Statistical Evaluation of Gene Expression Microarrays

Marcel Dettling

Seminar für Statistik, ETH Zürich

November 16, 2001

dettling@stat.math.ethz.ch

An Introduction to Molecular Biology (I)

→ All living organisms are made up of *cells*

→ In every cell we find a nucleus, containing the *chromosomes*

→ Chromosomes are large pieces of DNA which contain thousands of *genes*

→ A gene can be understood as “construction manual” for a single *protein*

Fact: Every cell in an organism has the same set of chromosomes and genes

But: Cells do have very distinct properties, think of skin, hair or blood cells

How can and does this happen?

An Introduction to Molecular Biology (II)

$\Rightarrow$ Answer: Variety of the cells is a result of differences in abundance and distribution of a cells proteins

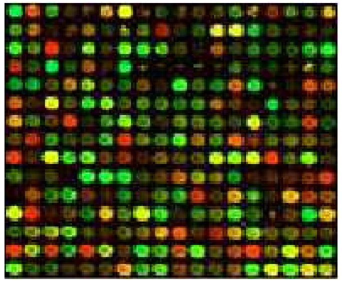
The central dogma of molecular biology

$$\boxed{\text{DNA} \longrightarrow \text{mRNA} \longrightarrow \text{protein} \longrightarrow \text{phenotype}}$$

$\Rightarrow$ Connection between the state of a cell and its protein and mRNA composition

Goal:	Distinction of phenotypes, e.g. normal/cancerous cells
Problem:	Abundance of proteins in a cell cannot be measured
Solution:	Microarrays allow quick and accurate measuring of relative mRNA abundance, also known as “gene activity”

The Microarray Technology (I)



Experimental Steps:

1) Attach various DNA target segments, each corresponding to a specific gene, in "overabundance" on distinct, tiny spots of a glass slide

2) Extract mRNA, e.g. from a healthy reference cell and a cancerous sample cell and mark it with red and green fluorescent labels

3) The fluorescent dyes are mixed and competitively hybridized on the glass slide. Each "construction manual" can only hybridize at a corresponding target gene.

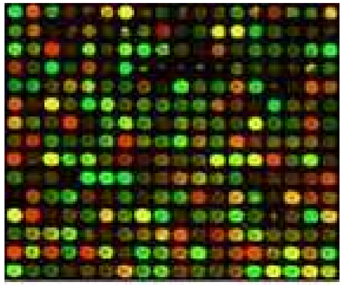
4) Laser scanning yields red and green intensities for each spot. Intensities are proportional to the amount of hybridization that occurred

The Microarray Technology (II)

5) Repeat steps 1) – 4) for n sample cells and one common reference cell

6) Apply background corrections, compute log-ratios of intensities and normalize

$\Rightarrow$ Microarray technology & preprocessing steps $\rightsquigarrow$ **gene expression matrix**



$$\rightsquigarrow \quad (x_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

- Contains data of n experiments, each showing expression levels of p genes.
- Often, we have an additional categorical response vector

$$(y_j) = (y_1, y_2, \ldots, y_n)$$

Microarrays: Purpose and Statistical Challenges

- *Functional genomics*

  Which genes are relevant for a particular disease? What is the function of a gene? How do the genes interact and which genes are co-regulated?

  $\rightsquigarrow$ Variable selection, clustering techniques

- *Clinical diagnostics*

  Precise and early diagnosis of diseases, e.g. cancer (sub)type. Crucial for successfull treatement, e.g. deciding between surgery, radio- and chemotherapy

  $\rightsquigarrow$ Classification techniques

Main statistical challenges with microarray gene expression data:

