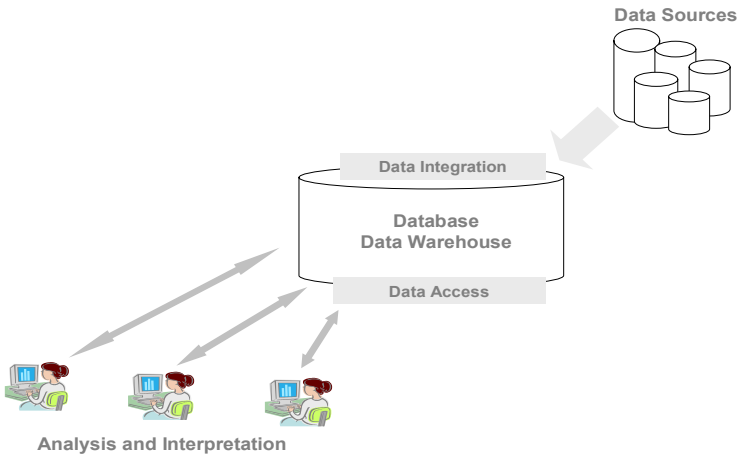
UNIVERSITÄT LEIPZIG

Leipzig Bioinformatics Working Paper
No. 1
November 2004

A Data Warehouse for Multidimensional Gene Expression Analysis

Toralf Kirsten, Hong-Hai Do, Erhard Rahm



published by the
Interdisciplinary Centre for
Bioinformatics

www.izbi.de/working_papers.html

ISSN 1860-2746

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
{tkirsten, do}@izbi.uni-leipzig.de,
rahm@informatik.uni-leipzig.de

Technical Report, November 2004

Abstract

We introduce the *GeWare* data warehouse system for microarray-based gene expression analysis. *GeWare* centrally stores expression data together with a variety of annotations to support different analysis forms. Compared to previous work, our approach is unique in several aspects. First, *GeWare* offers high flexibility with a multidimensional data model where expression data is stored in several fact tables which are associated with multiple hierarchical dimensions holding describing annotations on genes, samples, experiments, and processing methods. Second, all annotations are integrated and managed in a generic way, thus supporting easy evolution and extensibility. Especially, consistent experiment annotation is achieved by means of pre-defined annotation templates and controlled vocabularies. Finally, various analysis methods have been integrated using a flexible framework based on the exchange of experiment groups, gene groups and expression matrices. The system is fully operational and has been employed in several research projects.

1 Introduction

Microarrays measure the expression of thousands of genes simultaneously, producing huge amounts of data. 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 include ArrayDB[9], GeneX[16], M-CHIPS[10], RAD2[21], and SMD[20]. As evaluated by Do et al.[7], several limitations can still be observed in current solutions. First, gene annotations are either ignored or only integrated by web links thus preventing an automated analysis for subgroups of genes of interest. Second, sample and experiment annotations are mostly captured as free texts, whose heterogeneity essentially

*Corresponding postal address: University of Leipzig, Interdisciplinary Centre for Bioinformatics Leipzig, Kreuzstrasse 7b, Leipzig, 04103, Germany

complicates experiment comparisons and cross-experiment analysis. Third, expression analysis is mostly performed by stand-alone software tools outside the database system typically without involving all relevant annotations. The commercial solutions, such as LIMS by Affymetrix and GeneExpress[17] by GeneLogic are typically restricted to proprietary algorithms, e.g. for normalization and analysis, which do not necessarily reflect the current state of research[14].

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 platform called Gene Expression Warehouse (*GeWare*). The key aspects of our approach are the following:

- *GeWare* follows the data warehouse approach[13] to centrally integrate and store all relevant data, i.e. expression and annotation data. 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 have developed a multidimensional data model where expression data is stored in raw and preprocessed form and annotations on genes, samples, experiments, and processing methods are represented within multiple hierarchical dimensions. An advantage of the data model is the flexible support for focused analysis, for example to only consider experiments involving a given tissue and/or genes with a particular function by filtering the corresponding dimensions of expression data. Furthermore, the data model is easy to extend with new dimensions, e.g. to cover new annotations.

- We use a generic intermediate format to uniformly manage all relevant annotations, i.e. gene annotations imported from external sources and experiment annotations captured from user inputs. The generic format makes it easy to integrate new annotation sources and is robust against changes in external sources. From this format, the annotation-related dimensions in the multidimensional data model are dynamically derived for analysis.

- We enforce consistent experiment annotation by means of pre-defined *annotation templates* and *controlled vocabularies*. Both are also managed in the generic intermediate format, and thus can be easily adapted to new research requirements.

