3.4

Data analysis

Questions for microarray analysis

(1) Which are the different patterns of gene expression?
- Which genes closely share a pattern with gene X?

(2) What uncharacterized genes have similar expression patterns to well-characterized ones?
- What category of function might gene X belong to?

(3) Are there subtypes of disease X discernible by tissue gene expression?
- Which genes best differentiate between different tissue classes?

Data analysis ingredients

(1) Normalization
- Making values comparable across experiments

(2) Data mining
- Significant fold change
- Principal component analysis
- Clustering (unsupervised/supervised)

(3) Visualization
- Quality control
- Data visualization
- Visual data mining

Normalization

(1) Total intensity
- Assume that total DNA / protein content is always the same

(2) Regression
- Between two samples
- Linear / non-linear

(3) Set of known markers
- Housekeeping genes
- Added known RNA (corrects for purification efficiency)

→ Logarithmic expression ratio/level

(Absolute expression is restricted to positive - non-gaussian, logarithmic is assumed to be of gaussian distribution)

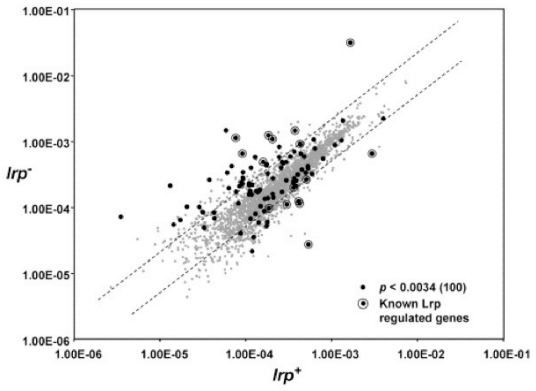
(N-)Fold analysis

(1) Comparison of two states
- Taking ratio between expression measurement
- Which ratios are significant (1.7, 2, 3, 10)?

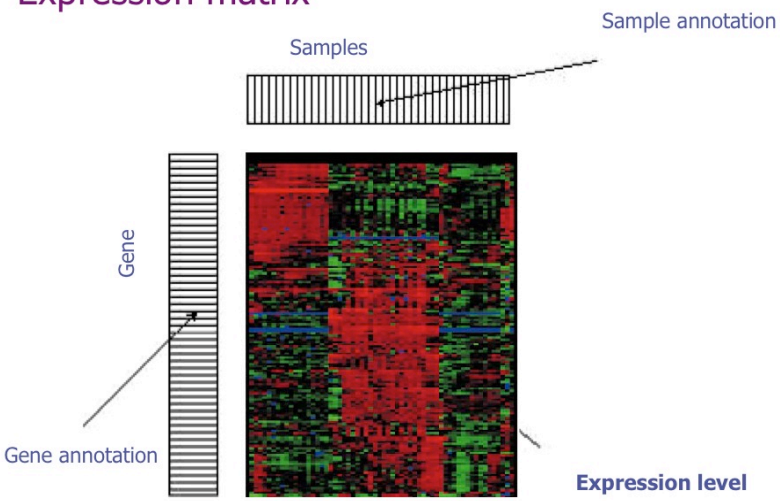
(2) Error sources
- Biological errors (fluctuation through regulation): often large, increase with expression levels, can be estimated by repeated measurements
- Technical errors (background noise): often smaller, but decrase with expression levels (signal-to-noise)

(3) Statistical estimates
- Depend on error model (of the particular array technology)
- t-test with desired level of confidence

Comparison plot of two states



Expression matrix



Cluster analysis

(1) Interpret differences in expression
- Finding groups which share a similar pattern of expression
- E.g. same reaction to different expression conditions (state, treatment)

