Classification of microarray samples

Tim Beißbarth

Mini-Group Meeting

8.7.2002

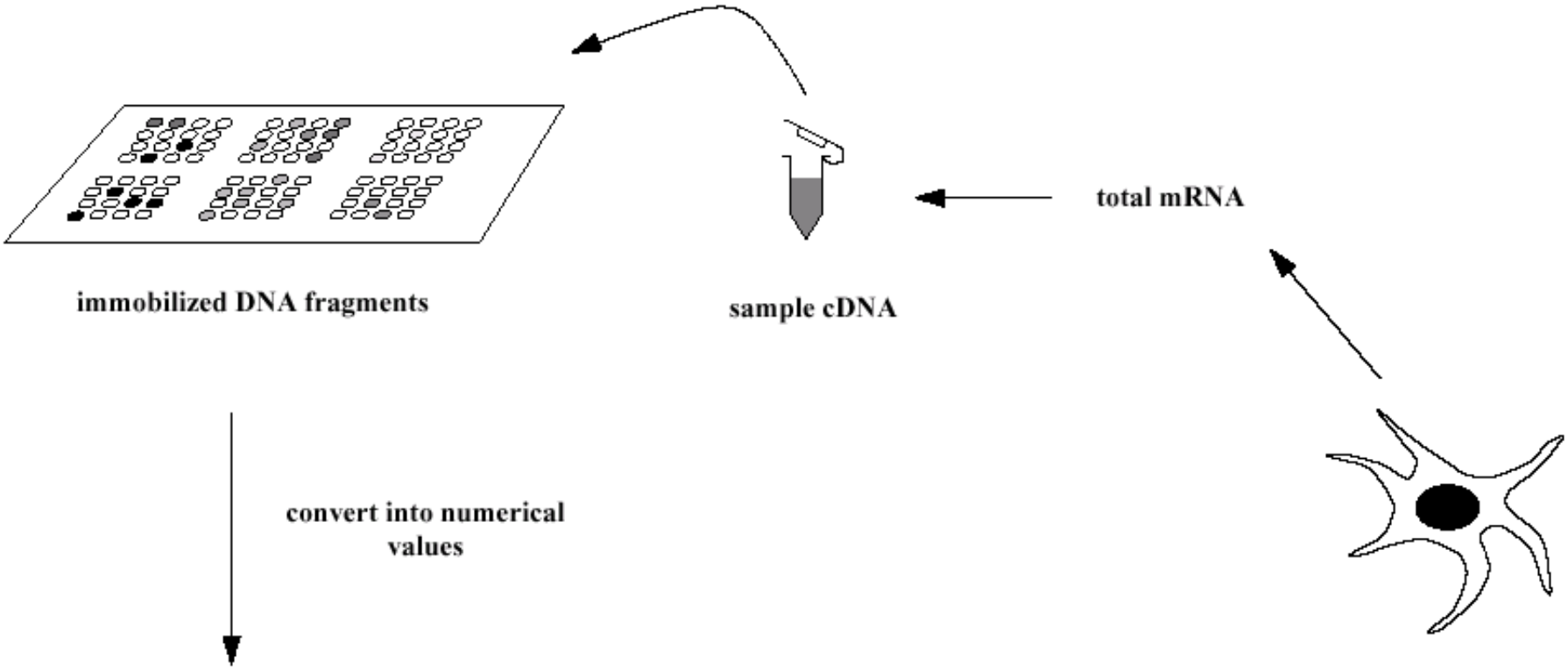
Papers in PNAS May 2002

- *Diagnosis of multiple cancer types by shrunken centroids of gene expression*

  Robert Tibshirani,Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu
- *Selection bias in gene extraction on the basis of microarray gene-expression data*