- *GeWare* provides different algorithms for preprocessing and analyzing expression data, e.g. to identify lists of interesting genes. The analysis methods are coupled in a simple, yet powerful way of exchanging *experiment groups*, *gene groups* and *gene expression matrices*. In particular, it not only supports the specification of complex automatic analysis workflows but also allows the user to flexibly continue analysis with pre-computed results.

The *GeWare* system is fully operational and has been employed in several research projects in Leipzig, which study, for example, the role of the transcription factor IL-6 on the survival of myeloma cells, the specific gene expression patterns in different kinds of thyroid nodules, and the factors influencing

the binding behavior of sequences on microarrays. Currently, *GeWare* manages data for more than 300 experiments. Interactive access is available under http://www.izbi.uni-leipzig.de/izbi/AG1/GeWare.html.

In the next section we give an overview of the *GeWare* architecture and the supported workflows. Section 3 describes the multidimensional data model. In Section 4, we first present our generic approach to manage annotation data and then discuss the mechanisms to enforce uniform experiment annotation and to integrate gene annotations from public sources. Section 5 illustrates the different analysis capabilities supported by *GeWare*. Section 6 concludes the paper.

2 System Overview

2.1 Architecture

Figure 1 shows the overall architecture of *GeWare*. Data is imported from several sources and transformed within a so-called staging area before it is integrated and stored in the central warehouse database for analysis. So-called data marts[13] support special analysis needs. They are either derived from the warehouse or are used to store the results of a particular analysis method for later reuse. All administration and analysis functions of *GeWare* are accessible via web interfaces.

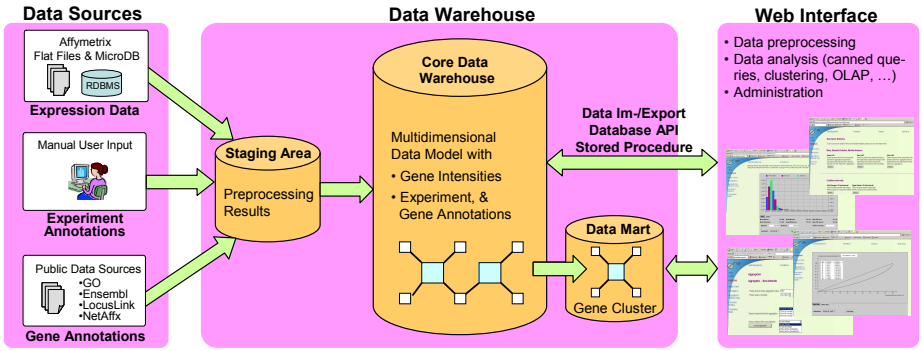


Figure 1: Overall architecture of GeWare

The following kinds of data are integrated in the warehouse:

- *Expression data*: Influenced by local requirements, we currently focus on expression data produced by Affymetrix microarrays. These oligonucleotide arrays use short sequences of 25 base pairs as *probes*. Each gene is represented by a so-called *probeset* consisting of 11-20 probes. A chip measures the expression level for thousands of such probes at the same time from which the corresponding probeset intensities are derived. *GeWare* has import interfaces for probe-level and probeset-level intensity values. Probe intensities are parsed and imported from so-called CEL files, while probeset intensities are from the Affymetrix MicroDB tool. Obviously, probe intensities are much more voluminous but allow us to apply different normalization methods to clean the data as well as different aggregation methods to combine the intensities of the probes to probeset intensities.

When importing expression data from the MicroDB we obtain the intensity values according to the Affymetrix normalization and aggregation methods.

- *Experiment annotations*: These annotations have to be specified by the user via a web interface together with the import of new expression data. According to the annotation template specified for the experiment, *GeWare* automatically generates corresponding web-pages, which allow the user to provide text inputs or to select possible values from controlled vocabularies.

- *Gene annotations*: We support parsing and import functions for gene annotations from the NetAffx[15] database of Affymetrix and other public sources, in particular GeneOntology[11], Ensembl[4] and LocusLink[19]. To keep these annotations up-to-date, a periodic refresh is performed from the corresponding sources.

Imported data first has to be cleaned and transformed for integration. The intermediate results of these preprocessing steps are stored in a dedicated staging area. For expression data, several methods for normalization and aggregation are supported to preprocess probe-level expression data, since there are not yet generally accepted approaches for these tasks. Experiment annotations specified by the users and gene annotations imported from public sources are also first stored in the staging area using a generic format (see Section 4). Materialized views are then dynamically generated from this format to populate corresponding annotation-related dimension tables in the data warehouse.