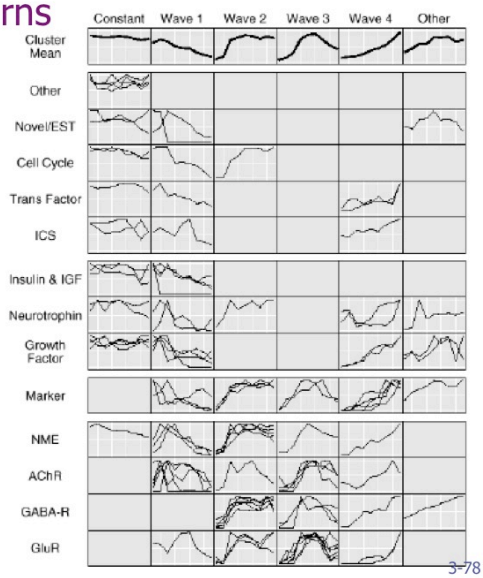
(2) Cluster analysis is a probabilistic process
- Clusters depend on a measure of distance/difference
- Clustering methods are somewhat arbitrary as they represent basically no prior knowledge

(3) Better analysis using error models
- Incorporating estimates of experimental error

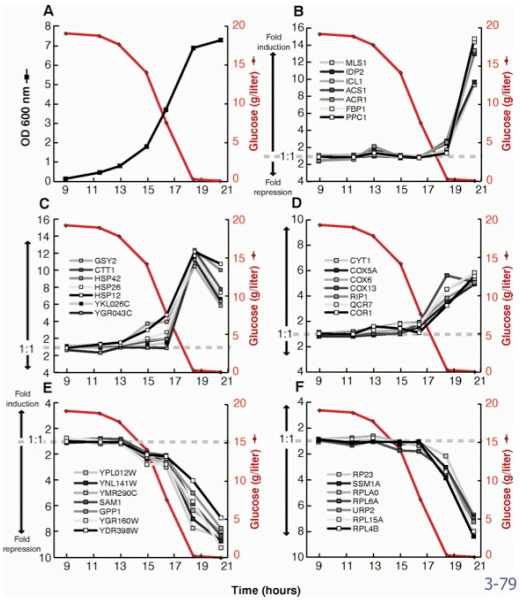
(4) New methods emerging
- Incorporating models of the biological behaviour studied

Expression patterns

Comparing different „waveforms" with different fuctional gene classes

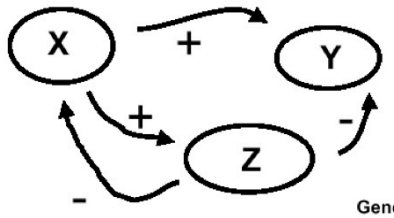


Diauxic shift



3-79

Regulatory networks



Gene Functions: X, Y, Z
Pathways: X-Y, X-Z-Y, X-Z-X
Networks: XYZ

(c) Russ B. Altman

3-80

Distance measures for clustering

(1) Each gene in the expression matrix is an n-dimensional vector
- Can be ordered (time series)
- Can be independent coordinates
- Can be arbitrary attributes

(2) Geometric distance
- Euclidian distance (dimensions must be normalized)

(3) Similarity
- Correlation
- Mutual information (suitable also for inverse regulation)

K-means clustering

(1) Given number of clusters k
- With coordinate (centroid)
- With set of assigned genes

(2) Initialize randomly to genes
- Select k genes randomly and assign each coordinate to one cluster

(3) While (clusters have changed)
- Assign each gene to its nearest centroid
- Calculate new centroids as the mean of the coordinates of the assigned genes