  Christphe Ambroise, and Geoffrey J. McLachlan

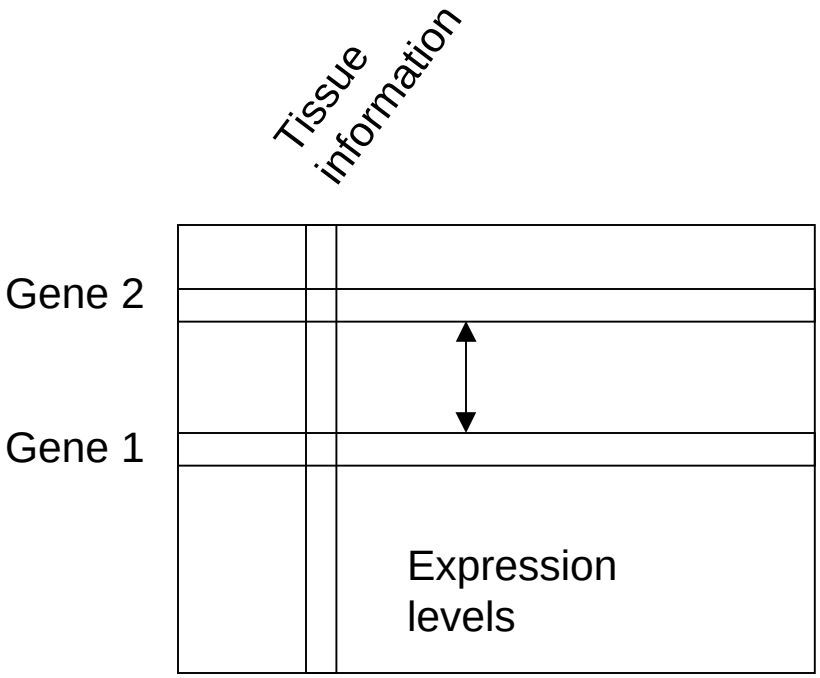
DNA Microarray Hybridization



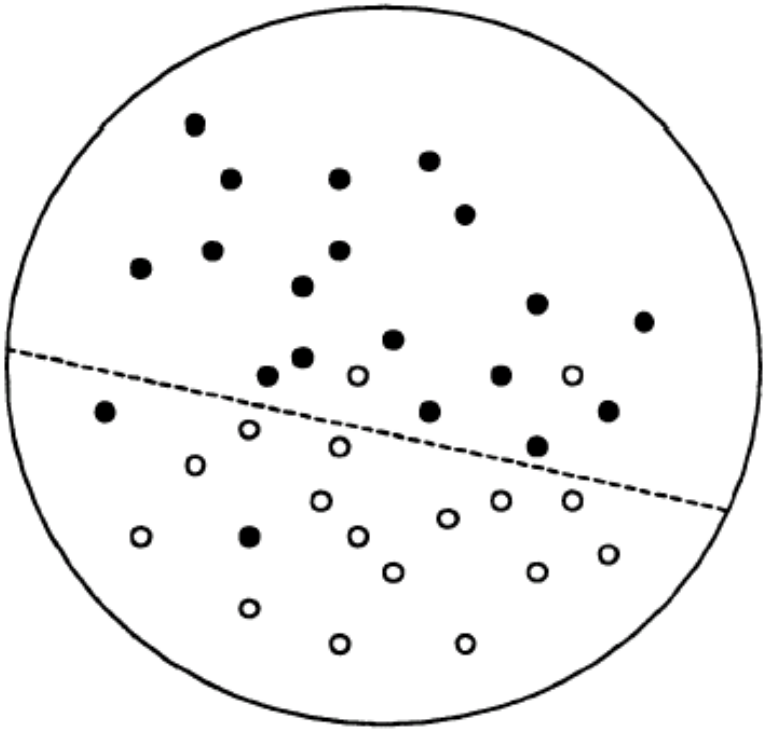
gene 1	14,243
gene 2	5,323
gene 3	10,300
gene 4	1,007
gene 5	100,232

Tables of Expression Data

Table of expression levels:



The Classification Problem



Classification Methods:

Support Vector Machines, Neural Networks, Fishers linear descriminant, etc.

Diagnosis of multiple cancer types by shrunken centroids of gene expression

Robert Tibshirani[†‡], Trevor Hastie[§], Balasubramanian Narasimhan[§], and Gilbert Chu[¶]

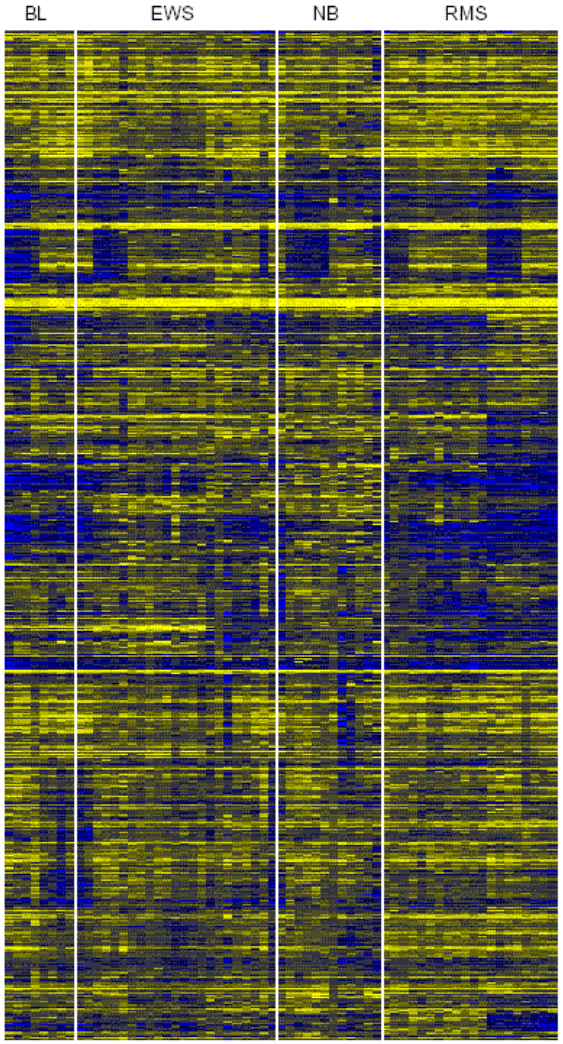
Departments of [†]Health, Research and Policy, and Statistics, [§]Statistics and Health, Research and Policy, and [¶]Medicine and Biochemistry, Stanford University, Stanford, CA 94305