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 Linux 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. For example, expression matrices can be extracted for relevant gene groups and experiment groups and saved in corresponding data marts so that they can quickly be reused and visualized without recalculation.

The web interface of *GeWare* is implemented using the Java Servlet technology to perform the resource-intensive analysis tasks centrally on the server. All functionalities, including user administration, data import and export, annotation management, and expression analysis, are accessible through web browsers. Data access in *GeWare* is authenticated through a sophisticated concept of user/user group and right management. In particular, access rights can be granted/revoked not only for the data (expression and annotation data), but also for the functions on the data, such as import, export, query etc. According to the user profile, the web interface is automatically generated to only cover the allowed functions.

2.2 Workflow

Figure 2a and b show the common workflow supported by *GeWare* for data import and expression analysis, respectively. To import new data, an experiment

first needs to be created in the data warehouse. Experiments serve as container objects to hold all data generated from single microarray chips, i.e. probe intensities, probeset/gene intensities, as well as relevant experiment annotations. Expression data can be imported in both raw or preprocessed form. *GeWare* also supports batch import of expression data for many experiments at the same time. Raw expression data further has to be normalized and aggregated using an integrated method. Independently from the import process, the experiment can be fully described by the experimenter using a selected annotation template.

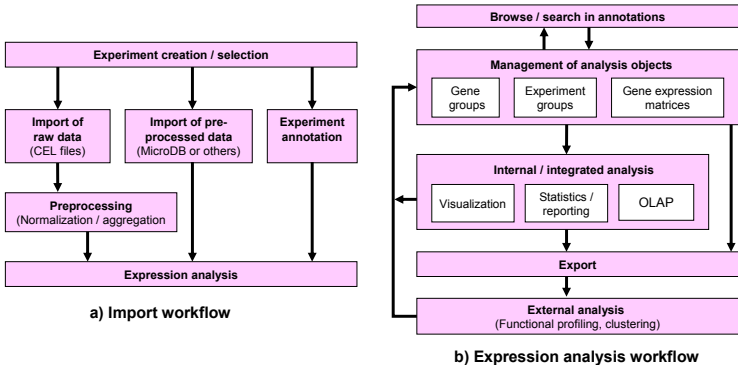


Figure 2: Operational workflow in GeWare

GeWare supports several forms of data analysis, such as visualization, predefined reports implementing statistical algorithms, and Online Analytical Processing (OLAP)[13]. Flexible combination and integration of the different methods is achieved by means of the uniform exchange of gene groups, experiment groups, and gene expression matrices, which are centrally created and managed by *GeWare*. Groups of interesting experiments and genes may be formed by manual user specification, by querying available annotations, or using some analysis algorithms. Focused analysis is possible with expression matrices generated for the relevant experiment and gene groups. The results, typically sub-groups of relevant genes, can in turn be saved for further queries and analysis. *GeWare* also supports the export of the pre-computed analysis results, e.g. gene groups and gene expression matrices, to perform analysis in external tools, such as for functional profiling and clustering.

3 Data Warehouse Model

Figure 3 shows a high-level view on the multidimensional data warehouse schema of *GeWare*. It is built of *dimension* and *fact* tables. 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 probe and gene (probeset) 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 probeset intensities imported from Affymetrix MicroDB and the results of other aggregation methods. Additional fact tables

are kept for data marts, e.g. *Expression Matrices*, to store the intensities of those genes participating in gene groups determined by a specific analysis method, such as clustering.

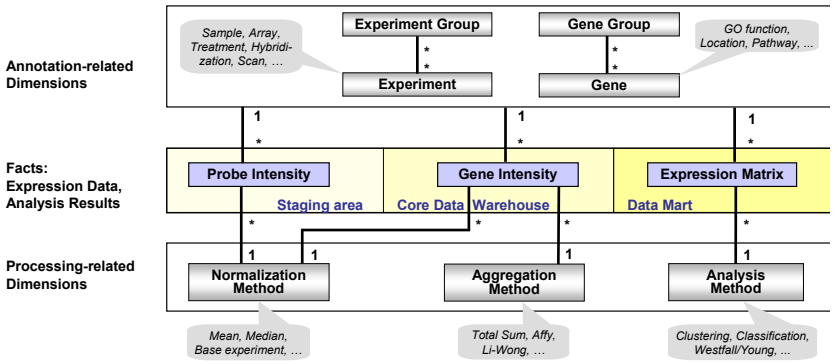


Figure 3: High-level data warehouse schema

The dimensions can be grouped into annotation- and processing-related dimensions, which are shown in Figure 3 together with some illustrating examples. Annotation-related dimensions include *Experiment*, *Gene*, and their groupings, *Experiment Group* and *Gene Group*, respectively. The experiment dimension holds the user-specified conditions of the experiments, while the gene dimension provides the facts currently known about the genes, such as their functions. Processing-related dimensions include *Normalization Method*, *Aggregation Method* and *Analysis Method* describing the computational methods and their parameters used to compute probe intensities, gene intensities, and to determine gene groups for expression matrices, respectively.

