Data integration for analyzing gene expression data

UNIVERSITÄT LEIPZIG

Hong-Hai Do[1], **Toralf Kirsten**[1], **Erhard Rahm**[1,2]
[1]Interdisciplinary Centre for Bioinformatics [2]Department of Computer Science
http://www.izbi.de http://dbs.uni-leipzig.de

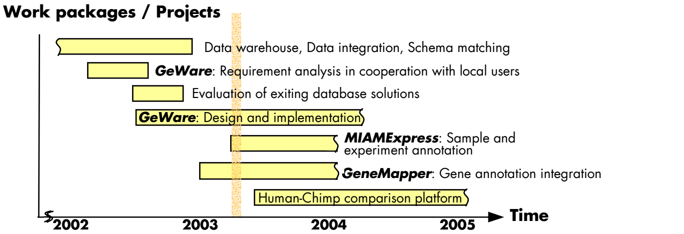
1. Motivation

- Microarrays: High-throughput method to monitor the expression of thousands of genes in a single experiment
  - Promising, however high requirements to a supporting data management and analysis solution
- Various user groups at University of Leipzig with different research focus, requiring gene expression analysis
  - Comparative genomics for different primates
  - Change detection in signal transduction in thyroid pathologies
  - Gene expression profiling of brain tumors
- Current capacity: about 300-500 experiments per year using Affymetrix microarrays

2. Gene Expression Warehousing: Goals and Objectives

- Data warehouse: Central data management and analysis platform supporting special local requirements
- Integration of annotation data with expression data
  - Experiment annotations: documenting the specific biological focus of an experiment and the technical process of conducting it
  - Gene annotations: describing all known aspects about the gene sequences on the microarray chips; to be exploited from public sources
- Integration of existing software and tools with the database for flexible data analysis

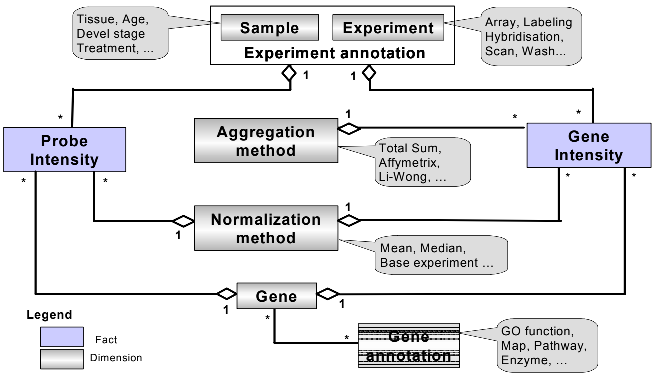
3. Project Context: A Chronology



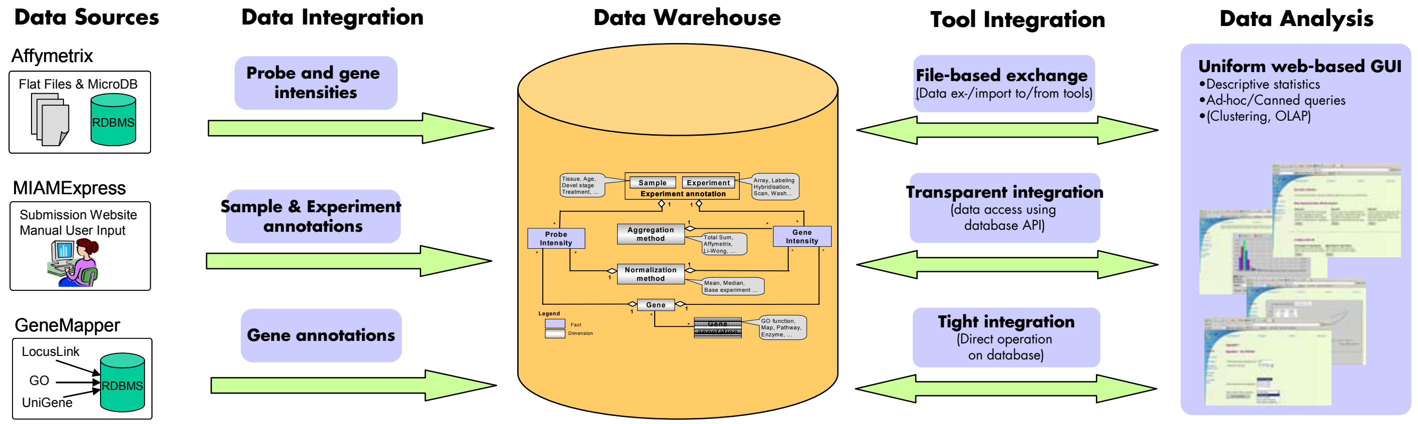
4. Evaluation of Existing Solutions [2]

