment annotations are provided by user input while gene annotations are vendor-provided or imported from public sources. In particular, the gene dimension holds information on all probe sets together with annotations integrated from sources such as the GeneOntology. In contrast to processing-related dimensions and facts which are

of static character, annotation-related dimensions represent a highly variable part of the data warehouse model. Annotation data exhibits a high degree of complexity and heterogeneity since the biological focus of experiments, the relevant annotation sources and vocabularies are frequently changing. This makes it impossible to use a fixed schema structure for annotations but necessitates a generic representation to uniformly manage different kinds of annotations. For this purpose, we developed a flexible approach Generic Annotation Model (GAM) to represent and integrate heterogeneous annotation data which is described in detail in [3].

Typically, dimensions are organized in generalization/specialization hierarchies thus providing different levels of abstraction for analysis. For example, the gene dimension may be organized according to the function taxonomy of GeneOntology so that we can analyze intensity values at any functional level. Furthermore, OLAP-like navigations known from business data warehouses [4] can be used to drill-down or roll-up along an annotation hierarchy to increase or decrease the level of detail for analysis.

The sketched multidimensional data model supports a tremendous flexibility for gene expression analysis. While current approaches typically evaluate a complete gene expression matrix containing the intensity values for all measured genes and several/all experiments, we now can focus an individual or comparative analysis to an arbitrary subset of intensity values determined by specific annotation values of interest. The selection may be based on a value at a specific level of a single dimension or any combination for several dimensions (e.g. compare aggregation methods A and B for experiment series X and genes at level Y).

The data model is easily extensible. Within each dimension new processing methods or annotations can be added without affecting the existing data organization. New data marts and fact tables can be added and associated to the existing or new dimensions. The number of entries in the fact tables is virtually unlimited. While we have not yet performed a detailed performance analysis, we expect that DB2's performance options tailored to data warehouses allow fast processing times for most analysis queries.

4 Outlook

The *GeWare* data warehouse system is operational in a first version and holds already the expression data of many microarray experiments. Its contents can be analyzed using our own functions and interfaces or by external tools provided with data files exported from *GeWare*. We are in the process of integrating additional analysis and data mining functions to fully exploit the potential of our warehouse approach. Capturing sample and experiment annotations needs further improvement. The available tool *MIAMExpress* still makes extensive use of free-text fields and provides only a static structure of predefined annotation categories. We are working on a GAM-based solution, which allows the flexible definition of annotation categories and their corresponding vocabularies, and supports automatic generation of interfaces for user specification.

References

1. Brazma, A., et al: Minimum Information about a Microarray Experiment (MIAME) – Toward Standards for Microarray Data. Nature Genetics 19, 2001
2. Do, H.H., Kirsten, T., Rahm, E.: Comparative Evaluation of Microarray-based Gene Expression Databases. Proc. 10th Conf. Database Systems for Business, Technology and Web (BTW), 2003
3. Do, H.H., Rahm, E.: Generic Management and Integration of Molecular-biological Annotation Data. Technical Report, University of Leipzig, July 2003
4. Jarke, M. et al. (eds): Fundamentals of Data Warehouses, Springer-Verlag, 2nd ed., 2003
5. Li, C., Wong, W.H.: Model-based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection, *Proc. Natl. Acad. Sci.* Vol. 98, 31-36, 2001