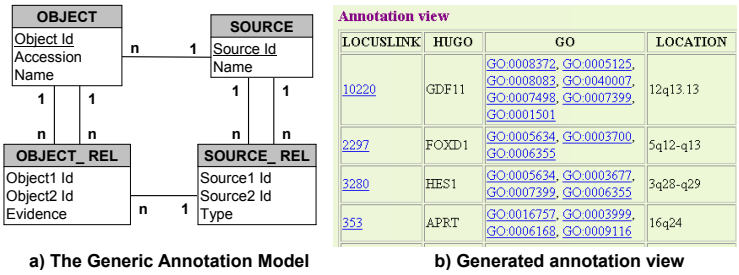
Typically, dimensions are organized in generalization / specialization hierarchies thus providing different levels of abstraction for analysis. For example, the gene dimension may be organized according to the function taxonomy of GeneOntology so that we can analyze intensity values at any functional level. Furthermore, OLAP-like navigations known from business data warehouses[13] 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 on 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 different aggregation methods for a given experiment group and a gene group with a particular GeneOntology function). Moreover, 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 also be added and associated to the existing or new dimensions.

4 Annotation Management and Integration

4.1 Annotation management

While the fact tables and processing-related dimensions are of a rather static structure, the 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 impractical to use a fixed schema structure for annotations but necessitates a generic representation to uniformly manage different kinds of annotations.



a) The Generic Annotation Model b) Generated annotation view

Figure 4: GAM and example of an integrated annotation view

Figure 4a presents our (simplified) generic representation for annotations, the Generic Annotation Model (GAM)[8], which is implemented in the staging area to first capture all annotation data. In order to avoid discrepancies in free-text annotations, we allow annotations to be split into predefined *objects*, i.e. items with clearly defined semantics. In this generic representation, an object can denote a category in an annotation template, a term in a controlled vocabulary, or an entry in a public source etc. According to their origin, objects are grouped into single so-called *sources*, such as the corresponding annotation templates, controlled vocabularies, and public sources. Furthermore, GAM is able to capture and manage different kinds of relationships between the objects as well as between the sources (*Object_rel* and *Source_rel*). A relationship between two sources comprises of relationships between their objects. Relationships can hold between different sources but also within a single source, for example to model the hierarchical structure of an annotation template or the GeneOntology (GO) taxonomy.

GAM makes it easy to add a new source and integrate it with existing ones by specifying corresponding source and object relationships. An existing source can also be quickly altered by removing irrelevant objects and adding new ones. However, the trade-off for this flexibility and robustness is the complexity of queries when directly accessing GAM. Hence, applications and users are typically provided with views tailored to their analysis needs. In particular, the annotation-based dimensions in the multidimensional data model (Section 2) are dynamically derived from GAM as materialized views. Here, we employ the same method implemented and described in Do et al[8]. In particular, by navigating along the relationships from the source to be annotated to the source providing annotations, we can obtain the required annotations for the objects of interest. Figure 4b shows an example of such a generated annotation view

for some LocusLink genes.

4.2 Specification of experiment annotations

Depending on the biological focus, microarray experiments can be conducted and documented in different ways. For example, annotating the time points in a time-series experiment is not necessary for an experiment comparing normal and diseased tissues. Such different experiment designs make it difficult to achieve a single unique schema for experiment annotation. Addressing this problem, the MIAME standard[6] gives a recommendation about the minimal information to be captured about a microarray experiment. While specifying the categories to be filled in, such as for array design, sample description, hybridization procedure, etc., MIAME leaves open how the values are to be filled in for those categories. This can easily lead to conflicting values, such as "Homo sapiens" and "Human" for an *Organism* category.

Therefore, to achieve consistent experiment annotation we support so-called *annotation templates* to prescribe the categories to be annotated and *controlled vocabularies* to constrain the values for the categories. An annotation template is typically hierarchically organized and consists of *pages*, which group related *categories* together, such as for array design, sample description etc. The definition of an annotation template consists of constructing such pages, specifying relationships between the pages, and eventually, defining categories for the single pages. A category is either of type free text or, preferably, controlled. In the latter case, an existing controlled vocabulary has to be chosen to provide corresponding choices for user input. It is possible to copy and modify an existing template to speed up the construction of a new, similar template. Currently, a template consisting of MIAME categories is provided as a basis to construct new project-specific templates. *GeWare* supports user-defined controlled vocabularies as well as existing standard ontologies, such as NCBI taxonomy. Once defined or imported, a controlled vocabulary can be shared by different templates.

