Gene Expression Warehousing in Leipzig

Toralf Kirsten[1], Hong Hai Do[1], Dr. Dieter Sosna[2]
Prof. Dr. Erhard Rahm[1,2]
[1]Interdisciplinary Centre for Bioinformatics http://bioinformatik.uni-leipzig.de
[2]Database Group http://dbs.uni-leipzig.de

Dr. Knut Krohn[3], Markus Eszlinger[3]
Prof. Dr. Ralf Paschke[3,4]
[3]Interdisciplinary Centre for Clinical Research http://www.uni-leipzig.de/~izkf/
[4]Medical Department III http://www.uni-leipzig.de/~innere/

UNIVERSITÄT LEIPZIG
Interdisciplinary Centre for Bioinformatics Leipzig
Interdisciplinary Centre for Clinical Research Leipzig

Motivation

- Different technologies to detect gene expression
  - EST clustering, Microarray, SAGE, etc.
- (Qualitative/quantitative) gene expression analysis: answers to many questions
  - determination of gene functions as genome response to changes of environmental conditions, developmental stages, or between different tissues, etc.
  - co-expression of genes
  - discovery of new genes
- Microarray: most promising technology for gene expression analysis, simultaneous study of thousands of genes, producing huge amounts of valuable data with every single experiment
- Data management: major bottleneck in applications for gene expression analysis; impact on effectiveness of microarray experiments
- Open problems:
  - integrating data on experiments with existing, publicly available data, such as sequences, pathways
  - integrating gene expression data produced by different technologies, experiments
  - flexible analysis and data mining approaches (statistical evaluation, detection of relevant patterns, ...)

Molecular biology research at University of Leipzig

- 15 user groups with different research focus; all requiring gene expression analysis
  - Change detection in signal transduction in thyroid pathologies
  - Gene expression profiling of brain tumors
  - Comparative genomics for different primates
- Technologies: Affymetrix Oligonucleotide Microarrays (Chip fabrication, Wash, Scan, Data management and analysis)
- Experiments: about 300-500 experiment series / years
- Current situation:
  - Service center responsible for array processing and distribution of array results (raw data or Excel sheets) to respective users
  - Data analysis locally by single users
  - Limited data management and analysis capabilities provided by standard software (Affymetrix)

Related work

- Microarray Gene Expression Database (MGED) Group. International consortium, suggestion of standards for storing and representing data on gene expression experiments
- Several databases for gene expression data, RDBMS-based, publicly accessible through WWW, high detail level: image and related information for single array spots:
  - ***ArrayDB*** *(NHGRI)*: Sybase. Covering the entire process of array fabrication, wash and scan. Cy3 / Cy5 glass slide data. Probe sequences are linked with external databases, e.g. UniGene, KEGG, using their corresponding identifier. Flexible query tool with support for cross-experiment analysis based on visualization. No clustering algorithms.
  - ***GeneX*** *(NCGR)*: PostgresSQL / Sybase. Integration with external databases (dbEST, KEGG) through hyperlinks. Interoperability by means of GeneXML. Two clustering algorithms: hierarchical and permuation-based. Integrated tool for statistical analysis
  - ***Stanford Microarray Database*** *(SMD – University of Stanford)*: Oracle. Cy3 / Cy5 glass slide data. Data submissions from different scientific publications. Hierarchical clustering and SOM.
- Advantages:
  - Experiences from building databases for gene expression data
  - Open schemas; great help for constructing a customized data model meeting specific local needs
- Drawbacks:
  - Specifically developed tools for data access and analysis. No ad-hoc query / reporting tools, or integration of commercial tools, e.g. for data mining
  - Limited analysis capabilities: analysis of data from single experiments; main reason: lack of a uniform experiment / sample annotation mechanism to identify comparable experiments.
  - No „real" integration with external data, such as sequences, pathways; navigational access per hyperlinks
  - No (semantic) integration with other types of gene expression data, e.g. generated by EST cluster profiling

