Comparison of discrimination methods for the classification of tumors using gene expression data

Sandrine Dudoit[1], Jane Fridlyand[2] and Terry Speed[2,3]

1. Mathematical Sciences Research Institute, Berkeley
2. Department of Statistics, UC Berkeley
3. Walter and Eliza Hall Institute, Melbourne

Outline

1. Genetic background
2. Tumor classification using microarray experiments
3. Discrimination methods
4. Datasets
5. Results
6. Conclusions

Genetic background

A **gene** consists of a segment of **DNA** which codes for a particular **protein**.

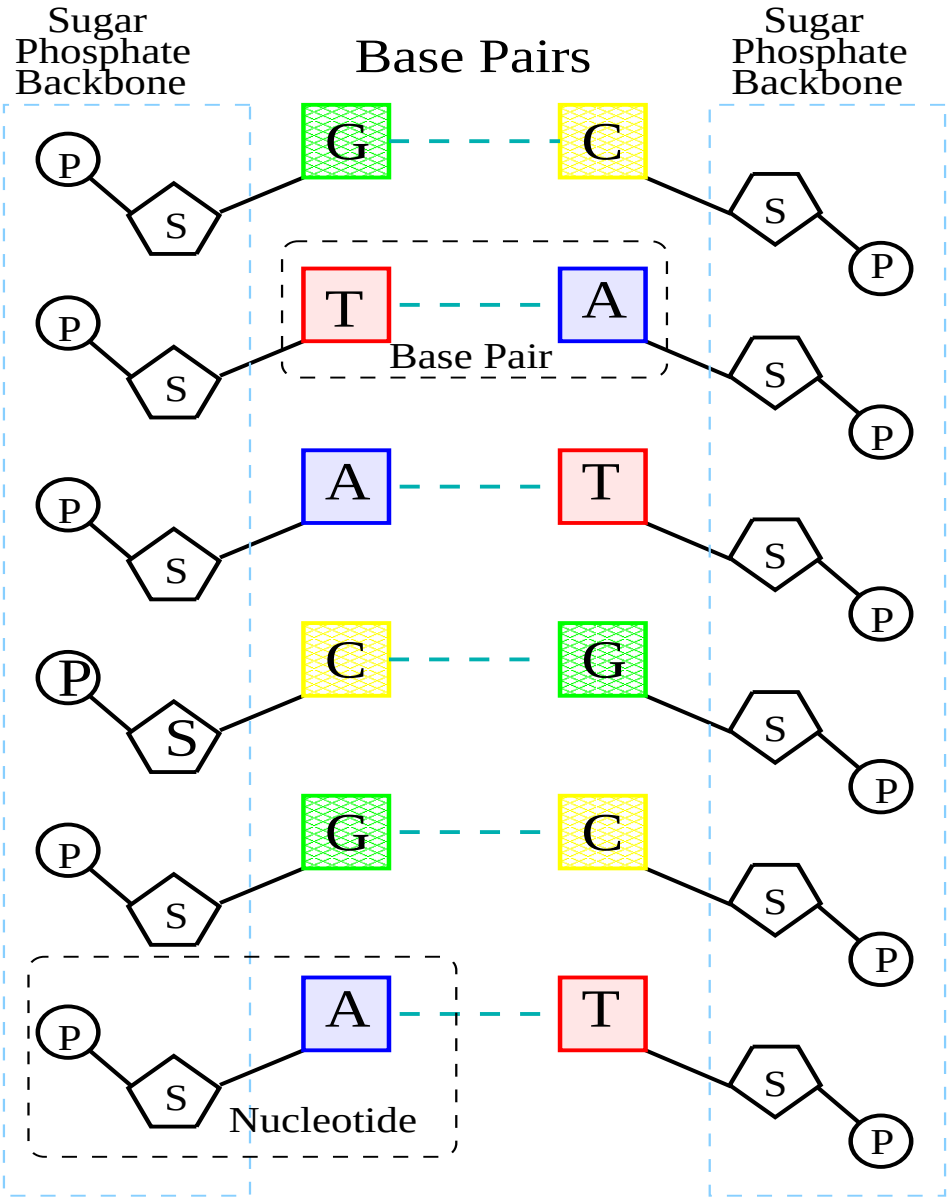
The expression of the genetic information stored in the DNA molecule occurs in two stages:

(i) **transcription**, during which DNA is transcribed into **mRNA**;

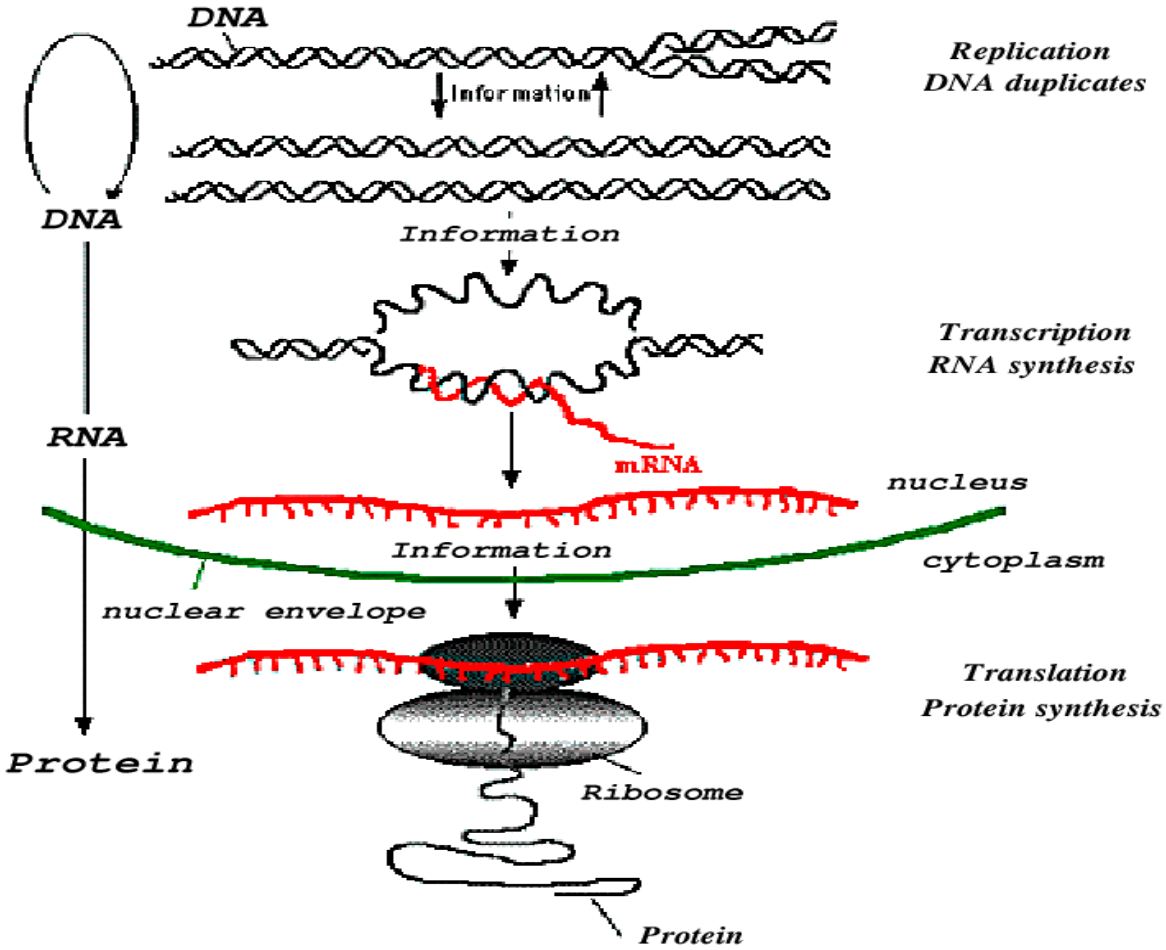
(ii) **translation**, during which mRNA is translated to produce a **protein**.