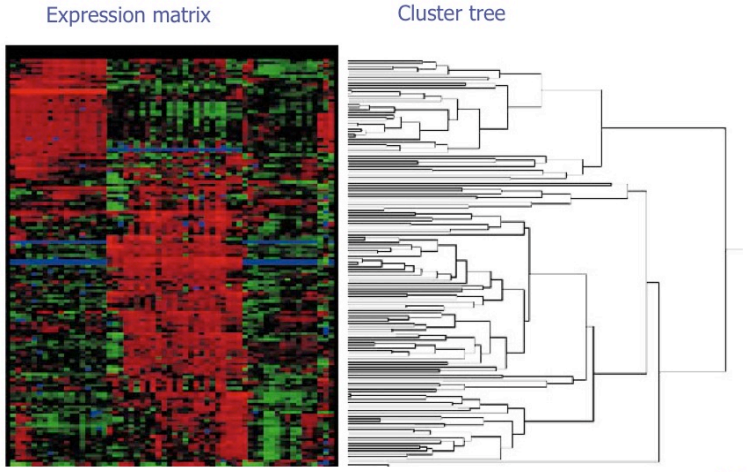
Dendrograms

(1) Built like a phylogenetic tree
- UPGMA clustering of genes

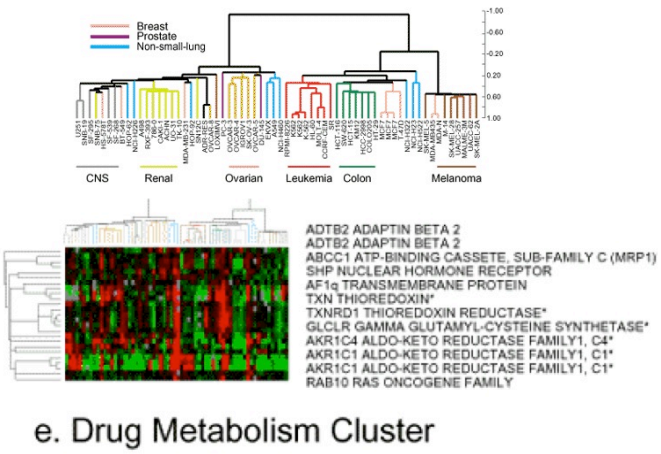
(2) Sort genes hierarchically according to tree
- Can be shown together with sorted expression matrix

(3) 2-dimensional dendrograms
- Sort samples according to clustering on attributes

Example clustering



2-dimensional dendrograms



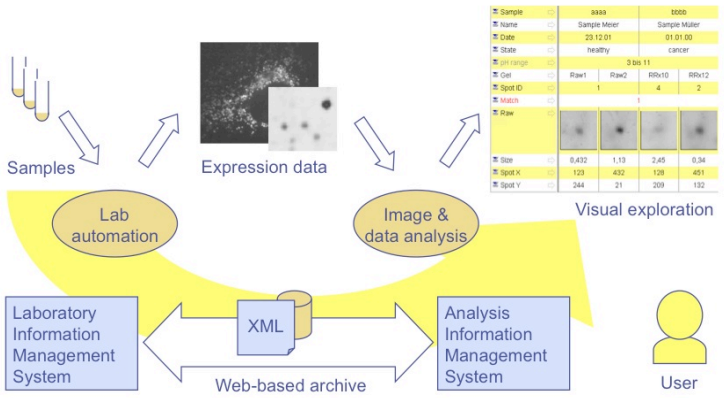
e. Drug Metabolism Cluster

From: Ross et al. Nature Genetics 24(3):227-35, 2000

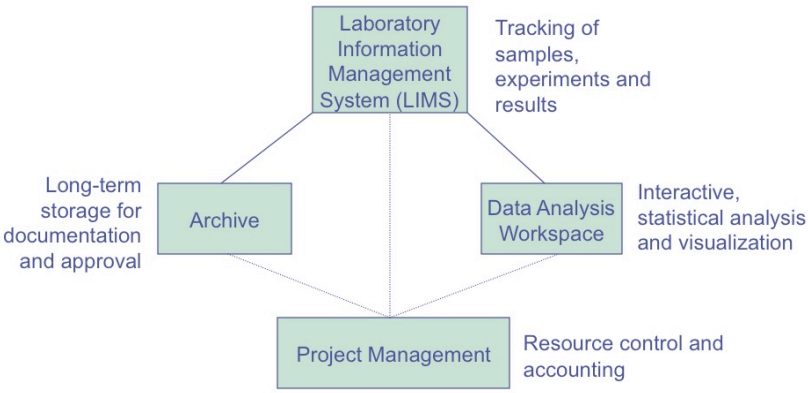
3.5

Data management

Discovery knowledge management



Data management overview



Raw Data Archiving

(1) Instrument data need to be archived
- FDA approval etc.
- Proof of authenticity needed (sigital signatures?)
- Long-term archive
  (ASCII formats like XML, or Adobe PDF)

(2) Laboratory information management system
- Track samples and experiments
- Link to archive
- Link to data analysis/project management

