Comparative evaluation of microarray-based gene expression databases

Hong-Hai Do,
Toralf Kirsten,
Erhard Rahm

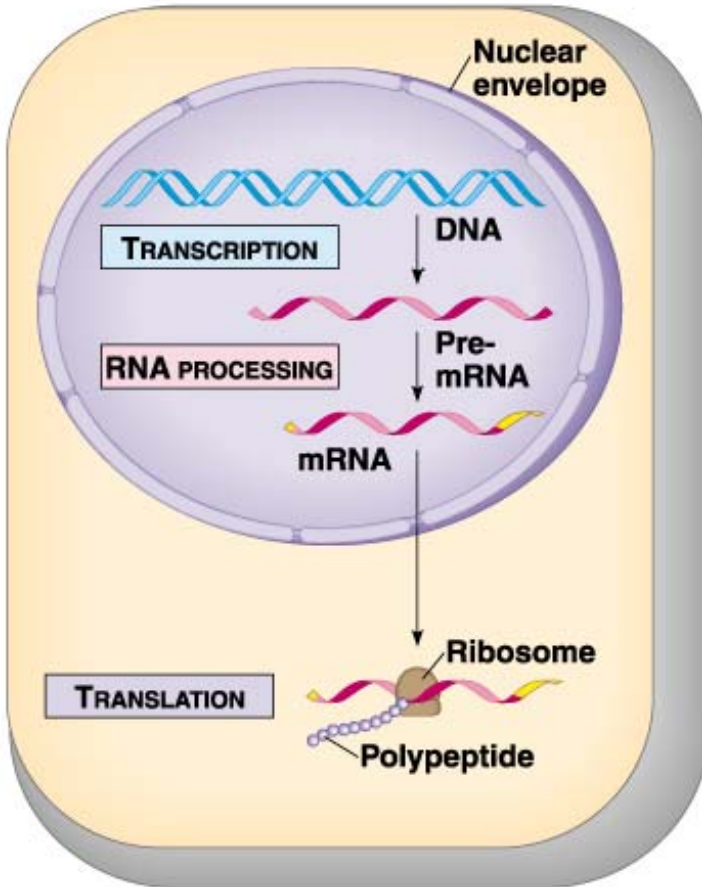
University of Leipzig, Germany
www.izbi.de, dbs.uni-leipzig.de

Content

- Gene expression analysis
- Microarray database requirements
- Evaluation of 8 database solutions
- The GeWare project in Leipzig
- Conclusions

Gene Expression Analysis

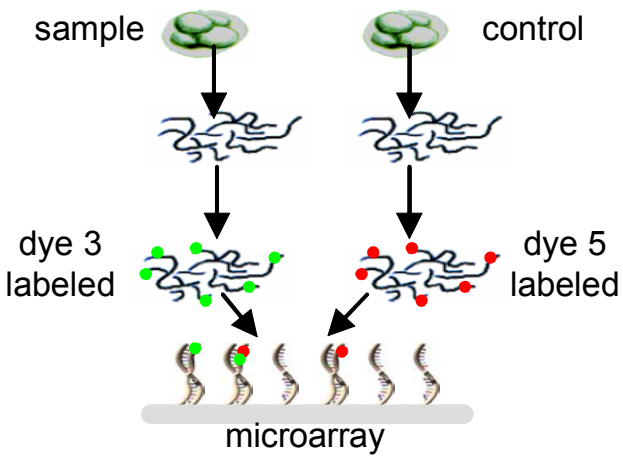
- Goal: Characterization of functions of genes and their mutual influence in the regulatory network



- Measuring mRNA amount in cells under different conditions
- Microarrays
  - Measuring expression of thousands of genes simultaneously
  - Large amounts of data with every experiment

Microarray Experiment

cDNA Arrays



	0	1	2	3
0	18,2	63	78,4	483
1	15,4	33,6	58,8	340,2
2	5,6	43,4	2,8	64,4
3	495,6	638,4	33,6	764,4
4	327,6	106,4	74,2	50,4
5	4,2	44,8	105,28	33,6
6	44,8	43,4	109,2	49
7	172,2	16,8	64,4	16,8
8	32,2	1369,2	454,16	75,6
9	120,4	44,8	327,6	18,2
10	63	16,8	903	64,4
11	79,8	75,6	77	19,6