Communicated by Bradley Efron, Stanford University, Stanford, CA, February 19, 2002 (received for review October 10, 2001)

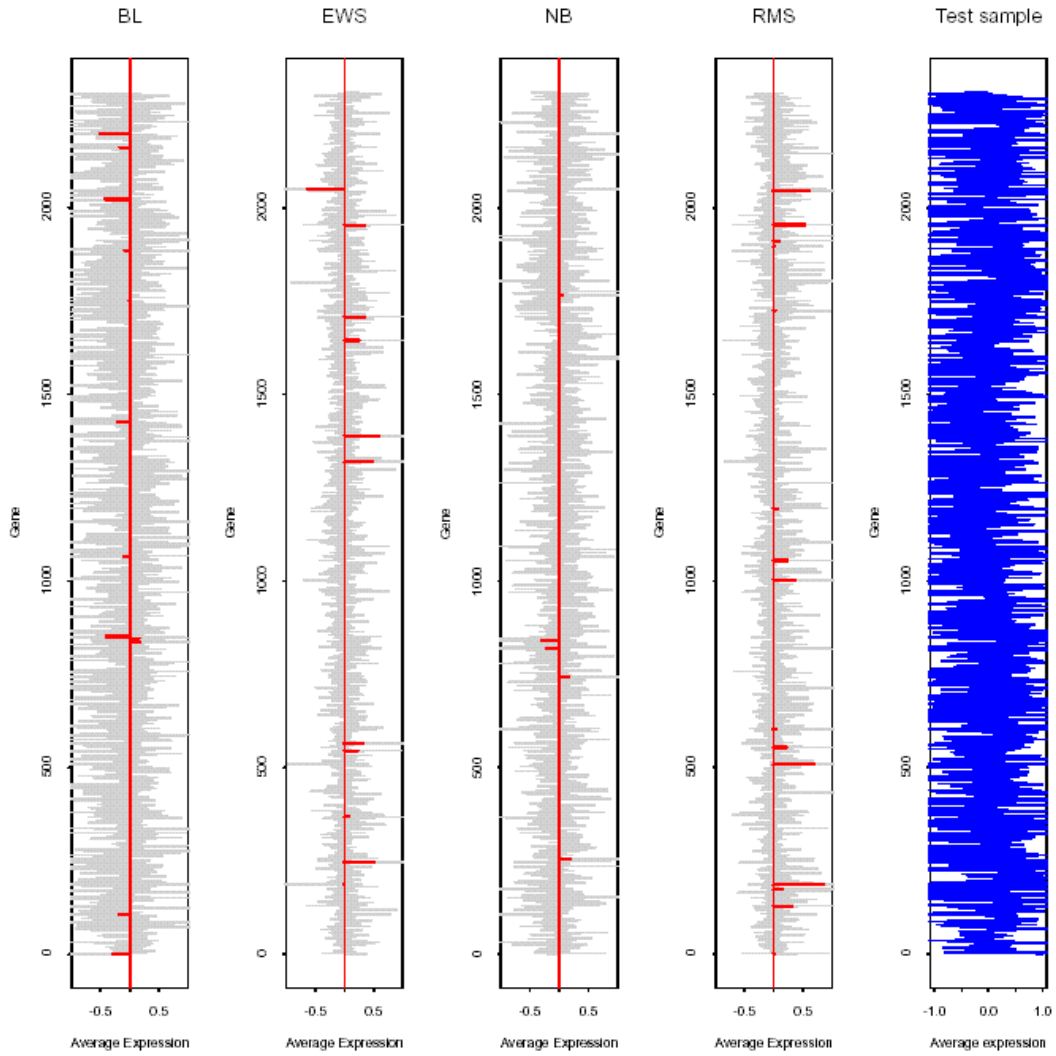
Example: small round blue cell tumors; Khan et al, Nature Medicine, 2001

- Tumors classified as BL (Burkitt lymphoma), EWS (Ewing), NB (neuroblastoma) and RMS (rhabdomyosarcoma).
- There are 63 training samples and 25 test samples, although five of the latter were not SRBCTs. 2308 genes
- Khan et al report zero training and test errors, using a complex neural network model. Decided that 96 genes were "important".
- Upon close examination, network is linear. It's essentially extracting linear principal components, and classifying in their subspace.
- But even principal components is unnecessarily complicated for this problem!

Khan data



Class centroids



Nearest Shrunken Centroids

Idea: shrink each class centroid towards the overall centroid. First normalize by the within-class standard deviation for each gene.

Details

- Let x_{ij} be the expression for genes $i = 1, 2, \ldots p$ and samples $j = 1, 2, \ldots n$.
- We have classes $1, 2, \ldots K$, and let C_k be indices of the n_k samples in class k.
- The ith component of the centroid for class k is $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$, the mean expression value in class k for gene i; the ith component of the overall centroid is $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$.

