STATISTICAL EVALUATION OF GENE EXPRESSION MICROARRAYS

Marcel Dettling

Seminar für Statistik, ETH Zürich

March 6, 2002

dettling@stat.math.ethz.ch

http://stat.ethz.ch/~dettling

A Simplified 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*

To first approximation:	Every cell in an organism has the same set of chromosomes and genes
On the other hand:	Cells do have very distinct properties, think of skin, hair or blood cells

How can and does this happen?

A Simplified 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

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

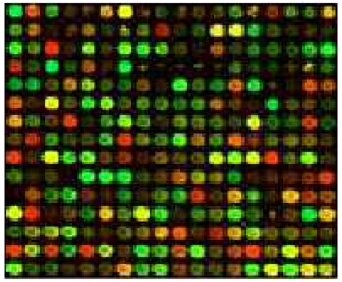
$\Rightarrow$ Connection between the state of a cell and its protein and mRNA composition

Goal: Distinction of phenotypes, e.g. (sub-)types of cancer

Problem: Abundance of proteins in a cell is very difficult to measure

Solution: Microarrays allow quick and accurate measuring of (relative) mRNA abundance, also known as “gene activity” or “gene expression”

The Microarray Technology



Goal: Measuring the abundance of mRNA transcripts for each gene (DNA segment)

1) Attach DNA segments (*probes*), each corresponding to a part of a specific gene in “overabundance” on distinct, tiny spots of a glass slide.

2) Extract mRNA from a tissue of interest and mark it with fluorescent labels.

3) Let the fluorescent dye hybridize on the microarray. Each mRNA transcript preferably attaches to the DNA segment it originated from.

4) Laser scanning yields intensities for each spot, which are proportional to the amount of hybridization that occurred.

cDNA Arrays vs. GeneChips

2 main microarray technologies:

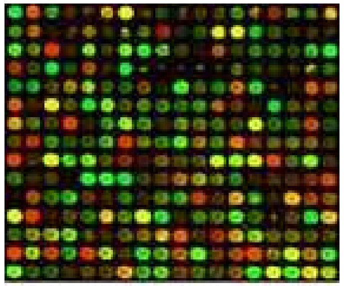
cDNA Microarrays	vs.	**Affymetrix Gene Chips**
– open, more flexible		– proprietary, commercial
– 1 longer, single probe		– 20 shorter paired probes
– competitive hybridization with reference tissue		– only experimental tissue mismatch probe as control
– expressions are relative (log-ratios)		– expressions are averaged differences: $\sum$(PM-MM)

The Microarray Technology (II)

5) Repeat steps 1) – 4) for n experimental tissues of interest

6) Apply background corrections, carry out normalization and data transformation

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



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

- Contains data of n experiments, each showing expression levels of p genes.
- We focus on "**supervised problems**" with additional categorical response vector

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

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 successful treatment, e.g. deciding between surgery, radio- and chemotherapy

  $\rightsquigarrow$ Classification techniques, pattern recognition

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

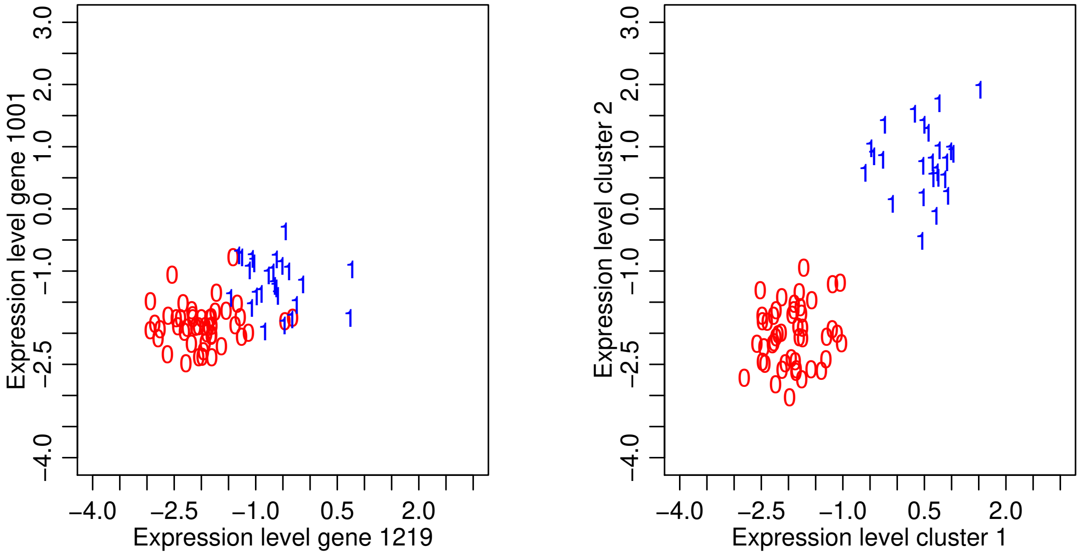
Gene Clustering for Functional Genomics

Previous work in gene clustering:

- *Unsupervised, hierarchical clustering*
  - building groups of genes that regulate similar across experiments
  - not precise enough and yielding too many clusters
- *"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 hierarchical clustering, but not the right way to do it!
- *"Supervised Clustering of Genes" by MD & P. Bühlmann (2001)*
  - forming a few groups of genes that optimally separate 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 Stochastic Framework

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

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

where $x_{\mathcal{C}} = (x_{\mathcal{C}_1}, \ldots, x_{\mathcal{C}_k})^T \in \mathbb{R}^k$ with $k \ll p$ contains the "representative" values of k gene clusters. We define them here as simple arithmetic means

$$x_{\mathcal{C}_\ell} = \frac{1}{|\mathcal{C}_\ell|} \sum_{g \in \mathcal{C}_\ell} x_g \ \in \mathbb{R}$$

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

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

The Partitioning Problem

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

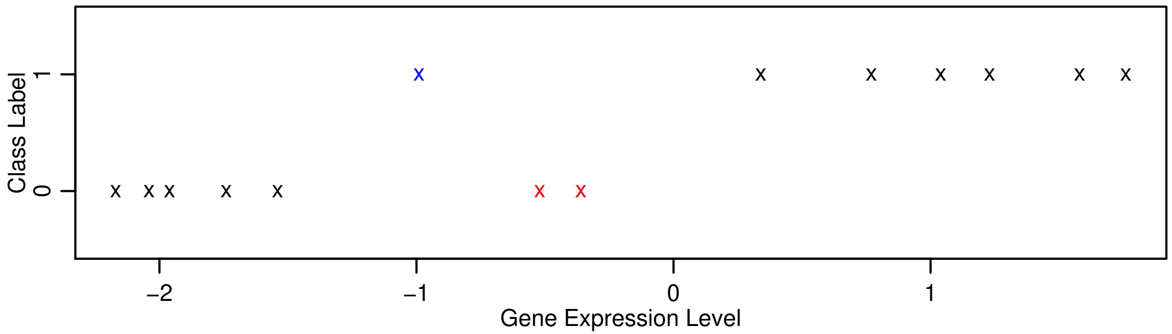
– working with a greedy algorithm makes things much easier

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”



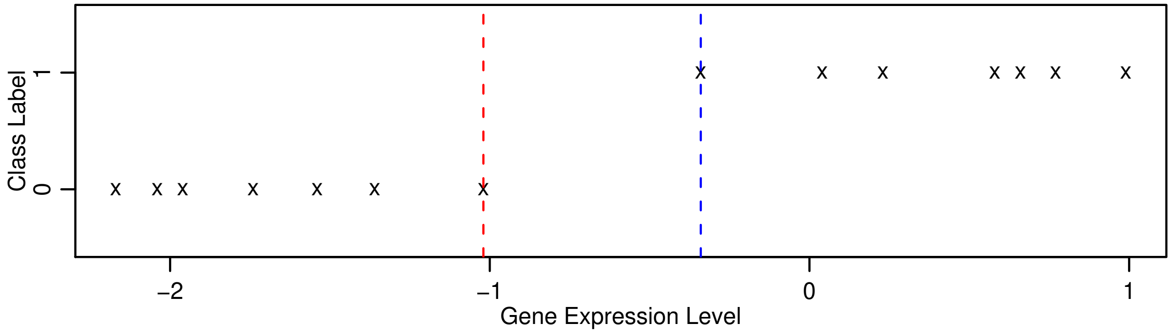
$\Rightarrow$ Severeness of the “overlap” for a particular gene or cluster expression vector $x_{\mathcal{C}_\ell} = (x_{\mathcal{C}_{\ell 1}}, \ldots, x_{\mathcal{C}_{\ell n}})$ across experiments is measured by the score function

$$\text{Score}(x_{\mathcal{C}_\ell}) = \sum_{r \in \mathcal{N}_0} \sum_{s \in \mathcal{N}_1} 1_{[x_{\mathcal{C}_{\ell s}} - x_{\mathcal{C}_{\ell r}} \leq 0]},$$