spot intensities

(1) Cell selection

(2) RNA/DNA preparation

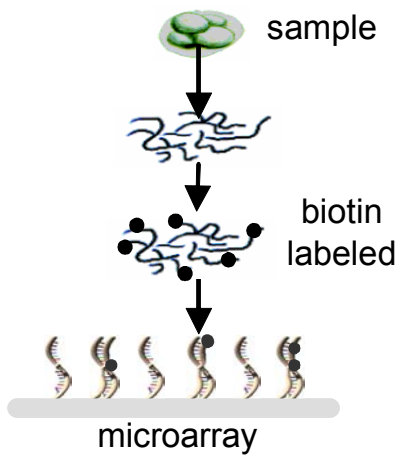
(3) Hybridization

(4) Array scan

(5) Image analysis

(6) Expression analysis

Oligonucleotide Arrays



	0	1	2	3
0	10	3,8	10	10
1	396,7	475,4	388,5	294,6
2	170,3	172,4	50,7	74,4
3	10	10	32,7	10
4	42,2	10	10	10
5	416,3	263,5	724,7	605,4
6	95,2	79,5	36,6	32,7
7	10	10	42,3	56,3
8	10	4,6	23,4	10
9	10	10	10	9,3
10	8,5	101,8	2,1	7,6
11	718,2	384,2	225	231,4

spot intensities

Local Situation

- In Leipzig: ~15 different user groups:
  - Comparative primate genomics (human vs. chimpanzee)
  - Change detection in signal transduction in thyroid pathologies
  - Gene expression profiling of brain tumors
  - ...
- Affymetrix oligonucleotide microarrays
  - About 300-500 experiment series / year (trend ↗)
- Current data management and analysis:
  - Handling of flat files produced by Affymetrix software
  - Data analysis using Affymetrix tools, MS Excel
  - Manual search for annotations in public sources

Database Requirements

- Storage of different types of data
- Data integration
- Annotation management
- Data normalization
- Data analysis
- Tool integration

Data Characteristics

- Various kinds of data with different characteristics and requirements

Data		Source	Type	Characteristics	Usage
Image Data		Array scan	binary	large files	Generation of expression data
Expression Data		Image analysis	number	fast growing volume	Visualization, statistical and cluster analysis
Annotation Data	Gene	External public sources	text	regularly updated	Interpreting / Relating / Inferring gene functions
	Experiment	User input		user-specified, often free text	

Annotation Integration

- Various public sources with gene annotations:
  - LocusLink and RefSeq: GO annotations, homology, organism, reference sequence
  - UniGene, GeneCards, GeneLynx, Tigr, ...
  - Vendor-specific sources: NetAffx

  - However, often different gene identifiers !!!

- Manual specification of experiment annotations
  - Free text to be limited/avoided for better analysis support

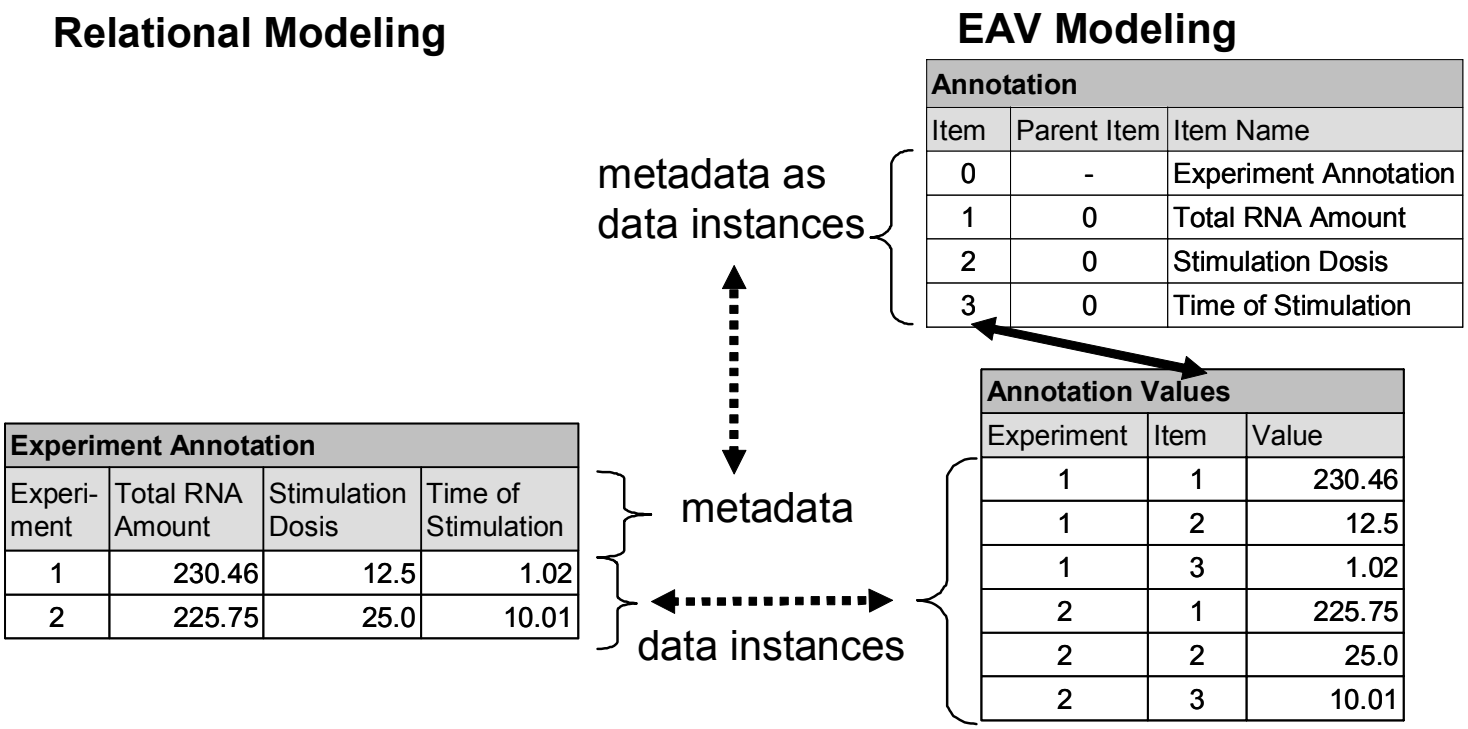
- Standard support necessary, e.g., MIAME, MAGE-ML, GeneOntology, ...

