IZBI

UNIVERSITÄT LEIPZIG

A Data Warehouse for Multidimensional Gene Expression Analysis

Toralf Kirsten[1], Hong-Hai Do[1], Erhard Rahm[1,2]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig http://www.izbi.de

[2] Department of Computer Science, University of Leipzig http://dbs.uni-leipzig.de

Objectives

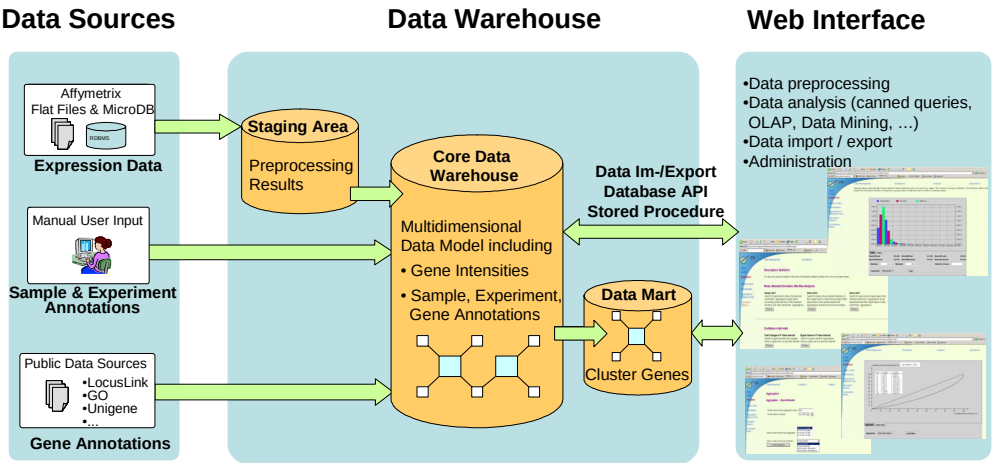
- Management and analysis of complex molecular-biological data of users for research networks with fast growing amount of data
- Design and implementation of flexible databases and analysis platforms for interdisciplinary projects and clinical studies
- Database research topics:
  - Integration of molecular-biological data and metadata (e.g. annotations)
  - Database coupling / integration of analysis algorithms and tools
  - Flexible, high performance data organization and querying

Main Results

- Comparative evaluation of microarray-based gene expression databases showed limitations of previous approaches [Do 2003a]
- **GeWare**: Design and implementation of a data warehouse for gene expression analysis; first version of warehouse operational
- **GenMapper**: Integration of gene annotations from different public sources; first version operational

GeWare System Architecture

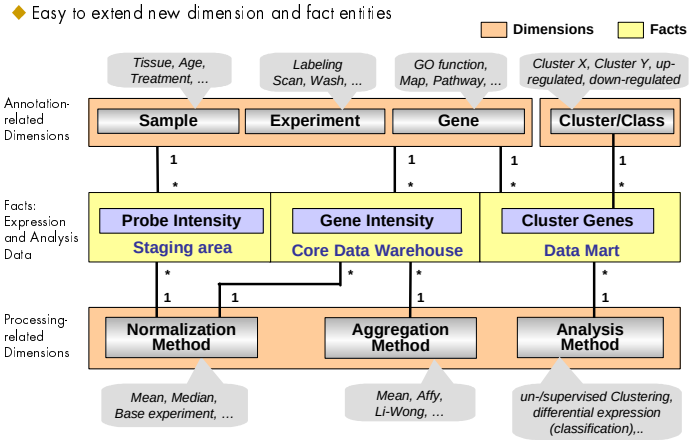
- Flexible data management for gene expression analysis based on Affymetrix oligonucleotide arrays
- Large amounts of data (around 500 experiment series per year) generated by local user groups
- Innovative data warehouse approach:
  - Multidimensional data organization
  - Integration of sample/experiment and gene annotation data with expression data
  - Support for several normalization, aggregation and analysis methods
  - Integration of existing analysis tools



Data Warehouse Model

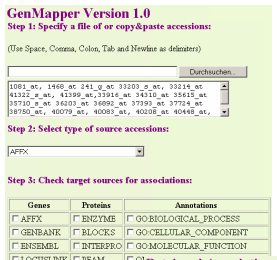
Multidimensional data model allowing:

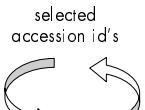
- OLAP-like navigations
- Individual / comparative analysis using subsets of data determined by
  - Specific dimensions (genes, experiments etc. and its annotation)
  - Limited expression values
- Flexible structure to add new processing methods without any model change
- Easy to extend new dimension and fact entities



Annotation Integration / Management

- Public sources with annotations refer to different gene representations, i.e. identifiers
  - Public sources: LocusLink, Human Genome Browser, Ensembl, UniGene, GeneCards, GeneLynx, ...
  - Vendor-based sources, e.g. NetAffx (Affymetrix): annotations of proprietary genes, i.e. probe sets
- Flexible management of sources its vocabularies and intra- / interdependencies
- Goal: Providing gene-oriented views on annotations by matching between different gene representations [Do 2003b, Mützel 2003]





Analysis Support

- Tight integration of several preprocessing methods, such as for background subtraction, normalization and aggregation
- Several analysis reports, i.e. canned queries for descriptive statistics to detect outlier and differential expression
- Advanced analysis using Insightful ArrayAnalyzer coupled with GeWare compromising large and valuable function libraries
- Export interface gene expression matrix for all or subset of genes due to analysis in external tools without database access
- Sequence-dependent sensitivity analysis of oligonucleotide probes [Binder 2003] by means of user defined database functions
  - Single sequence functions, e.g. baseCount, sequenceQuality, sequenceComplement
  - Match functions, e.g. extendSequence
  - Probe grouping function

References

[Binder 2003] Binder, H., Hofacker, I., Kirsten, T., Löffler, M., Richter, P., Stadler, P.: Sequence specific sensitivity of oligonucleotide probes. EUNITE Workshop: Intelligent Technologies for Gene Expression Based Individualized Medicine , Jena, May 2003

[Do 2003a] 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

[Do 2003b] Do, H.H. Rahm, E.: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Technical Report, University of Leipzig, July 2003

[Do 2001] Do, H.H., Rahm, E., Krohn, K., Paschke, R.: DBMS-based EST Clustering and Profiling for Gene Expression Analysis. First Workshop Computational Biology in Saxony: Problems and Perspectives. Dresden, November 2001

[Mützel 2003] Mützel, B., H.H. Do, P. Khaitovich, G. Weiß, E. Rahm, S. Pääbo: Functional Profiling of Genes Differently Expressed in the Brains of Humans and Chimpanzees. Abstract, Proc. 2nd Biotechnology Day, Leipzig, May 2003

Printed by Universitätsrechenzentrum Leipzig