Equivalent to the test statistic of Wilcoxon’s two sample test

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

$$\text{Margin}(x_{\mathcal{C}_\ell}) = \min_{s \in \mathcal{N}_1}(x_{\mathcal{C}_{\ell s}}) - \max_{r \in \mathcal{N}_0}(x_{\mathcal{C}_{\ell r}}),$$

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 $\mathcal{C}_\ell$ with each individual gene $g = 1, \ldots, p$

   $$x_{\mathcal{C}_\ell}^{+g} = \frac{1}{|\mathcal{C}_\ell| + 1} \left(x_g + \sum_{h \in \mathcal{C}_\ell} x_h \right)$$

   b) Identify the winning gene g^* such that the new cluster average $x_{\mathcal{C}_\ell}^{+g^*}$ has best score and margin values

   c) Accept and repeat until cluster quality worsens

The Clustering Algorithm (II)

3. **Backward search**

   a) Exclude each gene $g \in \mathcal{C}_\ell$ and average the remaining ones

$$x_{\mathcal{C}_\ell}^{-g} = \frac{1}{|\mathcal{C}_\ell| - 1} \left(\sum_{h \in \mathcal{C}_\ell \setminus \{g\}} x_h \right)$$

   b) Identify the winning gene g^* such that $x_{\mathcal{C}_\ell}^{-g^*}$ has best score/margin values

   c) Exclude and repeat until cluster quality worsens

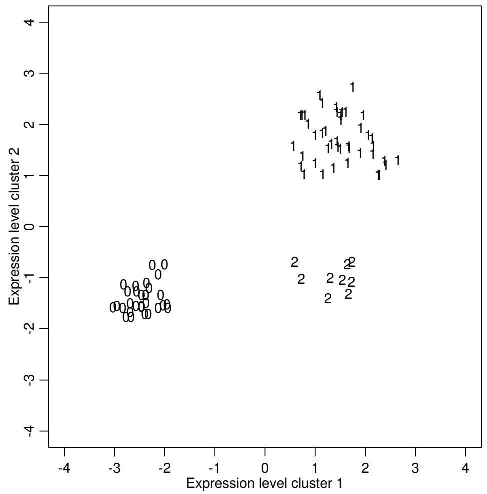
4. **Stagewise clustering**

   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 6 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 “reasonably stable” and certainly more than a noise artifact

Tumor Classification for Clinical Diagnostics

Situation: We are given a preprocessed gene expression matrix (x_{ij}) containing expression values of p genes for n different individuals which are equipped with an additional class label y, describing a cancer phenotype.

Task: For an expression profile $\vec{x} = (x_1, \dots, x_p)^T$ of a new patient, we have to predict the corresponding unknown class label y . This can be very important for cancer treatment, i.e. deciding between surgery, chemotherapy or radiotherapy.

Stochastically: Based on the n training data pairs $(x_1, y_1), \dots, (x_n, y_n)$ with known class label try to accurately estimate a classifier function

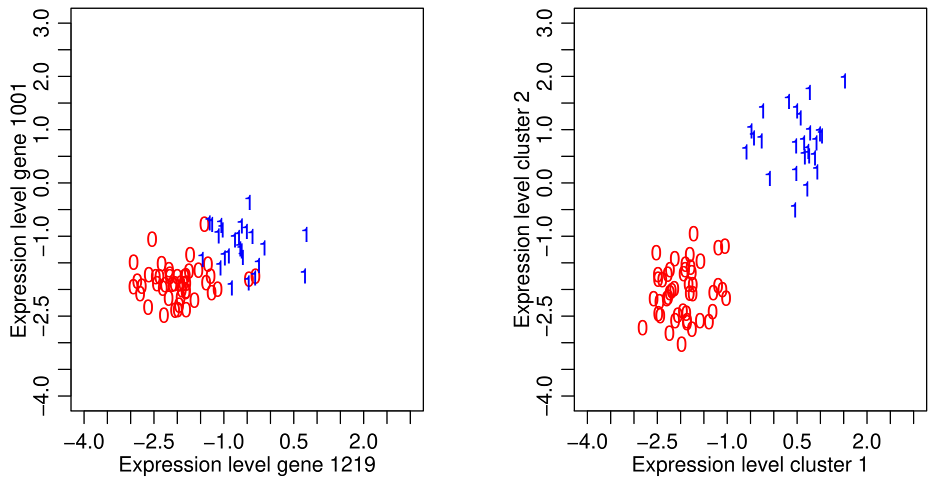
$$C(\cdot) : X \mapsto C(X) \in \{0, 1\},$$

which predicts the class label of a new individual, based on its expression profile

Tumor Classification with Gene Clusters

Idea: Use cluster expression profiles as input for the classifier function $C(\cdot)$

Classification based on best two *single genes* (left) and on best two *clusters* (right)



– classification seems much easier with clusters as with single genes

– but: the best two clusters average here the values of 12 genes

– accurate classification with single genes is possible in higher dimensional spaces

Classification Potential of the Clusters

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
- often better than sophisticated classification based on single genes
- setting new benchmark performance on some of the datasets.