- Comparison of 8 published microarray databases:
  - ArrayDB (NHGRI), ExpressDB (Harvard Univ.), GeneX (NCGR), GIMS (Univ. Manchester), M-CHIPS (DKZF), RAD2 (Univ. Pennsylvania), SMD (Stanford Univ.), YMD (Yale Univ.)
- Major evaluation criteria
  - Data storage and management: Image and expression data, annotation data
  - Data integration: Integration of annotation data, especially gene annotations, with expression data
  - Tool integration: Integration of tools/algorithms for interactive and iterative data analysis
- State of the art
  - Sample and experiment annotations: Mostly free-text fields, no controlled vocabularies used
  - Gene annotations: Mostly not locally integrated but linked through web links, not sufficient for analysis
  - Data analysis and tool integration: Large variety of data mining approaches available, however advanced analysis outside of database by means of stand-alone tools

5. GeWare Data Model


Tissue, Age, Devel stage Treatment, ...
Sample
Experiment
Experiment annotation
Array, Labeling Hybridisation, Scan, Wash...
1
1
*
Probe Intensity
Aggregation method
Total Sum, Affymetrix, Li-Wong, ...
*
Gene Intensity
*
*
*
*
1
Normalization method
1
1
Mean, Median, Base experiment ...
Gene
1
1
*
*
Gene annotation
GO function, Map, Pathway, Enzyme, ...
Legend
Fact
Dimension

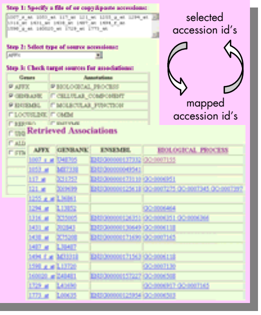
6. GeWare Overall Architecture



7. Contributing/Built-upon Projects

Integrating gene annotations with GeneMapper

- Various public sources with different annotations, however related to different gene representations, i.e. identifiers
  - Public sources: LocusLink (NCBI), Human Genome Browser (UCSC) and Ensembl (EBI-Sanger), UniGene, Tigr, GeneCards, GeneLynx, ...
  - Vendor-based sources, e.g. NetAffx (Affymetrix): annotations of proprietary genes, i.e. probesets
- Project goals: Providing gene-oriented views on annotations by means of matching between different gene representations



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

Step 2: Select type of source accessions:

AFFX

Step 3: Check target sources for associations:

Genes	Annotations
☑ AFFX	☑ BIOLOGICAL_PROCESS
☑ GENBANK	☐ CELLULAR_COMPONENT
☑ ENSEMBL	☐ MOLECULAR_FUNCTION
☐ LOCUSLINK	☐ OMIM

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	[illegible]	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

Human-Chimpanzee Comparison Platform (in cooperation with MPI-EVA)

- Recent availability of draft versions of the human and chimpanzee genomes: first example of two closely related mammalian genomes
- Project goals: Design and implementation of an integrated platform for comparative analysis between humans and chimpanzees [3]
  - Genome-wide comparisons at both sequence and expression levels
  - Determining genes with expression drastically changed during human evolution: Identification of traits specific to humans
- Collecting data at the MPI-EVA in cooperation with Charite Berlin
  - Application of various techniques: Microarray, mass spectrometry, gel electrophoresis
  - High volume of data expected: currently ca. 3TB data available stored in flat files associated with ca. 800 microarray-based experiments

8. Cooperation Partners

Local partners

❑ **University of Leipzig**
- Medical Department (F. Horn)
- Dept. of CS (E. Rahm/P. Richter)
- Carl-Ludwig-Institute (W. Briest/A. Deten)

❑ **MPI-EVA** Leipzig (Svante Pääbo's Lab)
❑ **IZKF Leipzig** (K. Krohn/M. Eszlinger)

IZBI
WG 1 ↔ WG 3

International / National partners

❑ **EBI** Cambridge, UK (A. Brazma/U. Sarkans)
❑ **Lion BioScience**, UK (T. Etzold)
❑ **Humboldt Uni. Berlin**
- Dept. of CS-DB (J.-C. Freytag/S. Heymann)
- Dept. of CS-KM (U. Leser)
- Dept. of CS-DI (F. Naumann)

9. References

1. Brazma, A. et al.: *Minimum Information about a Microarray Experiment (MIAME) – Toward Standards for Microarray Data.* **Nature Genetics 29**, 2001
2. Do, H.-H., Kirsten, T., Rahm, E.: *Comparative Evaluation of Microarray-based Gene Expression Databases.* Proc. **10. Conf. Database Systems for Business, Technology and Web** (BTW), Leipzig, Feb. 2003
3. Enard, W. et al.: *Intra- and Interspecific Variation in Primate Gene Expression Patterns.* **Science 296**, 2002
4. Kirsten, T., Do, H.-H., Rahm, E.: *GeWare: A Data Warehouse for Integrated Gene Expression Analysis*, Work in progress, 2003