- Let

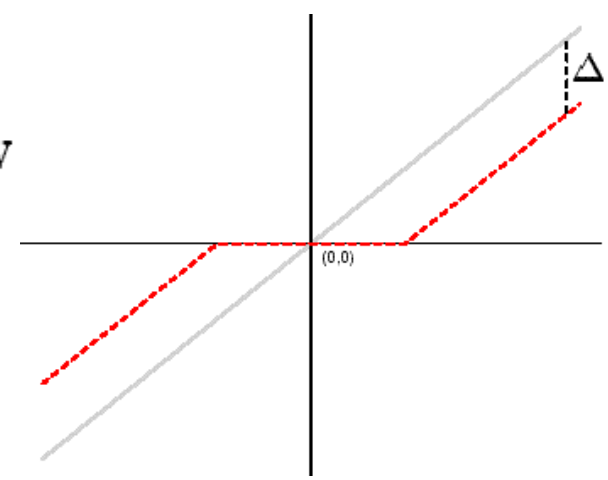
$$d_{ik} = (\bar{x}_{ik} - \bar{x}_i)/s_i$$

where s_i is the pooled within-class standard deviation for gene i:

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{i \in C_k} (x_{ij} - \bar{x}_{ik})^2.$$

- Shrink each d_{ik} towards zero, giving d'_{ik} and new shrunken centroids or prototypes

$$\bar{x}'_{ik} = \bar{x}_i + s_i d'_{ik}$$



- The shrinkage is by soft-thresholding:
- Choose Δ by cross-validation.

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$$

K-Fold Cross-Validation

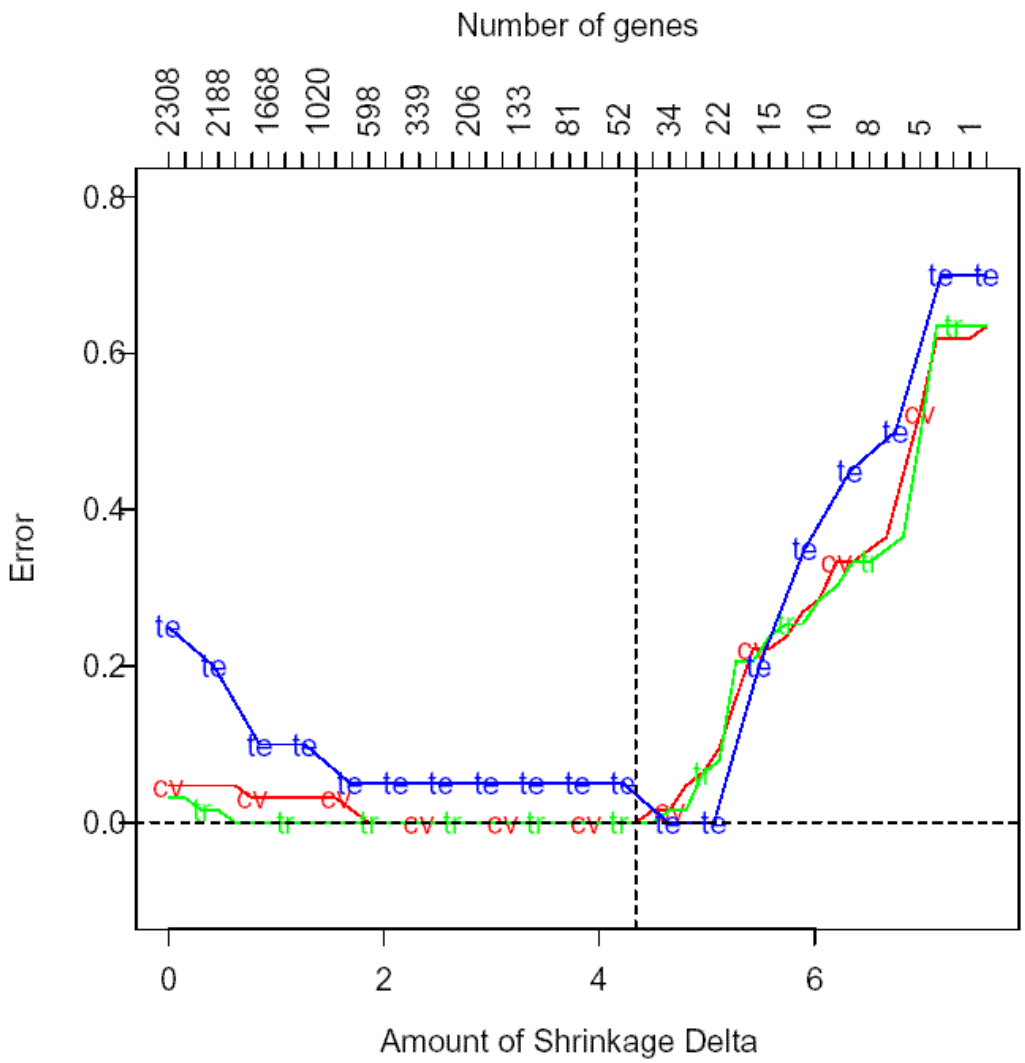
Primary method for estimating a tuning parameter λ.
Divide the data into K roughly equal parts.