The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the **genetic code**, which relates nucleotide triplets to amino acids.

DNA


Sugar
Phosphate
Backbone
Base Pairs
Sugar
Phosphate
Backbone
G
C
T
A
A
T
C
G
G
C
A
T
P
S
Base Pair
Nucleotide

Central dogma



cDNA microarrays

Although the regulation of protein synthesis in a cell is by no means regulated solely by mRNA levels, mRNA levels sensitively reflect the type and state of the cell.

cDNA microarrays allow the monitoring of mRNA levels in different types of cells for thousands of genes simultaneously.

Microarrays derive their power and universality from a key property of DNA molecules: **complementary base-pairing**.

A cDNA microarray consists of thousands of individual DNA sequences or **probes** printed in a high density array on a glass microscope slide.

cDNA microarrays, cont'd

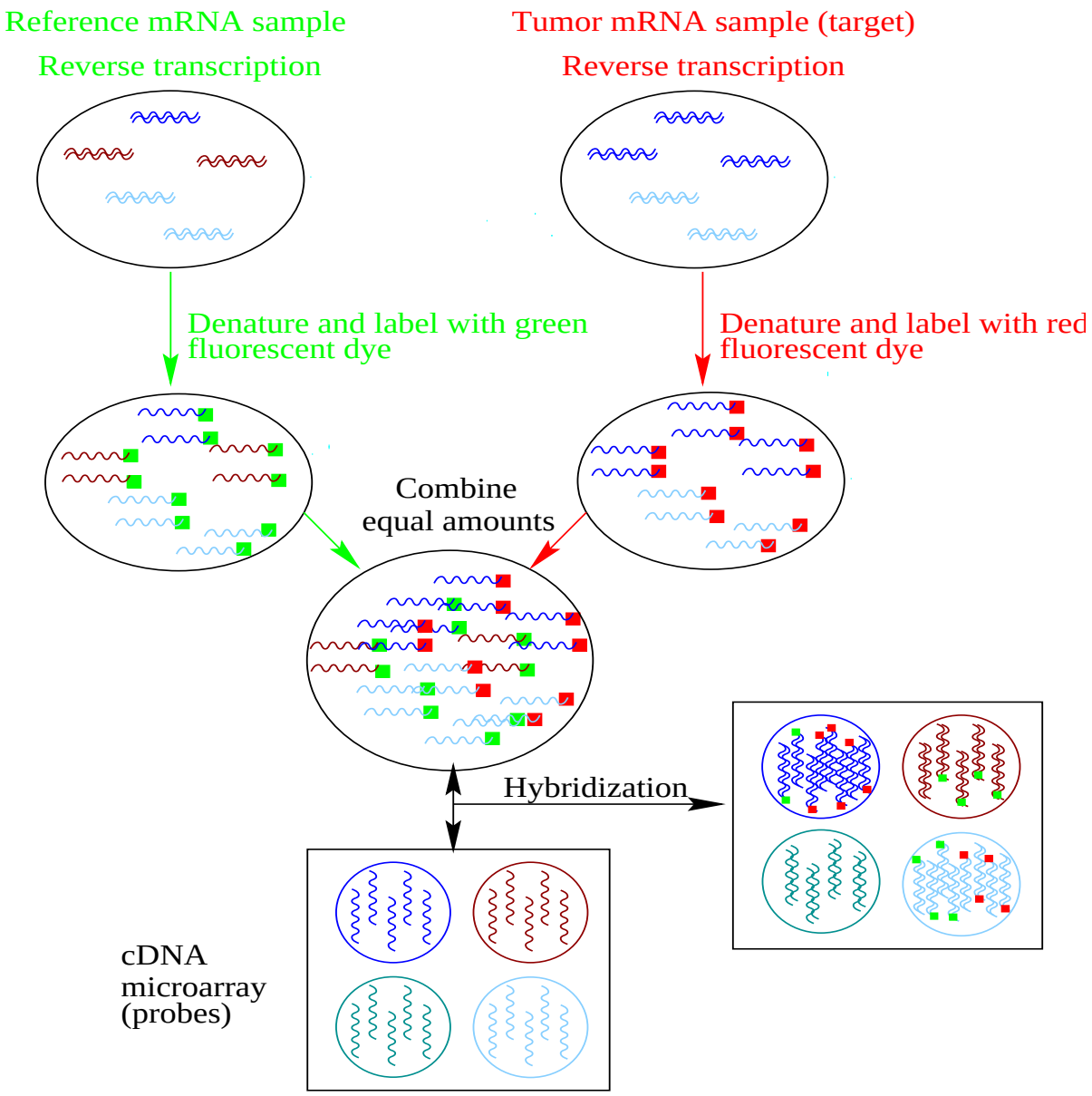
The relative abundance of a set of probes in two mRNA (cDNA) samples or **targets** may be assessed by monitoring the differential hybridization of the two targets to the sequences on the array.