(3) Both should be based on a database service
- Commercial products available

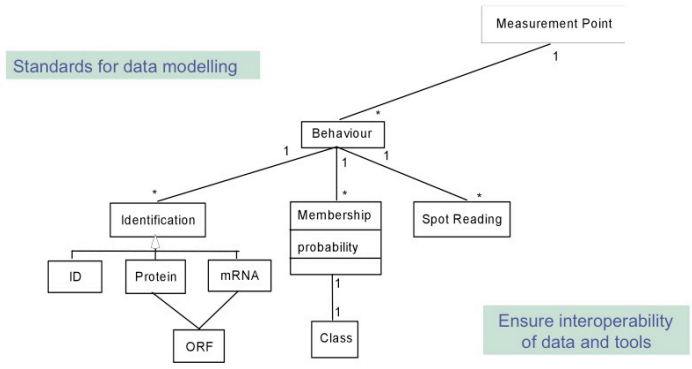
Data Design for Individual Studies

(1) Depends on experiment
- Screening (fixed protocol)
- Variation experiment (experiment-specific input parameters)

(2) Varies often
- New experimental designs
- New devices / protocols

(3) Data model needs to be planned during experiment design

Object Models



Databases vs. XML Integration

(1) Database
- Single schema, large amount of similarly-structured data
- Queries to find particular records (but overview difficult)
- Easy to update selected records
- Often monolithic and expensive

(2) Single files (Tabular ASCII, Excel, XML)
- More complex structure possible, medium amount of data
- Queries often clumsy (but browsing and overview possible)
- Selected update needs rewriting whole file
- Individual and flexible

XML Example

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE BinaryComparator PUBLIC "-//AVENTIS//DTD BinaryComparator 1.0//EN"
  "BinaryComparator.dtd">
<BinaryComparator
  numRefSpots="507"
  numTestSpots="487"
  refBioFilename="ctrl-output-bio50withBest.xml"
  refTechFilename="ctrl-output-tech50withBest.xml"
  testBioFilename="outBioIseRef50withBest.xml"
  testTechFilename="outTechIseRef50withBest.xml"
>

....

</BinaryComparator>
```

```
<Comparator
  refLabel="2"
  testLabel="2"
  refCentX="385"
  refCentY="63"
/>
<Comparator
  refLabel="3"
  testLabel="1"
  refCentX="364"
  refCentY="63"
/>
...
```

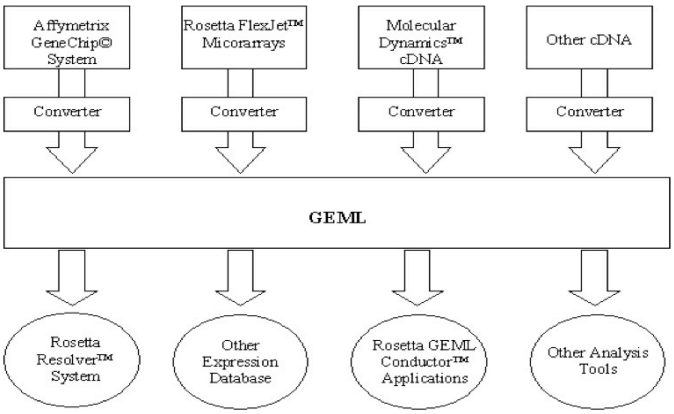

Modeling and Processing with XML

(1) XML files may contain sections of different type and structure
- Nested
- Defined in a Document Type Definition (DTD)

(2) Processing can be defined for individual sections
- Component-oriented software development

(3) Sections can be extended later
- Similar to subclassing
- Old code ignores new data, new code can handle both

Gene Expression Markup Language



GEML document types

(1) Pattern files
- Project
- Chip layout
- Probes

(3) Profile files
- Signal
- Background
- Channel information (optical)
- Log ratio

XML for expression data management

GENE				
GENE_ID	*CONTIG_ID*	*CONTIG_START*	*CONTIG_END*	*CONTIG_STRAND*
GB2VN	NT_0106058.3	2354807	2360778	Complement
GB2VN32	NT_0106051.3	2308745	2321072	Complement

```
<gene_features>
   <gene_id>GBVN32</gene_id>
   <contig_id>NT_010651</contig_id>
   <contig_start>2354807</contig_start>
   <contig_end>2360778</contig_end>
   <contig_strand>Complement</contig_strand>
</gene_features>
```