From the definition of the annotation template, *GeWare* automatically generates web pages for user input. To annotate an experiment, the user simply walks through the generated pages to provide values for the corresponding categories. If a similar experiment has already been annotated, it is also possible to copy all its annotations into the new experiment for modification. Figure 5a shows the index of all pages defined in an existing template, *Human Cell Culture*. A page is displayed in Figure 5b and contains categories to capture culture conditions of an in-vivo experiment. The values for vocabulary-based categories can be chosen from corresponding select and check boxes.

By using a single template, experiments can be uniformly annotated, making it easy to identify related experiments by searching in their annotations. As indicated in Figure 5c, the terms from controlled vocabularies are shown for the corresponding categories so that the user can specify search criteria. The criteria are combined using the logical operators AND, OR, and NOT. The identified experiments can then be saved as an experiment group for later analysis.

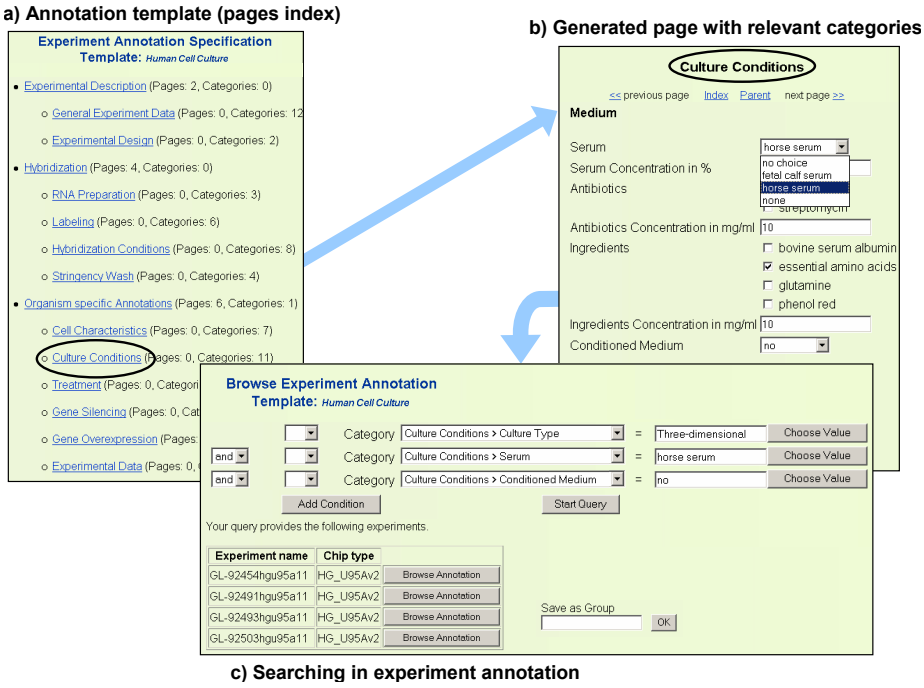


Figure 5: Capturing experiment annotations

4.3 Integration of gene annotations

Currently, we integrate gene annotations from three public sources LocusLink, GeneOntology (GO), Ensembl, and the vendor-based NetAffx. Typically, weblinks are provided in one source to refer to annotations offered by another source. We explicitly capture these semantic correspondences and store them in the GAM format so that relationships between objects and their annotations can be flexibly queried and combined[8].

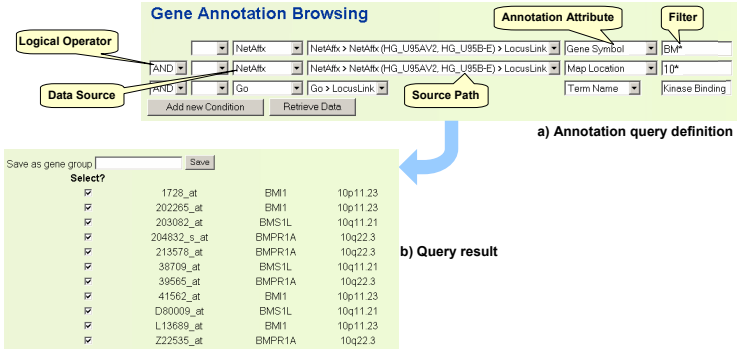


Figure 6: Querying integrated gene annotations

In *GeWare*, gene annotations can be queried to identify groups of interesting genes. Conversely, given a gene group the user can query the associated annotations. Figure 6 shows an example of a search in gene annotations. A

query can involve multiple criteria, which are combined by the logical operators AND, OR, and NOT. For each criterion, the user has to select the source to be annotated, the path leading to the source providing annotations, the required annotation attributes, and the filter conditions on the attributes. The query in Figure 6 returns a list of Affymetrix probesets, whose gene symbol starts with "BM", are located on the tenth chromosome and have the GO function "Kinase Binding". All annotations, i.e. gene symbol, chromosome, and GO, are provided by LocusLink.