The two targets are labeled using different fluorescent dyes, then mixed and hybridized with the arrayed probes.

Fluorescence measurements are made separately for each dye at each spot on the array.

The ratio of the fluor intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two targets.

cDNA microarrays, cont'd



Tumor classification

A reliable and precise classification of tumors is essential for successful treatment of cancer.

Current methods for classifying human malignancies rely on a variety of morphological, clinical, and molecular variables.

In spite of recent progress, there are still uncertainties in diagnosis. Also, it is likely that the existing classes are heterogeneous.

DNA microarrays may be used to characterize the molecular variations among tumors by monitoring gene expression profiles on a genomic scale.

This may lead to a more reliable classification of tumors.

Tumor classification, cont'd

There are three main types of statistical problems associated with tumor classification:

1. the identification of new/unknown tumor classes using gene expression profiles - *cluster analysis / unsupervised learning*;
2. the classification of malignancies into known classes - *discriminant analysis / supervised learning*;
3. the identification of "marker" genes that characterize the different tumor classes - *variable selection*.

Gene expression data

Gene expression data on p genes (variables) for n mRNA samples (observations)

$$X_{n\times p} = \begin{array}{c} \text{Genes} \\ \left[\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{array}\right] \end{array} \begin{array}{l} \\ \\ \\ \\ \text{mRNA samples} \end{array}$$

$$\begin{aligned} x_{ij} &= \text{gene expression level of gene } j \text{ in mRNA sample } i \\ &= \begin{cases} \log\left(\frac{\text{Red intensity}}{\text{Green intensity}}\right), \\ \log(\text{Avg. PM} - \text{Avg. MM}). \end{cases} \end{aligned}$$

Gene expression data, cont'd

In some situations, the mRNA samples are known to belong to certain classes (*e.g.* follicular lymphoma).

Label the classes by $\{1, 2, \ldots, K\}$.

Then, the data for each observation consist of:

$$\begin{aligned} \mathbf{x}_i &= (x_{i1}, x_{i2}, \ldots, x_{ip}) \\ &\quad \text{- gene expression profile / predictor variables} \\ y_i &= \text{tumor class / response.} \end{aligned}$$

The prediction problem

Want to predict a **response** y given **predictor variables** $\mathbf{x}$.

Task: construct a prediction function f, such that $f(\mathbf{x})$ is an accurate predictor of y.

E.g. digit recognition for zipcodes;

prediction of binding peptide sequences;

prediction of tumor class from gene expression data.

Predictors

A **predictor** or **classifier** for K tumor classes partitions the space $\mathcal{X}$ of gene expression profiles into K disjoint subsets, $A_1, \ldots, A_K$, such that for a sample with expression profile
$\mathbf{x} = (x_1, \ldots, x_p) \in A_k$ the predicted class is k.