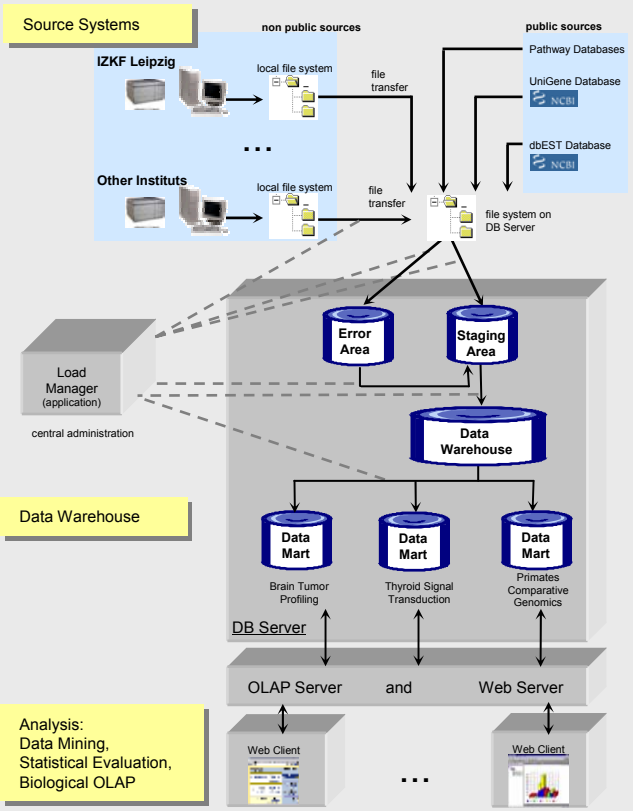
References

- Brazma, A. et al.: *Minimum Information about a Microarray Experiment (MIAME) toward Standards for Microarray Data.* **Nature Genetics**, Vol. 29, 2001
- Ermolaeva, Olga et al.: *Data Management and Analysis for Gene Expression Arrays.* **Nature Genetics**, Vol. 20, 1998
- Mangalam, H. et al.: *GeneX: An Open Source Gene Expression Database and Integrated Tool Set.* **IBM System Journal**, Vol. 40, No. 2, 2001
- Masys, Daniel: *Database Design for Microarray Data.* **The Pharmacogenomics Journal**, Vol. 1, No. 4, 2001
- Sherlock, G. et al.: *The Standford Microarray Database.* **Nucleic Acids Research**, Vol.29, No.1, 2001

Goals

- Systematic management of expression data generated by (local) microarray experiments
- Flexible support for different comparative gene expression studies
- Uniform experiment/sample annotation, especially for cross-experiment analysis
- Integration with gene expression data from other local (EST cluster profiling) projects
- Integration with external data from public databases
- Use of tools for analysis, data mining and visualization purposes

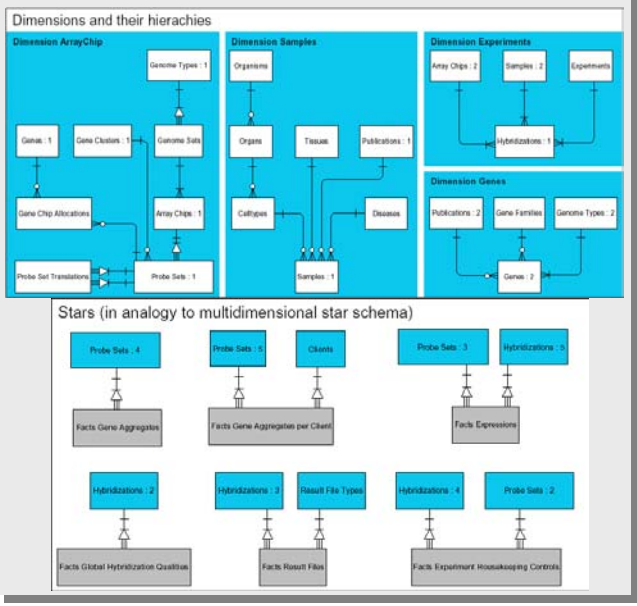
System architecture



Project steps

1. **Requirement Analysis**: Analysis of user requirements, especially functionality, report examples, security levels; Analysis of internal and external source systems; Determining a quantitative structure for daily / monthly / yearly data
2. **Warehouse Modeling**: Construction of a multi-dimensional data model for gene expression data, construction of ontologies to annotate genes (probe sets), experiments and sample sets; Implementation of the data warehouse in a commercial relational or object-relational DBMS
3. **Tool Evaluation:** Specification, order and installation of required Warehouse specific hardware (server, net, etc.) incl. software (OS, DBMS, Report tools, etc.); Decision „Make or Buy" – individual software vs. standard software
4. **Warehouse Population**: Design and implementation of load management routines for expression data generated by local experiments and related data from public databases
5. **Data Access**: Design and implementation of user-friendly web interfaces for common analysis, data submission and administration tasks; Integration of commercial and other public domain tools for searching, querying, data mining
6. **Warehouse Pilot and Test**
7. **Roll-out**: Evaluation of different analysis procedures for different molecular biological studies

First approach of a multidimensional data model



Milestones and perspective