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
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

- **Sample**
- **Control**
- **Dye 3 labeled**
- **Dye 5 labeled**

### Diagram

#### Sample

- **Cell selection**
- **RNA/DNA preparation**
- **Hybridization**
- **Array scan**
- **Image analysis**

### Table

#### Sample

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15.2</td>
<td>15.4</td>
<td>25.6</td>
<td>45.6</td>
</tr>
<tr>
<td>1</td>
<td>78.4</td>
<td>58.4</td>
<td>33.4</td>
<td>76.4</td>
</tr>
<tr>
<td>2</td>
<td>440.2</td>
<td>440.2</td>
<td>440.2</td>
<td>440.2</td>
</tr>
<tr>
<td>3</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>5</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>6</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>7</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>8</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>9</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>10</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
</tbody>
</table>

#### Control

#### Array Image

### Oligonucleotide Arrays

- **Sample**
- **Biotin labeled**

### Diagram

#### Sample

- **Hybridization**
- **Array scan**
- **Image analysis**
- **Expression analysis**

### Table

#### Sample

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Control

#### Array Image

### 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

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

### Relational Modeling

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Time of Stimulation</th>
<th>Stimulation Dosis</th>
<th>Total RNA Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>12.5</td>
<td>230.46</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>25.0</td>
<td>225.75</td>
</tr>
</tbody>
</table>

### EAV Modeling

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Item</th>
<th>Parent Item</th>
<th>Item Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>0</td>
<td>Experiment Annotation</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Total RNA Amount</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>Stimulation Dosis</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>Time of Stimulation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation Values</th>
<th>Experiment</th>
<th>Item</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>230.46</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>225.75</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>25.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>10.01</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Database</th>
<th>Organization</th>
</tr>
</thead>
</table>
| ArrayDB  | National Human Genome Research Institute – NHGRI  
            [http://genome.nggrt.nih.gov/arraydb](http://genome.nggrt.nih.gov/arraydb) |
| ExpressDB | Harvard University  
            [http://arep.med.harvard.edu/ExpressDB](http://arep.med.harvard.edu/ExpressDB) |
| GeneX    | National Center for Genome Resources – NCGR  
            [http://genebox.ncgr.org/genex](http://genebox.ncgr.org/genex) |
| GIMS     | University of Manchester  
            [http://www.cs.man.ac.uk/~norm/gims](http://www.cs.man.ac.uk/~norm/gims) |
| M-CHIPS  | German Cancer Research Center  
            [http://www.mchips.de](http://www.mchips.de) |
| RAD2     | University of Pennsylvania  
            [http://www.cbil.upenn.edu/RAD2](http://www.cbil.upenn.edu/RAD2) |
| SMD      | Stanford University  
| YMD      | Yale University  
            [http://info.med.yale.edu/microarray](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

Source systems
- Experimental data
  - Raw chip intensities
  - Expression matrix
- Experiment annotations
  - experiment, sample, ...
  - MIAME
- External annotations
  - Netaffx data
  - Gene ontology (GO)
  - LocusLink

Data warehouse
- Core data warehouse
  - multidimensional data model (star schema)

Analysis
- Loose integration
  - Export
  - Download
- Transparent integration
  - Use of API's
  - Insightful ArrayAnalyzer
  - OLAP Tools
- Tight integration
  - Special UDF's
  - DB procedures

DWH uniform web-based interface
Data Warehouse Model

- Multidimensional data model (star schema)

Sample tissue, Disease, References, ...

MIAME - Experiment annotation

Experiment

Aggregation methods

Total Sum, Affymetrix, Li-Wong, ...

Probe expression

Mean / median Base experiment ...

Normalization method

Gene expression

Gene functions, Pathways, Bibliographic references, ...

Gene

Gene annotations

Sample tissue, Disease, References, ...

Gene annotations
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