– Sample size n is usually small, only between 20 and 80

– Feature space contains many genes, p is usually between 2'000 and 8'000

– Features are often highly correlated

$\Rightarrow$ Need for specifically developed statistical procedures

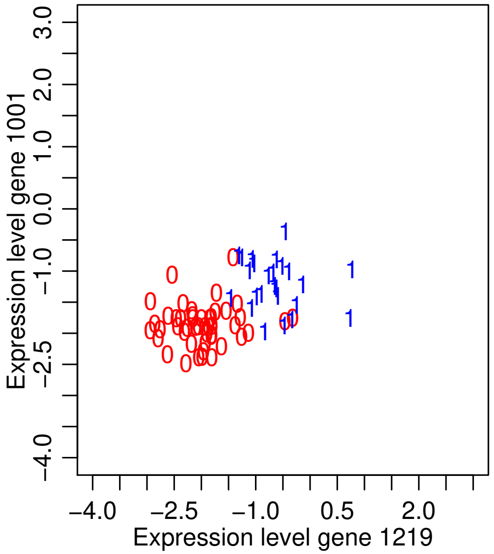
Gene Clustering for Functional Genomics

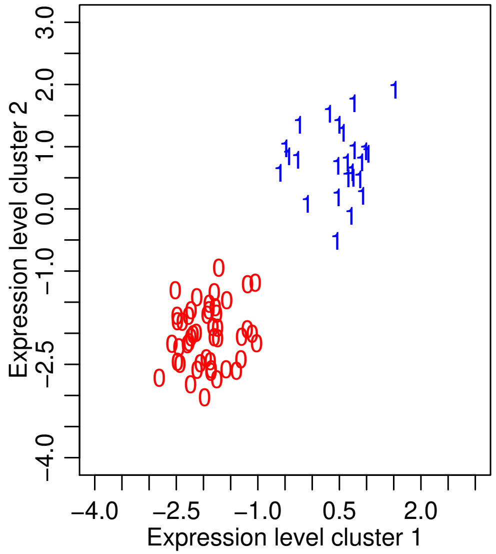
Previous work in gene clustering:

- *Unsupervised, hierarchical clustering*
  - building groups of genes that regulate similar across experiments
  - not precise enough
- *"Tree Harvesting" by Hastie et al. (2001)*
  - searching groups of similar genes that are of specific interest for classification
  - hierarchical clustering & subsequent selection of clusters by a response model
  - much more precise than hierachical clustering, but not the right way to do it!
- *"Supervised Clustering of Genes" by MD & P. Bühlmann (2001)*
  - forming groups of genes that optimally seperate different phenotypes
  - directly includes the response variables in the clustering process

Supervised Clustering of Genes

Projections (2-dim.) of gene expression data: *single genes* (left) and *clusters* (right)





Goal: Find clusters which make the separation of phenotypes as easy as possible

$\longrightarrow$ Need to define a “representative” cluster expression value

$\longrightarrow$ The problem: it’s hard to find the optimal clusters among thousands of genes

The Partitioning Problem

The basic stochastic model for microarray data with binary categorical response is

$$(X, Y) \text{ with } X \in \mathbb{R}^p \text{ and } Y \in \{0, 1\}$$

It's a fact: not individual, but groups of genes optimally discriminate the response Y

$$\mathrm{P}\left[Y = 1 | X = x\right] = f(\xi)$$

where $\xi = (\xi_1, \ldots, \xi_k)^T \in \mathbb{R}^k$ with $k \ll p$ contains the "representative" values of k gene clusters. We define them here as simple arithmetic means

$$\xi_j = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_i \ \in \mathbb{R}$$

$\longrightarrow$ Sensible choice since we anyway want to cluster similar genes

$\longrightarrow$ Optimizing more complex, general (non-)linear combinations is infeasible

The Partitioning Problem (II)

The problems we are facing:

– even by using arithmetic means, finding clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ is highly nontrivial

– exhaustive search is impracticable due to combinatorial complexity

– numerous “good” clusters may exist, and what is a “good” cluster?

A possible strategy:

– optimize an empirical objective function that quantitatively measures cluster quality

– aim for a quality measure which tells how accurately the samples are discriminated

– computation needs to be very efficient, since the function is “very often” evaluated

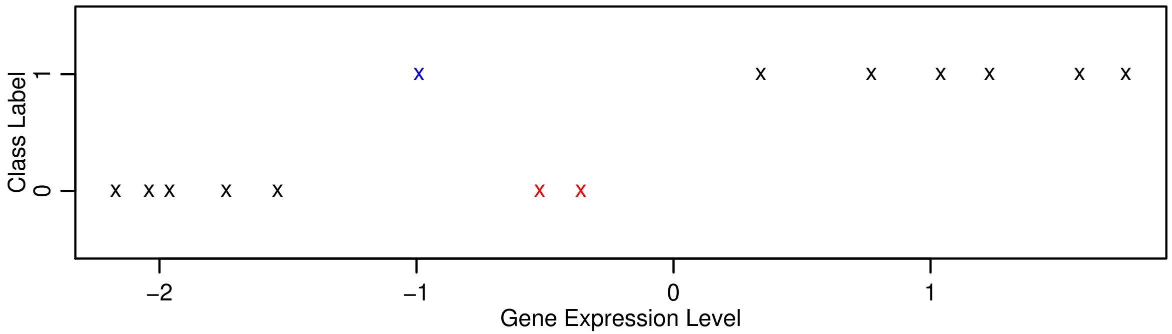
– working with a greedy algorithm makes things much easier

Implementation of the strategy:

$\longrightarrow$ **Clustering with score and margin**

Clustering with Score and Margin (I)

Basic Idea: Discrimination of the classes is easy if there is no “overlap”