1	2	3	4	5
Test	Train	Train	Train	Train

- for each $k = 1, 2, \ldots K$, fit the model with parameter λ to the other $K - 1$ parts, and compute its error in predicting the kth part. Average this error over the K parts to give the estimate $CV(\lambda)$.
- do this for many values of λ. Draw the curve $CV(\lambda)$ and choose the value of λ that makes $CV(\lambda)$ smallest.

Typically we use $K = 5$ or 10.

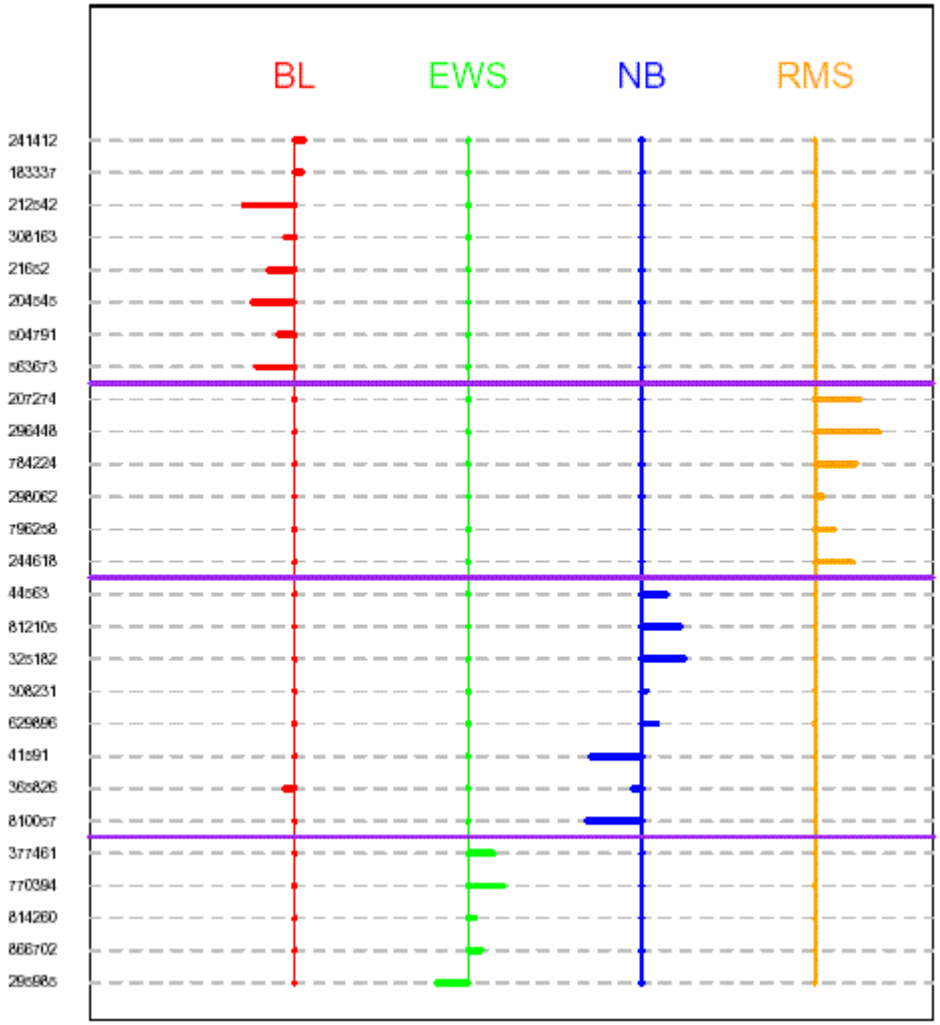
Results


Number of genes
2308
2188
1668
1020
598
339
206
133
81
52
34
22
15
10
8
5
1
Error
0.8
0.6
0.4
0.2
0.0
0
2
4
6
Amount of Shrinkage Delta
te
cv
tr