It is already known that Affymetrix probesets do not necessarily represent unique genes, and multiple probesets may represent a single known gene. Hence, in order to avoid duplicated annotations, the probesets should first be mapped to a generally accepted gene representation, such as LocusLink. The obtained unique genes can then be used to search for annotations. *GeWare* supports both tasks by exploiting the captured relationships. For example, we have integrated all associations between Affymetrix and LocusLink, and between LocusLink and GeneOntology. Thus, we are able to map Affymetrix probesets to LocusLink genes and retrieve unique GO functions for the relevant genes. Moreover, the analysis of other gene representations, such as Unigene[23], is also possible. In particular, we can map Affymetrix probesets to Unigene clusters and correlate Unigene clusters with GO entries, if they are related to the same LocusLink entries. This transitive combination of relationships represents a simple, yet powerful way to gain further annotation knowledge for gene expression analysis.

5 Expression Analysis

In the following, we present a selection of the different analysis methods supported by *GeWare*. We first focus on the preprocessing of raw data. We then discuss the routines for reporting and statistical analysis and present a first application of OLAP to expression data.

5.1 Preprocessing

Preprocessing consists of two main tasks, 1) normalization to clean noise signals from raw probe intensities and to make the probe intensities from different chips comparable, and 2) aggregation to produce gene/probeset intensities from normalized probe intensities. To save implementation effort, *GeWare* integrated the open-source BioConductor package, and thus has access to a library of different methods commonly used in each preprocessing step. *GeWare* allows to tailor a preprocessing strategy by combining different methods for the single steps. In addition, BioConductor also offers more sophisticated algorithms covering both steps. Such algorithms, including Speed's RMA[12], Li/Wong's model[14] and Affymetrix MAS5[1], can also be specified in *GeWare*.

On the other side, we observe that the determination of the intensity values from a hybridized chip is subject to many fluctuations. Apart from the external influences caused by the experimental procedure, the probe sequences with specific distributions of nucleotides may cause a bias to the binding behavior during hybridization. Hence, we are currently investigating these aspects in one project with the major goal to derive more accurate gene intensities and to improve the chip design with better probe sequences. *GeWare* supports this task

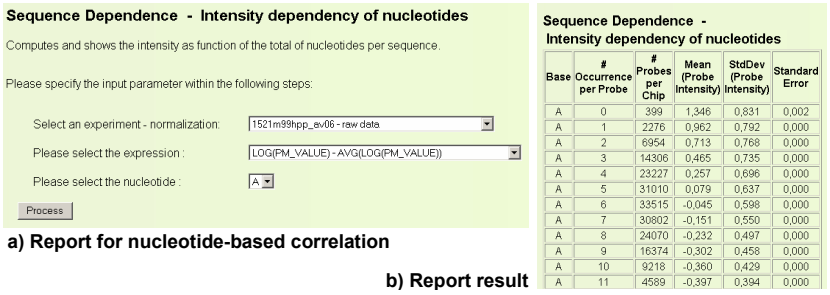


Figure 7: Report for correlation between nucleotides and intensities

by different reports showing correlations between probe sequences and obtained probe intensities. Figure 7 shows one report examining the relation between the occurrences of a nucleotide (e.g. A) and the distribution of probe intensities. In particular, it groups the probes with a particular number of occurrences of the corresponding nucleotide and computes the mean, standard deviation and standard error of intensity value for single probe groups. The detailed description of the analysis procedure and current results can be found in Binder et al[2, 3].

5.2 Reporting and statistical analysis

GeWare provides various functions for the interactive analysis of gene/probeset intensity values. These include visualization methods and reports to detect groups of genes with specific expression patterns.

GeWare supports several kinds of charts for visualizing expression data. Figure 8a shows a heat map and a line chart for gene intensities in an expression matrix. These charts help to quickly find (groups of) genes with different or similar expression patterns across different experiments. In addition to the charts, the distribution and concentration of gene intensities from a chip can also be examined by means of Lorenz curves.

