IZBI

UNIVERSITÄT LEIPZIG

Data Integration for Analyzing Gene Expression Data

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

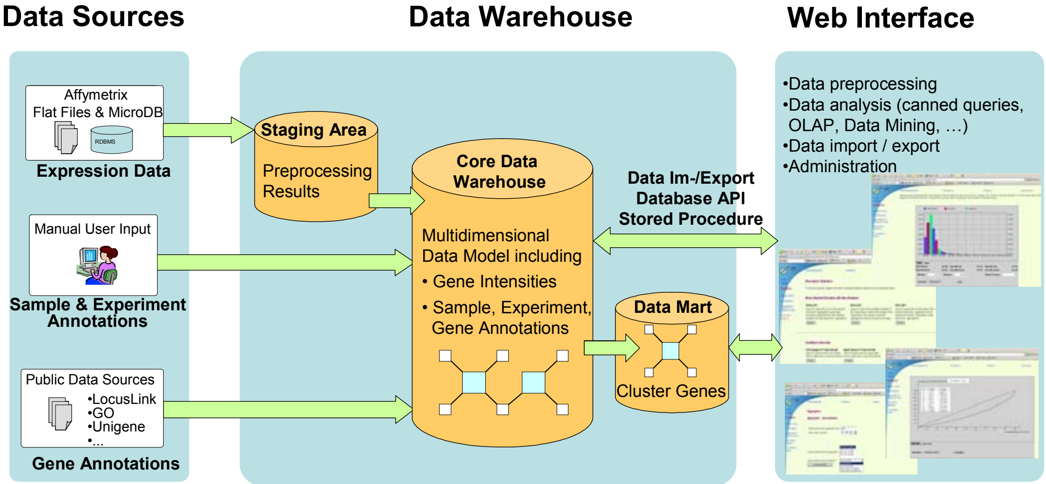
- Management and analysis of complex molecular-biological data of users for research networks with fast growing amount of data
- Design and implementation of flexible databases and analysis platforms for interdisciplinary projects and clinical studies
- Database research topics:
  - Integration of molecular-biological 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 2003b]
- **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

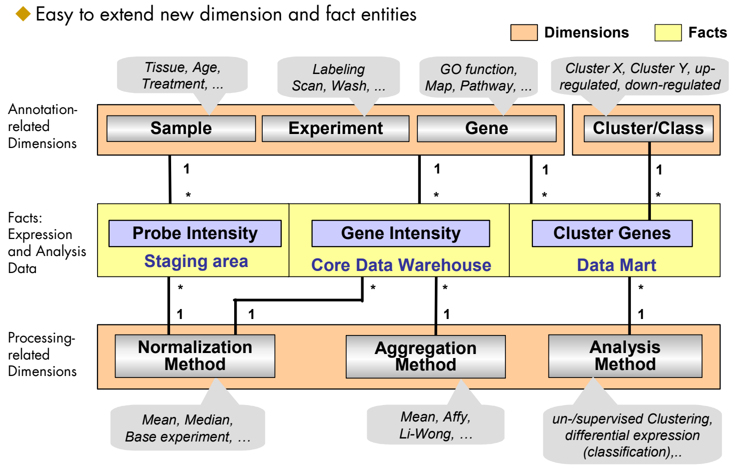
- Flexible data management for gene expression analysis based on Affymetrix oligonucleotide arrays
- Large amounts of data (around 500 experiment series per year) generated by local user groups
- Innovative data warehouse approach:
  - Multidimensional data organization
  - Integration of sample/experiment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



Data Warehouse Model

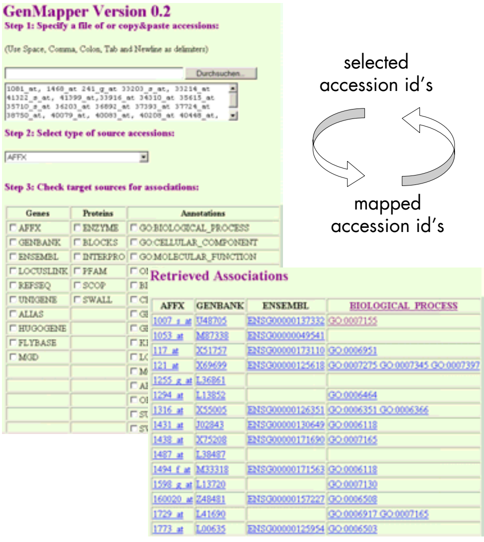
Multidimensional data model allowing:

- 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


Dimensions
Facts
Tissue, Age, Treatment, ...
Labeling Scan, Wash, ...
GO function, Map, Pathway, ...
Cluster X, Cluster Y, up-regulated, down-regulated
Annotation-related Dimensions
Sample
Experiment
Gene
Cluster/Class
1
*
1
*
1
*
1
*
Facts: Expression and Analysis Data
Probe Intensity
Staging area
Gene Intensity
Core Data Warehouse
Cluster Genes
Data Mart
*
1
1
*
1
*
1
*
1
Processing-related Dimensions
Normalization Method
Aggregation Method
Analysis Method
Mean, Median, Base experiment, …
Mean, Affy, Li-Wong, …
un-/supervised Clustering, differential expression (classification),..

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 2003a, Mützel 2003]



GenMapper Version 0.2

Step 1: Specify a file of or copy&paste accessions:

(Use Space, Comma, Colon, Tab and Newline as delimiters)

Durchsuchen...

```
1081_at, 1468_at 241_g_at 33203_s_at, 33214_at
41322_s_at, 41399_at,33916_at 34310_at 35615_at
35710_s_at 36203_at 36892_at 37393_at 37724_at
38750_at, 40079_at, 40083_at, 40208_at 40448_at,
```

Step 2: Select type of source accessions:

AFFX

Step 3: Check target sources for associations:

Genes	Proteins	Annotations
□ AFFX	□ ENZYME	□ GO:BIOLOGICAL_PROCESS
□ GENBANK	□ BLOCKS	□ GO:CELLULAR_COMPONENT
□ ENSEMBL	□ INTERPRO	□ GO:MOLECULAR_FUNCTION
□ LOCUSLINK	□ PFAM	□ OI
□ REFSEQ	□ SCOP	□ BI
□ UNIGENE	□ SWALL	□ CI
□ ALIAS		□ GI
□ HUGOGENE		□ GI
□ FLYBASE		□ KI
□ MGD		□ LC
		□ M
		□ AI
		□ OI
		□ SI
		□ SV

Retrieved Associations

AFFX	GENBANK	ENSEMBL	BIOLOGICAL_PROCESS
1007_s_at	U48705	ENSG00000137332	GO:0007155
1053_at	M87338	ENSG00000049541	
117_at	X51757	ENSG00000173110	GO:0006951
121_at	X69699	ENSG00000125618	GO:0007275 GO:0007345 GO:0007397
1255_g_at	L36861		
1294_at	L13852		GO:0006464
1316_at	X55005	ENSG00000126351	GO:0006351 GO:0006366
1431_at	J02843	ENSG00000130649	GO:0006118
1438_at	X75208	ENSG00000171690	GO:0007165
1487_at	L38487		
1494_f_at	M33318	ENSG00000171563	GO:0006118
1598_g_at	L13720		GO:0007130
160020_at	Z48481	ENSG00000157227	GO:0006508
1729_at	L41690		GO:0006917 GO:0007165
1773_at	L00635	ENSG00000125954	GO:0006503

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 integrated 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. Rahm, E.: Generic Management and Integration of Molecular-biological Annotation Data. Technical Report, University of Leipzig, July 2003

[Do 2003b] 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 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