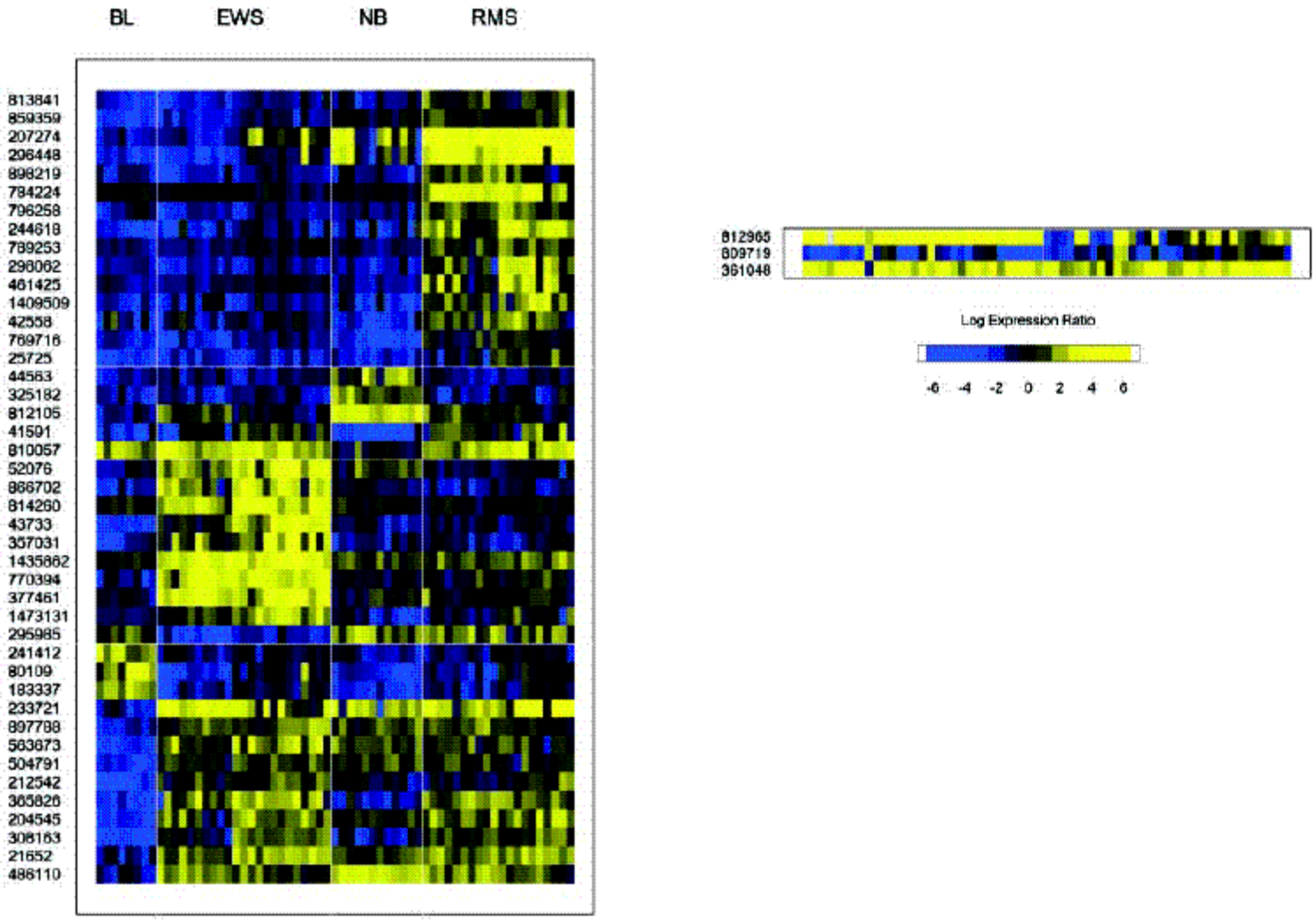
Advantages

- Simple, includes nearest centroid classifier as a special case.
- Thresholding denoises large effects, and sets small ones to zero, thereby selecting genes.
- with more than two classes, method can select different genes, and different numbers of genes for each class.

The genes that matter

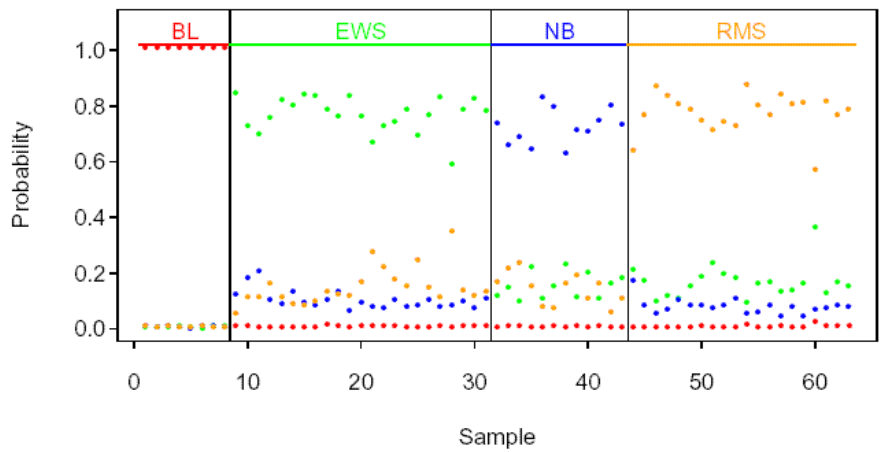


Heat map of the chosen 43 genes.