Data Integration Mechanisms

- Virtual integration
  - Web linkage based on accession keys
    - Navigational access
    - Annotation data not queryable
    - Little integration effort
  - Federated systems (Mediator-based)
    - Schema integration
    - On-the-fly data integration: transformation, cleaning, merging
    - Performance/Availability/Rudimentary query capabilities of public sources
- Materialized integration (Data warehousing)
  - All relevant annotation data + expression data locally stored
  - Advantages for data analysis: all data directly queryable, performance
  - High integration and update effort
- Hybrid approaches, e.g. SRS

Management of Annotation Data

- Flexible management required
  - Coping with attribute changes / fast-evolving schemas and vocabularies
- Database representation:
  - Relational vs. EAV (Entity-Attribute-Value)



Data Normalization

- Necessary for expression data due to
  - Fluctuations in technical experiment process
  - Comparison between multiple experiments
- Normalization for 1 experiment
  - Division by average intensity of all spots on array
  - Of control genes (housekeeping or spiked genes)
- Normalization for multiple experiments (series)
  - Normalization against a control experiment
- Storage of raw data for re-normalization

Data Analysis

- Navigation/Querying/Reporting
- Online analytical processing
  - Multidimensionality of expression data
- Statistics and data mining
  - Descriptive statistics: mean, standard deviation, …
  - Probability calculation: distributions, regression, correlation, …
  - Inductive statistics: random sampling, estimation, tests, …
  - Clustering: Hierarchical, K-mean, Self-Organizing Maps, …
  - Classification: Support Vector Machines, Decision trees, …
- Visualization
  - Display of statistical and clustering results
  - Scatter plots, dendrograms, charts, graphs, …

Tool Integration

- File exchange
  - Export from database, import in tool for analysis (tab-delimited ASCII format, XML etc.)
  - No integration effort, but restricted / static information
- API access to DBS by tools
  - Use of DBS is transparent to user
  - Access to current data using query language
- Tight integration: Direct analysis in database systems
  - Analysis / data mining approaches implemented by DBMS or as stored procedures
  - Potential for high performance
  - High implementation effort

Evaluation of 8 database solutions

Database	Organization
ArrayDB	National Human Genome Research Institute – NHGRI *http://genome.nggri.nih.gov/arraydb*
ExpressDB	Harvard University *http://arep.med.harvard.edu/ExpressDB*
GeneX	National Center for Genome Resources – NCGR *http://genebox.ncgr.org/genex*
GIMS	University of Manchester *http://www.cs.man.ac.uk/~norm/gims*
M-CHIPS	German Cancer Research Center *http://www.mchips.de*
RAD2	University of Pennsylvania *http://www.cbil.upenn.edu/RAD2*
SMD	Stanford University *http://genome-www4.stanford.edu/MicroArray/SMD*
YMD	Yale University *http://info.med.yale.edu/microarray*