Predictors are built from past experience, *i.e.* from observations which are known to belong to certain classes. Such observations comprise the **learning set (LS)** $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

Classifier built from a learning set $\mathcal{L}$:

$$C(\cdot, \mathcal{L}) : \mathcal{X} \to \{1, 2, \ldots, K\}.$$

Predicted class for an observation $\mathbf{x}$: $C(\mathbf{x}, \mathcal{L})$.

Predictors, cont'd

- Fisher linear discriminant analysis.
- Gaussian maximum likelihood discriminant rules (*e.g.* gene voting of Golub *et al.*)
  - Constant or variable class covariance matrices;
  - Diagonal or general class covariance matrices.
- Nearest neighbor classifier.
- Classification trees (CART)
  - Single tree;
  - Bagging: non-parametric, parametric, convex pseudo-data;
  - Boosting.

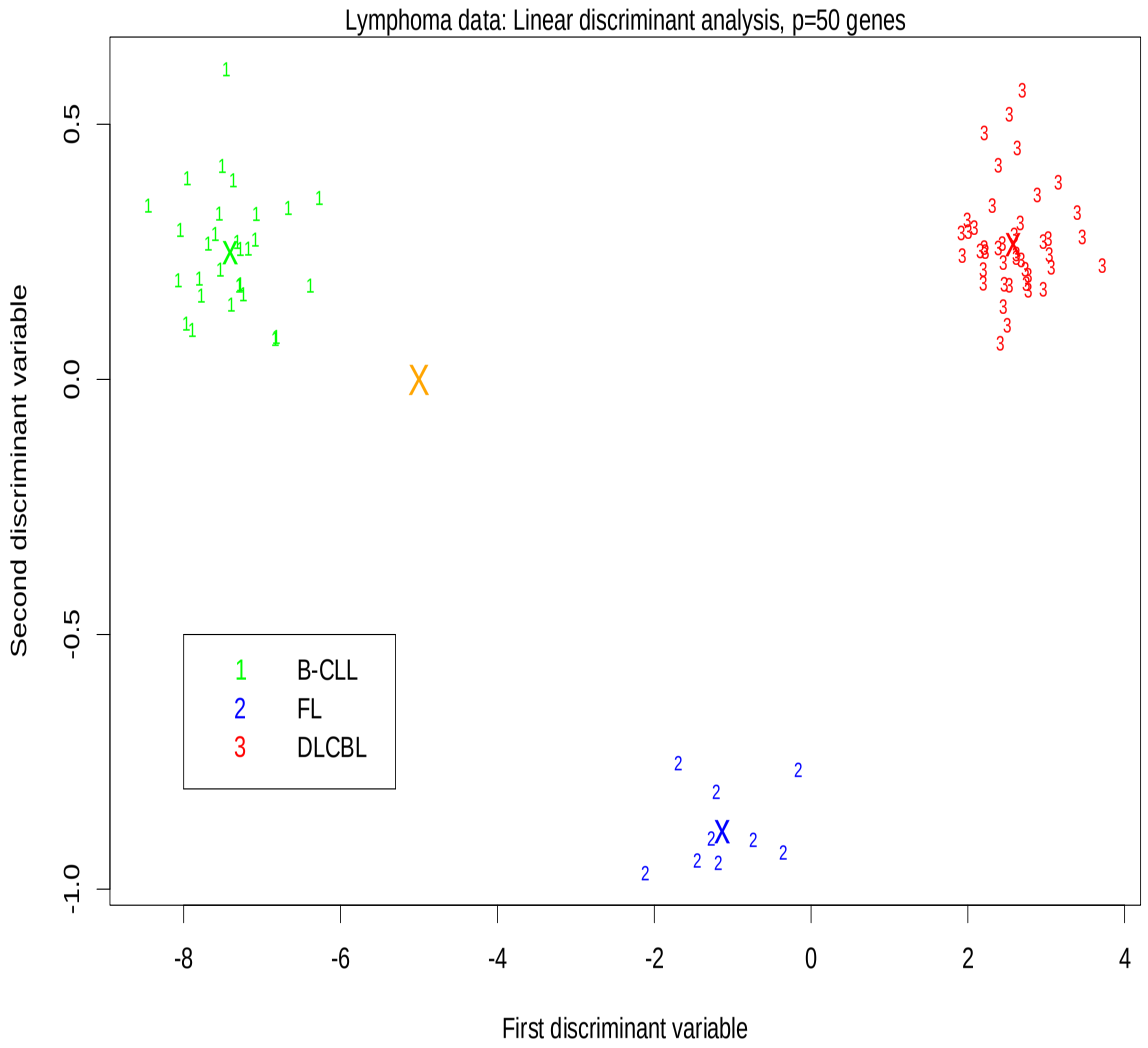
Fisher linear discriminant analysis

First applied in 1935 by M. Barnard at the suggestion of R. A. Fisher (1936), **Fisher linear discriminant analysis (FLDA)** consists of

1. finding linear combinations $\mathbf{x}\,\mathbf{a}$ of the gene expression profiles $\mathbf{x} = (x_1, \ldots, x_p)$ with large ratios of between-groups to within-groups sum of squares - **discriminant variables**;

2. predicting the class of an observation $\mathbf{x}$ by the class whose mean vector is closest to $\mathbf{x}$ in terms of the discriminant variables.

Fisher linear discriminant analysis, cont'd



Maximum likelihood discriminant rules

When the class conditional densities $pr(\mathbf{x}|y = k)$ are known, the **maximum likelihood (ML) discriminant rule** predicts the class of an observation $\mathbf{x}$ by $\mathcal{C}(\mathbf{x}) = \text{argmax}_k \; pr(\mathbf{x}|y = k)$.