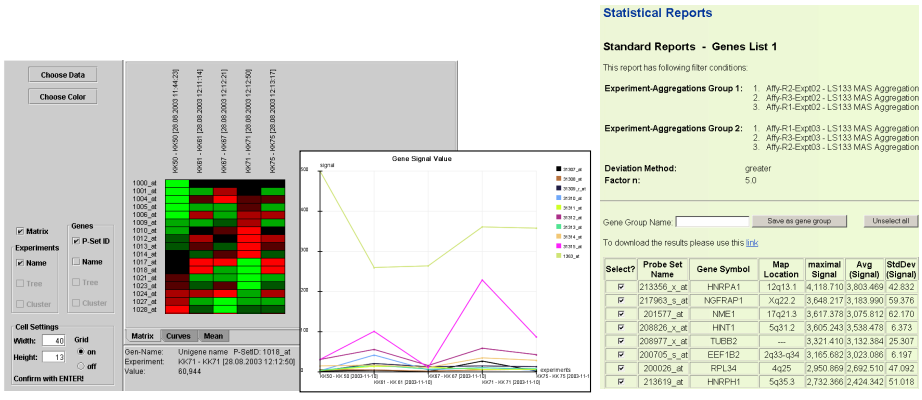


Figure 8: Interactive expression analysis

GeWare provides several built-in statistical approaches to identify interest-

ing genes from expression data. Here, we focus on outlier genes and differentially expressed genes, which typically are the target of expression analysis. Outliers are genes showing extreme intensity values in one experiment or experiment group compared to another experiment or experiment group. *GeWare* supports different statistical criteria for outlier detection, e.g. based on standard deviation, in corresponding reports. Figure 8b shows the result of such a report: outlier genes detected for two experiment groups are displayed along with the selected annotations, i.e. gene symbol and map location. Similarly, the detection of differently expressed genes has also been realized in several reports. They are based on more complex statistical tests following the Westfall/Young approach[22].

Similar to searching in gene annotations, the reports produce groups of interesting genes. The gene groups can be used as input in further analysis steps. Hence, iterative analysis is possible by continuing with pre-computed results, e.g. to successively filter the genes of interest. Furthermore, gene groups and expression matrices can be exported for analysis in stand-alone tools. While expression matrices are the typical input format of existing clustering tools, gene groups can be used to perform functional profiling in tools such as FUNC[18] and GenMapper[8].

5.3 Applying OLAP to expression data

In business warehouses, **O**n**L**ine **A**nalytical **P**rocessing (OLAP) is a common approach to analyze multidimensional data[13]. The user can interactively navigate across different dimension levels and aggregate facts to find interesting patterns in data. For example, the sale trend of a product can be summarized by determining the max, min, average, or sum of the sales of the product over a series of time periods, such as weeks or months.

Expression data is also highly multidimensional. In particular, annotations on genes, experiments, and processing/analysis methods, provide many interesting aspects, according to which gene intensities can be filtered and summarized. Figure 9 shows a screenshot of our application using an commercial OLAP tool. On the left, the hierarchies of dimensions as well as the available facts are shown which can be flexibly chosen for the analysis shown on the right hand side. In particular, two dimension hierarchies Experiment: ChipType → HG-U95Av2 → Experiment and Gene: Probeset Name, and the fact attribute Detection P-value are chosen. Furthermore, a pre-defined filter is applied to identify the top-ten genes ranked according to its Signal values. The tool then visualizes the distribution of Detection P-values of the filtered probesets in the corresponding experiments as bar charts and also shows the values in tabular form.

We believe OLAP is a promising analysis approach for expression data. It optimally supports the ad-hoc nature of interactive analysis and can be very helpful to quickly get an impression about the data and to identify a subset of interesting genes for closer examination. Furthermore, OLAP is already a matured technology with many supporting commercial tools.