Results: Data Management

- Supported types of data
  - Often no images stored
  - Expression data from different techniques (microarray-based and non-microarray)
- Gene annotations
  - not locally integrated/available in most cases
- Experiment annotations
  - Different content and varying degree of detail between the databases
  - Mostly free-text fields, no controlled vocabularies
- Data exchange
  - Tab-delimited used in many cases
  - XML not yet supported

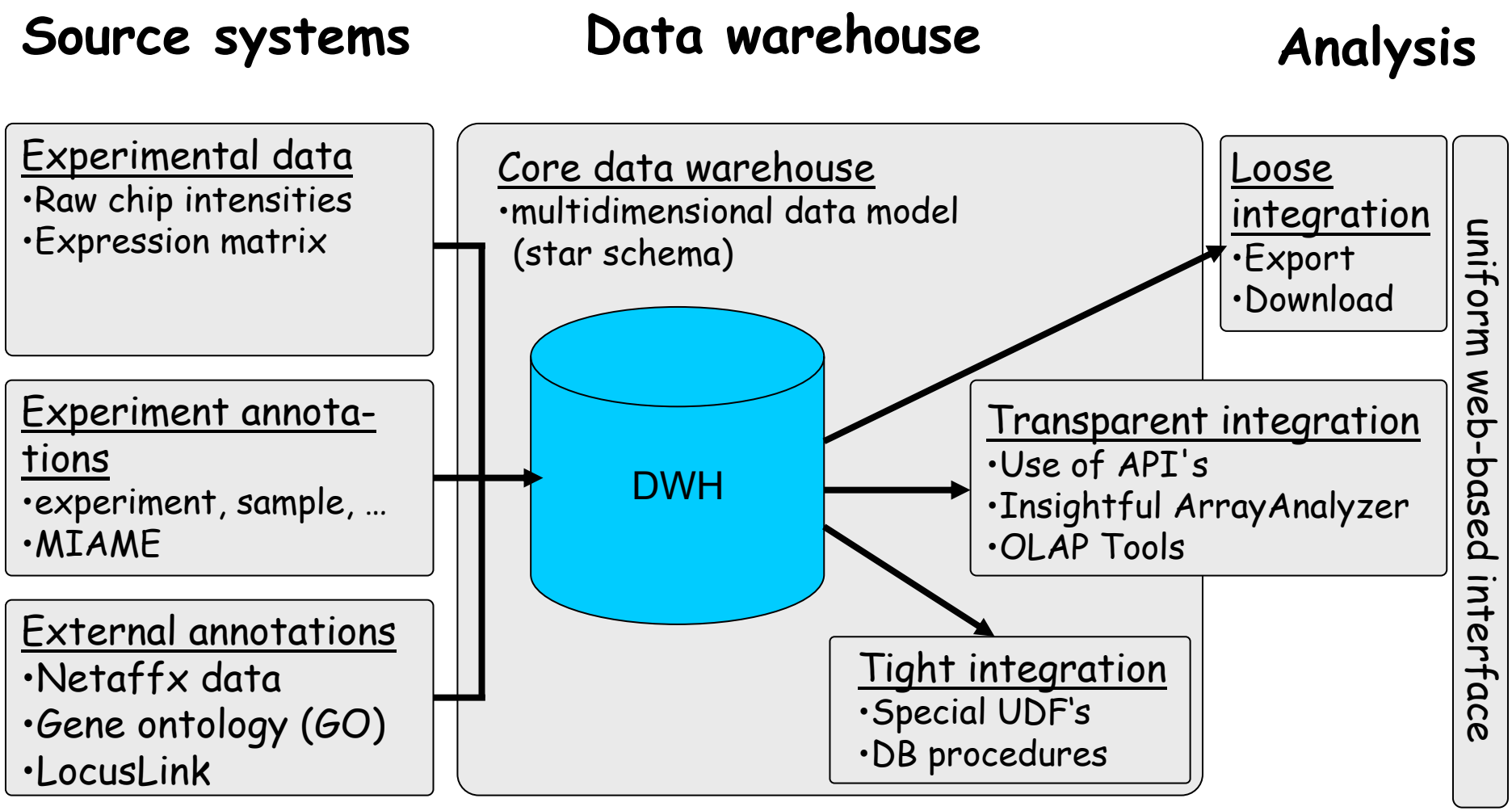
Results: Data and Analysis Integration

- Data integration
  - Web-link integration in most cases, but not sufficient for analysis
  - Federated and materialized integration not yet fully exploited
- Data analysis
  - Canned queries widely used
  - OLAP not yet applied despite multidimensionality
  - Large variety of data mining approaches
- Tool integration
  - Advanced analysis mostly outside of database by means of stand-alone tools