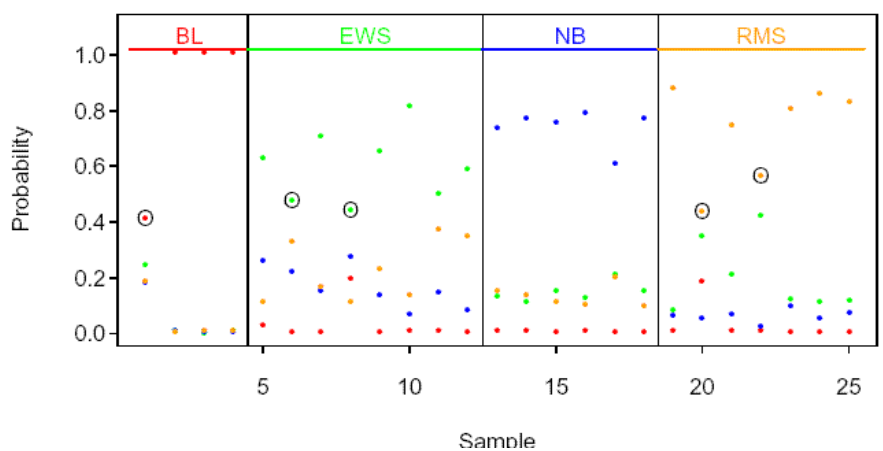


Estimated Class Probabilities

Training Data



Test Data



Selection bias in gene extraction on the basis of microarray gene-expression data

Christophe Ambroise[†] and Geoffrey J. McLachlan[‡§]

[†]Laboratoire Heudiasyc, Unité Mixte de Recherche/Centre National de la Recherche Scientifique 6599, 60200 Compiègne, France; and [‡]Department of Mathematics, University of Queensland, Brisbane 4072, Australia

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 21, 2002 (received for review February 20, 2002)

Steps in classification

- Feature selection
- Training a classification rule

Problem:

- For microarray data there are many more features (genes) than there are training samples and conditions to be classified.
- Therefore usually a set of features which discriminates the conditions perfectly can be found (overfitting)

Feature selection

- Criterion is independent of the prediction rule (filter approach)
- Criterion depends on the prediction rule (wrapper approach)

Goal:

- Feature set must not be to small, as this will produce a large bias towards the training set.
- Feature set must not be to large, as this will include noise which does not have any discriminatory power.

Methods to evaluate classification

- **Split Training-Set vs. Test-Set:**
  Disadvantage: Looses a lot of training data.
- **M-fold cross-validation:**
  Divide in M subsets, Train on M-1 subsets, Test on 1 subset
  Do this M-times and calculate mean error
  Special case: m=n, leave-one out cross-validation
- **Bootstrap**

Important!!!

- **Feature selection needs to be part of the testing and may not be performed on the complete data set. Otherwise a selection bias is introduced.**