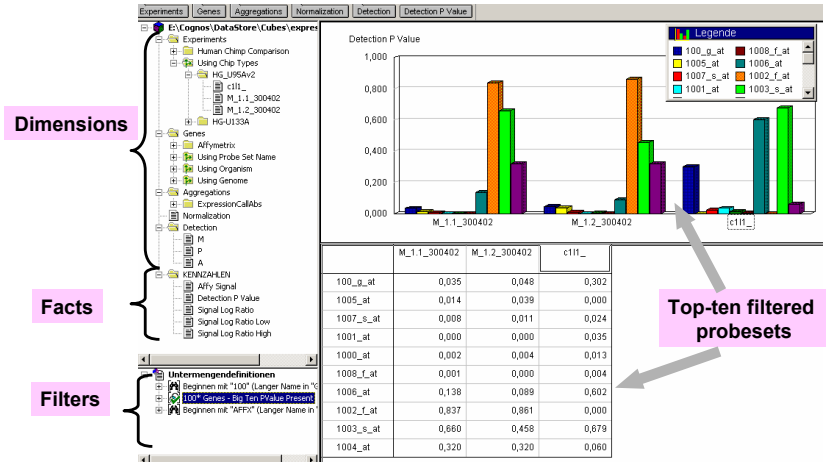


Figure 9: OLAP using expression values

6 Conclusions

We presented the *GeWare* system as an integrated platform for gene expression analysis. Compared to previous work, our approach exhibits a number of innovations. *GeWare* stores all relevant data, i.e. expression data and annotations, in a multidimensional data model. Annotations are integrated and managed in a generic way, thus supporting easy evolution and extensibility. Consistent experiment annotation is achieved by means of pre-defined annotation templates and controlled vocabularies. Furthermore, gene annotations are integrated from various public sources and are directly accessible for analysis. *GeWare* supports different types of analysis, in particular, visualization, statistics for detection of outlier and differentially expressed genes. The analysis results are uniformly stored in experiment and gene groups or expression matrices, which can be exchanged between different analysis steps. *GeWare* is operational and currently supports several research projects in Leipzig.

Acknowledgments

We thank Christine Körner and Jörg Lange for valuable help in implementing portions of *GeWare*. We also thank Hans Binder (IZBI, University of Leipzig), Friedemann Horn, Knut Krohn, Markus Eszlinger (Medical Department, University of Leipzig) and Philipp Khaitovich (Max-Planck-Institute for Evolutionary Anthropology, Leipzig) for continuous cooperation and helpful comments on *GeWare*. This work is supported by DFG grant BIZ 6/1-1.

References

[1] Affymetrix: Statistical Algorithms Description Document. White Paper, (http://www.affymetrix.com), 2002

[2] Binder, H. et al: The Sensitivity of Microarray Oligonucleotide Probes - The Effect of Base Composition and Saturation. Submitted.

[3] Binder, H. et al: Interactions in Oligonucleotide Duplexes Upon Microarray Hybridization - Insights from Probe Intensity Data. Submitted.

[4] Birney, E., et al: An Overview of Ensembl. Genome Research (14), 2004

[5] Bolstad, B. M., et al: A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2), 2003

[6] Brazma, A. et al: Minimum Information about a Microarray Experiment (MIAME) - Toward Standards for Microarray Data. Nature Genetics 19, 2001

[7] 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

[8] Do, H.H., Rahm, E.: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Proc. EDBT 2004, Heraklion, Greece, Springer LNCS, March 2004

[9] Ermolaeva, O. et al: Data Management and Analysis for Gene Expression Arrays. Nature Genetics 20, 1998

[10] Fellenberg, K. et al: Microarray Data Warehouse Allowing for Inclusion of Experiment Annotations in Statistical Analysis. Bioinformatics 18(3), 2002

[11] The Gene Ontology Consortium: The Gene Ontology (GO) Database and Informatics Resource. Nucleic Acids Research 32, 2004

[12] Irizarry, R. A. et al: Summaries of Affymetrix GeneChip Probe Level Data. Nucleic Acids Research 31(4), 2003

[13] Jarke, M. et al. (Eds): Fundamentals of Data Warehouses, Springer-Verlag, 2nd ed., 2003

[14] 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

[15] Liu, G., A.E. Loraine et al: NetAffx - Affymetrix Probeset Annotations. Proc. ACM SAC 2002

[16] Mangalam, H. et al: GeneX: An Open Source Gene Expression Database and Integrated Tool Set. IBM System Journal 40(2), 2001

[17] Markowitz, V.M., et al: Integration Challenges in Gene Expression Data Management. in Lacroix, Z., Critchlow, T.: Bioinformatics. Morgan Kaufmann Publishers, 2003

[18] Mützel, B et al: Functional Profiling of Genes Differently Expressed in the Brains of Humans and Chimpanzees. Poster/Abstract. Proc. 2nd Biotechnology Day, Leipzig, May 2003

[19] Pruitt, K.D., Maglott, D.R.: RefSeq and LocusLink: NCBI Gene-centered Resources. Nucleic Acids Research 29(1), 2001

[20] Sherlock, G. et al.: The Stanford Microarray Database. Nucleic Acids Research 29(1), 2001

[21] Stoeckert, C. et al: A Relational Schema for Both Array-based and Sage Gene Expression Experiments. Bioinformatics 17(4), 2001

[22] Westfall, P.H., Young, S.S.: Resampling-based multiple testing. Wiley&Sons Inc., 1993

[23] Wheeler D.L., et al. Database Resources of the National Center for Biotechnology. Nucl Acids Res 31:28-33, 2003.