Project GeWare

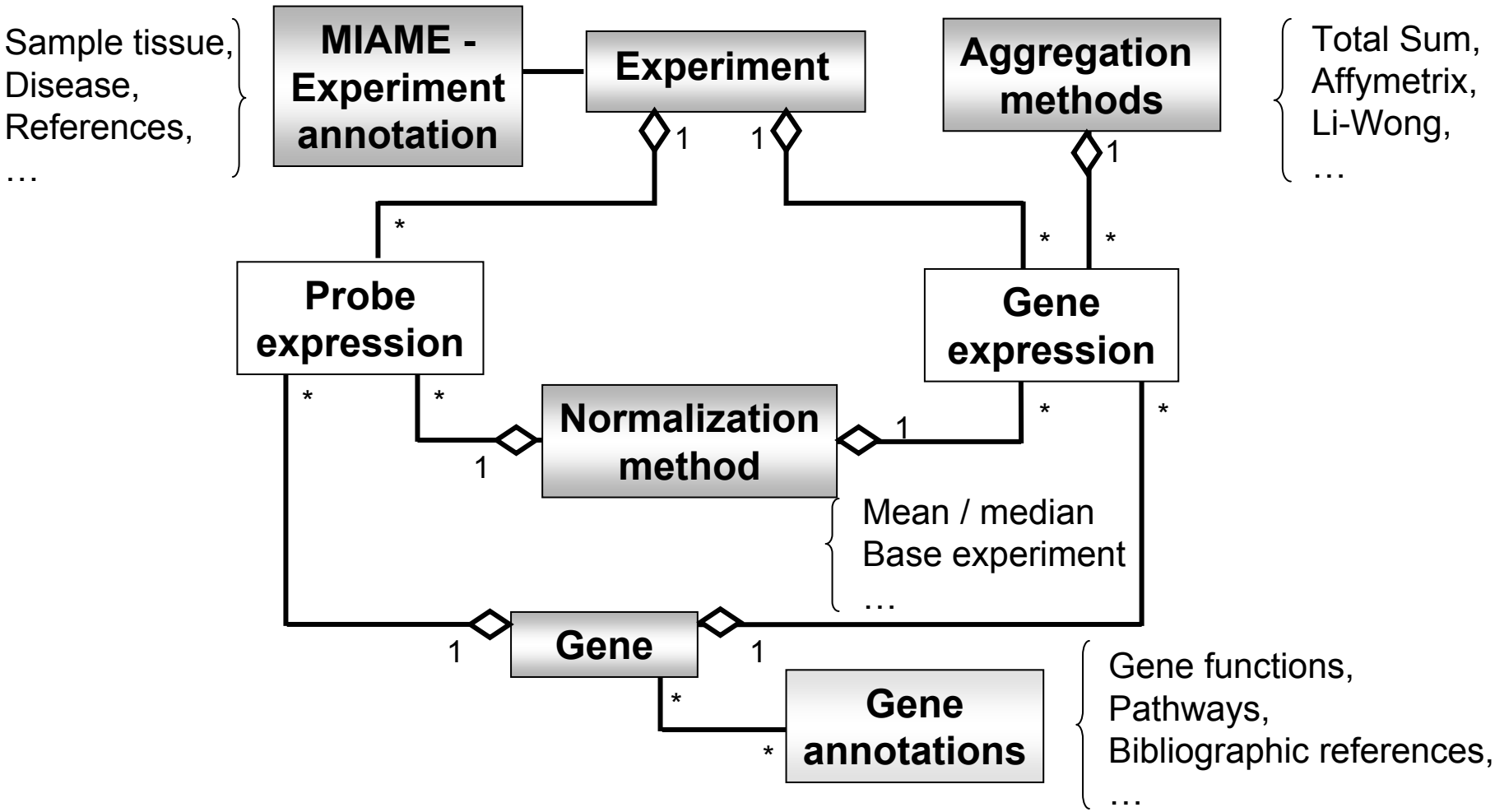
- Specific local requirements
- Central data management and analysis platform for local users
- Data Warehouse approach
  - Data import from Affymetrix system
  - Fact tables to store both raw and derived data
  - Uniform specification of experiment annotations
  - Integration of gene annotations from public sources
  - Integration of analysis and data mining algorithms/tools

System Architecture



Data Warehouse Model

- Multidimensional data model (star schema)



Conclusions

- Microarray-based gene expression analysis
  - Promising technique for a variety of biological problems
  - High requirements for data management
- State of the art: insufficient database integration of
  - Annotations
  - Analysis approaches
- GeWare