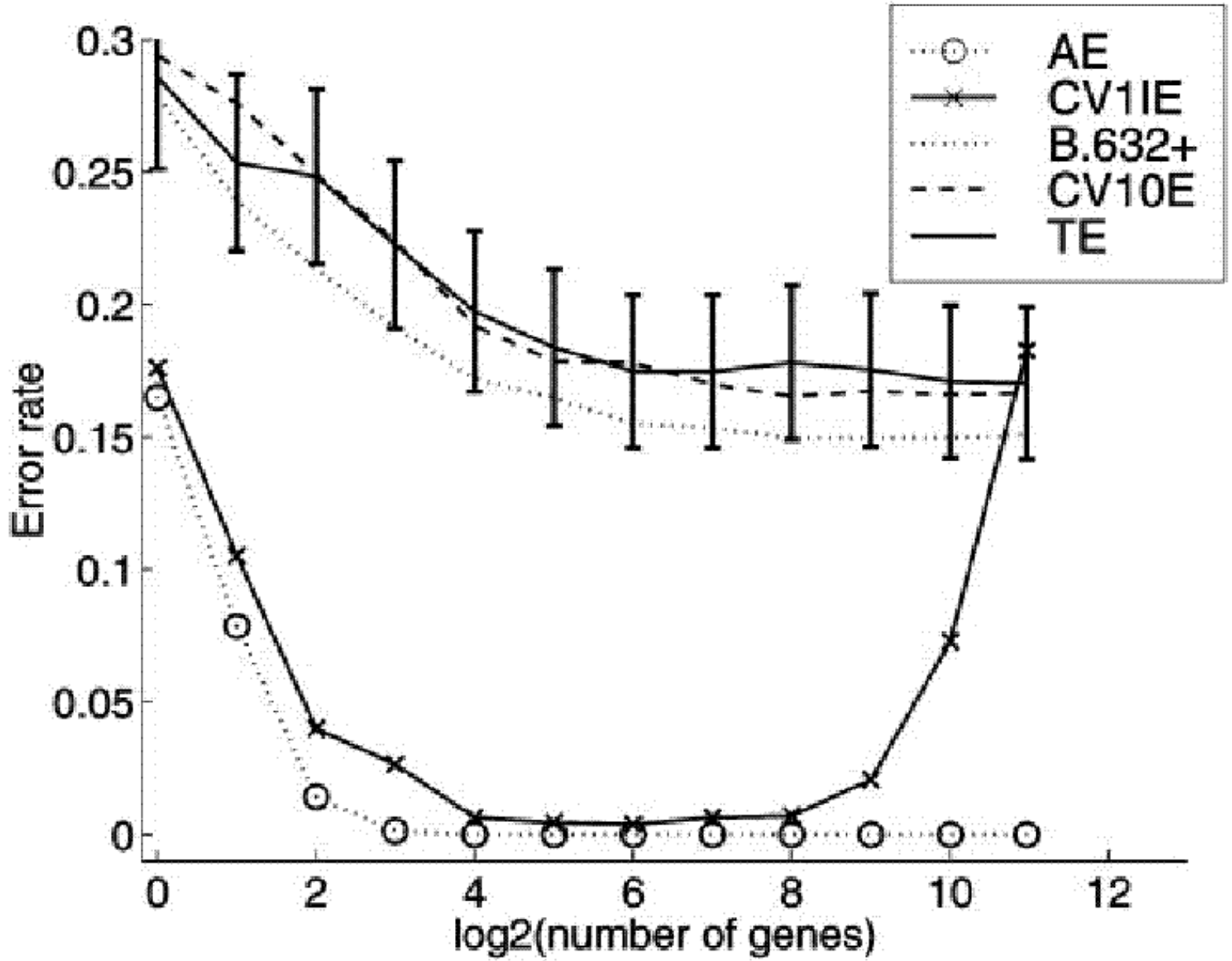
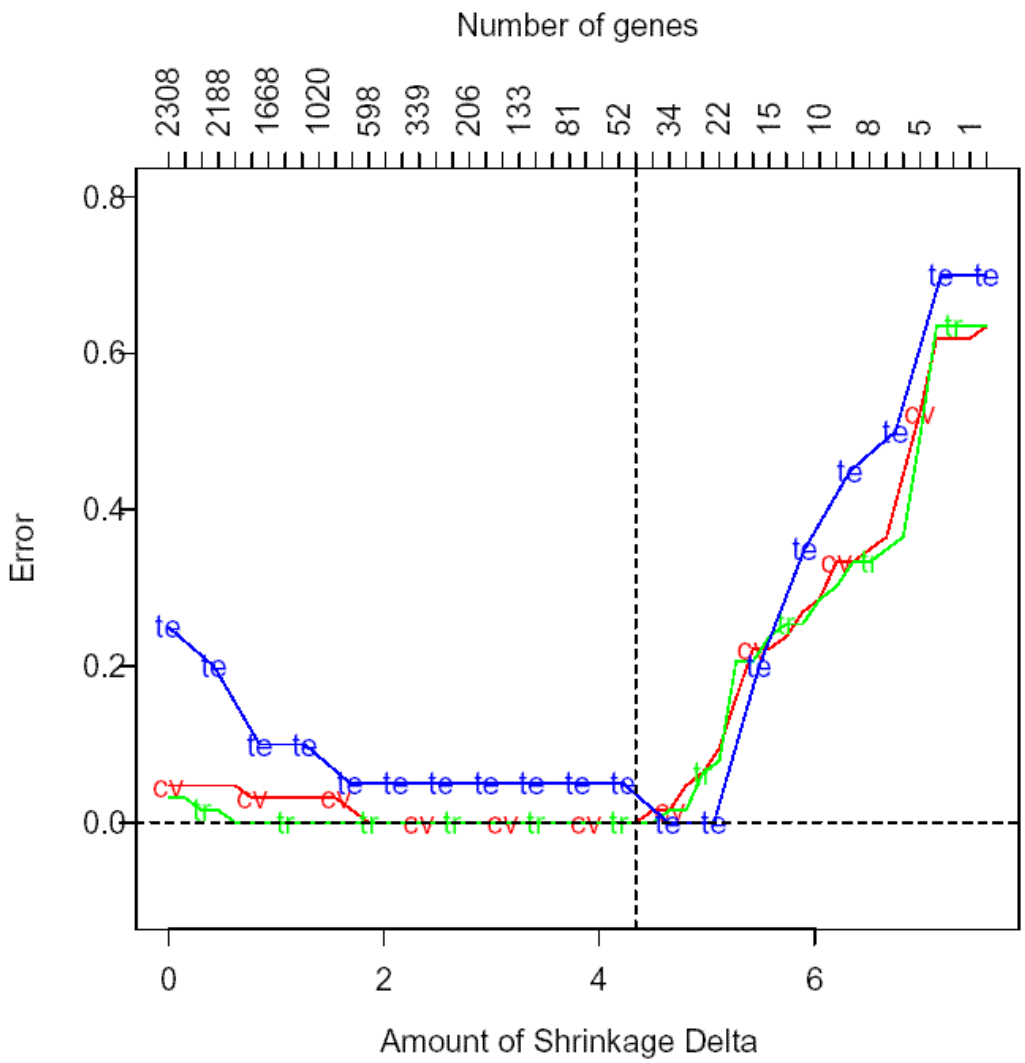


Fig. 1. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each. TE, test error.

Tibshirani et al, PNAS, 2002



Conclusions

- One needs to be very carefull when interpreting test and cross-validation results.
- The feature selection method needs to be included in the testing.
- 10-fold cross-validation or bootstrap with external feature selection.
- Feature selection has more influence on the classification result than the classification method used.

The End