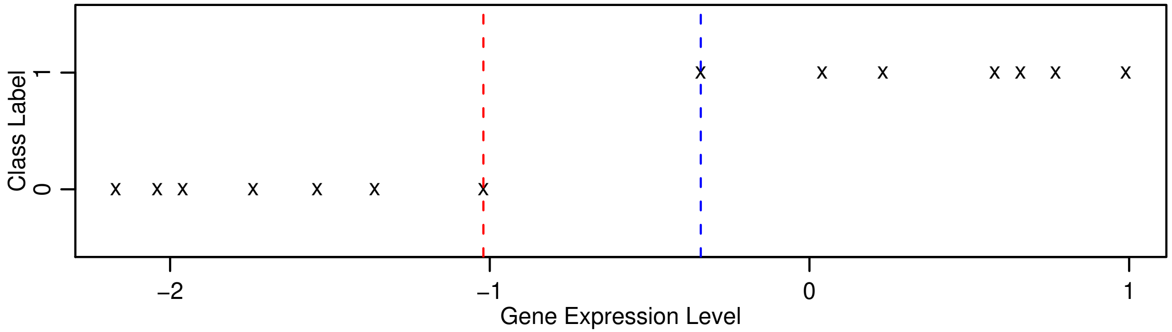
$\Rightarrow$ Severeness of the “overlap” for a particular gene or cluster expression vector $\xi_j = (\xi_{j1}, \ldots, \xi_{jn})$ is measured by the score function

$$\text{Score}(\xi_j) = \sum_{r \in \mathcal{N}_0} \sum_{s \in \mathcal{N}_1} 1_{[\xi_{js} - \xi_{jr} \leq 0]},$$

Equals the number of transpositions that are necessary for perfect separation

Clustering with Score and Margin (II)

Refinement by a second criterion: the margin function



$\Rightarrow$ Need margin function due to discreteness of the score function and to achieve uniqueness in decisions among genes/clusters with identical score

$$\text{Margin}(\xi_j) = \min_{s \in \mathcal{N}_1} (\xi_{js}) - \max_{r \in \mathcal{N}_0} (\xi_{jr}),$$

While clustering, scores are optimized with first, margins with second priority

The Clustering Algorithm (I)

1. **Starting genes**
   a) If given, compute their average expression profile
   b) If unknown, identify the best individual gene
2. **Forward search**
   a) Average the current cluster $\mathcal{C}_j$ with each individual gene $k = 1, \ldots, p$

$$\xi_j^{+k} = \frac{1}{|\mathcal{C}_j| + 1} \left(x_k + \sum_{i \in \mathcal{C}_j} x_i \right)$$

   b) Identify the winning gene k^*
   c) Accept and repeat until cluster quality worsens

The Clustering Algorithm (II)

3. **Backward search**

   a) Exclude each gene $k \in \mathcal{C}_j$ and average the remaining ones

$$\xi_j^{-k} = \frac{1}{|\mathcal{C}_j| - 1} \left(\sum_{i \in \mathcal{C}_j \setminus \{k\}} x_i \right)$$

   b) Identify the winning gene k^*

   c) Exclude and repeat until cluster quality worsens

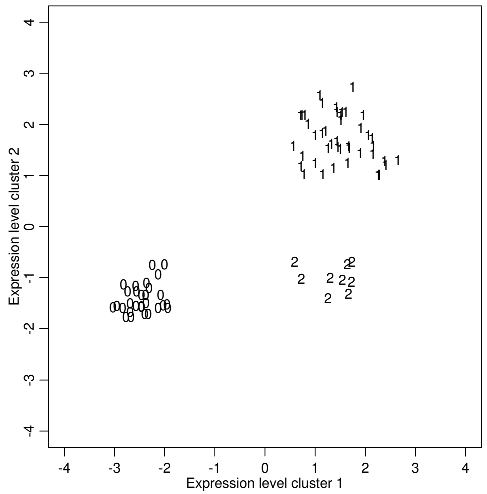
4. **Stagewise clustering**

   Repeat forward and backward stages until convergence

5. **Multiple clustering**

   Discard the clustered genes and restart the algorithm at step 1

Results from the Supervised Clustering Algorithm



From an empirical study:

– on 3 different microarray datasets with expression profiles of human cancer cells

Typical cluster output:

– cluster size is 3–9 genes with very clear separation: zero scores and large margins

Relevance and stability:

– clusters are “resonably stable” and certainly more than a noise artefact

More Results and Conclusions

Predictive potential of the clusters:

- error-free classification of training data already with simple classifiers
- perfect classification of test data if clusters are from the entire dataset
- very competitive classification performance for proper outsample predictions
- example: 27% error rate in a dataset with 8 human cancer types, compared to a best results of approximately 35% in the literature

Conclusions:

- quick and fairly easy algorithm for supervised clustering of genes
- better classification performance than with single genes
  $\Rightarrow$ "it is worth the effort"