$\Rightarrow$ *Classification with gene clusters is among the best which can be done*

Conclusions:

- already simple classifiers perform well with clusters due to low dimensionality
- averaging of single genes reduces variability and instability across experiments, thus improves classification performance

Tumor Classification with Boosting

Boosting:

- is a machine learning technique which is based on single genes as input
- has been successful on a wide variety of especially high-dimensional problems

$$\begin{array}{ccccccc} \boxed{\text{training sample}} & \longrightarrow & \boxed{\text{weighted sample}} & \longrightarrow & \ldots & \longrightarrow & \boxed{\text{weighted sample}} \\ \downarrow & & \downarrow & & & & \downarrow \\ C_1(X) & & C_2(X) & & & & C_M(X) \end{array}$$

$$C(X) = sign\left(\sum_{m=1}^{M} \alpha_m C_m(X)\right)$$

- aims at producing a powerful classifier from a sequence of *base* classifiers $C_m(\cdot)$
- gives higher weights to instances which were difficult to classify at earlier stages

Advantages of Boosting

Main advantages:

- *Provides class membership probabilities* $\mathrm{P}[Y = 1|X = x]$

  – natural quantization of uncertainty for class label assignment

  – allows construction of classifiers respecting non-equal misclassification costs

- *Suitable for high-dimensional problems*

  – can cope with large amounts of data and has the ability to do feature selection

  – resistance against overfitting due to the stagewise nature of the fitting process

- *Implementation is easy*

  – does not require much fine tuning and non-linear optimization to be successful

  – needs a few adaptions to be a precise, fully automatic classifier for gene data

$\Rightarrow$ *"How to Use Boosting for Tumor Classification with Gene Expression Data" by MD & P. Bühlmann (2002)*

Tailoring Boosting for Gene Expression Data (I)

1) *Gene preselection*

– can be important to improve classification performance

– eases computation and reduces the amount of noise

– use score function (Wilcoxon's test statistic) to identify best individual genes

– number of genes is determined by cross validation or bootstrap methods

– best results are usually obtained using between 50 and 200 genes

2) *The base classifier* $C_m(\cdot)$

– most popular are trees; we further restrict to 2-terminal-node-trees (stumps)

– larger trees, allowing for non-additive decision boundaries, were not necessary

3) *Data reweighting and aggregation weights*

– is the "art of the business", can be motivated by a functional gradient descent

– use of a robust boosting procedure due to data inhomogeneity and mislabeling

Tailoring Boosting for Gene Expression Data (II)

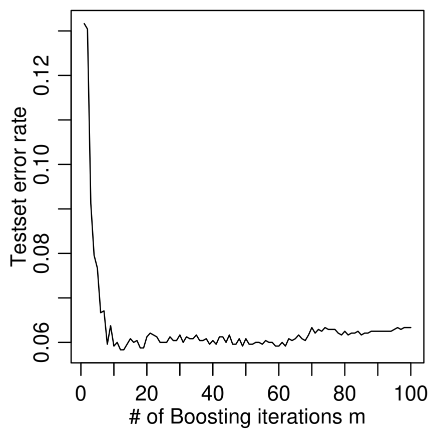
4) *Choice of the stopping parameter*

– data-driven estimation to avoid overfitting and to make boosting fully automatic

5) *Special treatment of multi-categorical problems*

– reduction to binary problems allows for different model complexity in each class

– improvement is most often substantial (40-70%) in various problems



Comparison of Results and Conclusions

	Leukemia	Colon	Estrogen	Nodal	Lymphoma	NCI 60
Gene Clusters	**0.00%**	16.13%	**2.04%**	28.57%	**0.00%**	24.59%
Boosting	2.78%	**12.90%**	**2.04%**	**16.33%**	1.61%	**22.95%**
Benchmark	1.39%	14.52%	14.29%	30.61%	4.84%	27.7%

$\longrightarrow$ both gene clusters and boosting outperform the benchmark

$\longrightarrow$ similar results for gene cluster and boosting

Gene Clusters	vs.	**Boosting**
better on easy problems		better on difficult problems
better for functional genomics		better for clinical diagnostics