For multivariate normal class densities, *i.e.*, for
$\mathbf{x}|y = k \sim N(\mu_k, \Sigma_k)$

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \; \left\{(\mathbf{x} - \mu_k)\Sigma_k^{-1}(\mathbf{x} - \mu_k)' + \log|\Sigma_k|\right\}.$$

In general, this is a **quadratic** rule.

In practice, the population mean vectors and covariance matrices are estimated by the corresponding sample quantities.

Maximum likelihood discriminant rules - Special cases

1. When the class densities have the same covariance matrix, $\Sigma_k = \Sigma$, the discriminant rule is based on the square of the Mahalanobis distance and is linear

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \ (\mathbf{x} - \mu_k)\Sigma^{-1}(\mathbf{x} - \mu_k)'.$$

   (FLDA for $K = 2$).

2. In this simplest case, when the class densities have the same diagonal covariance matrix $\Delta = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, the discriminant rule is linear and given by

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \sum_{j=1}^{p} \frac{(x_j - \mu_{kj})^2}{\sigma_j^2}.$$

   Diagonal linear discriminant analysis (DLDA).

Weighted gene voting of Golub *et al.* (1999)

This is a minor variant on DLDA for two classes, $k = 1,\ 2$.

DLDA classifies an observation $\mathbf{x}$ as 1 iff

$$\sum_{j=1}^{p} \frac{(\bar{x}_{1j} - \bar{x}_{2j})}{\hat{\sigma}_j^2} \left(x_j - \frac{(\bar{x}_{1j} + \bar{x}_{2j})}{2} \right) \geq 0.$$

The discriminant function can be rewritten as $\sum_j v_j$, where $v_j = a_j(x_j - b_j)$, $a_j = (\bar{x}_{1j} - \bar{x}_{2j})/\hat{\sigma}_j^2$, and $b_j = (\bar{x}_{1j} + \bar{x}_{2j})/2$.

In Golub *et al.* $a_j = (\bar{x}_{1j} - \bar{x}_{2j})/(\hat{\sigma}_{1j} + \hat{\sigma}_{2j})$ ... (Wrong units).

Nearest neighbor classifier

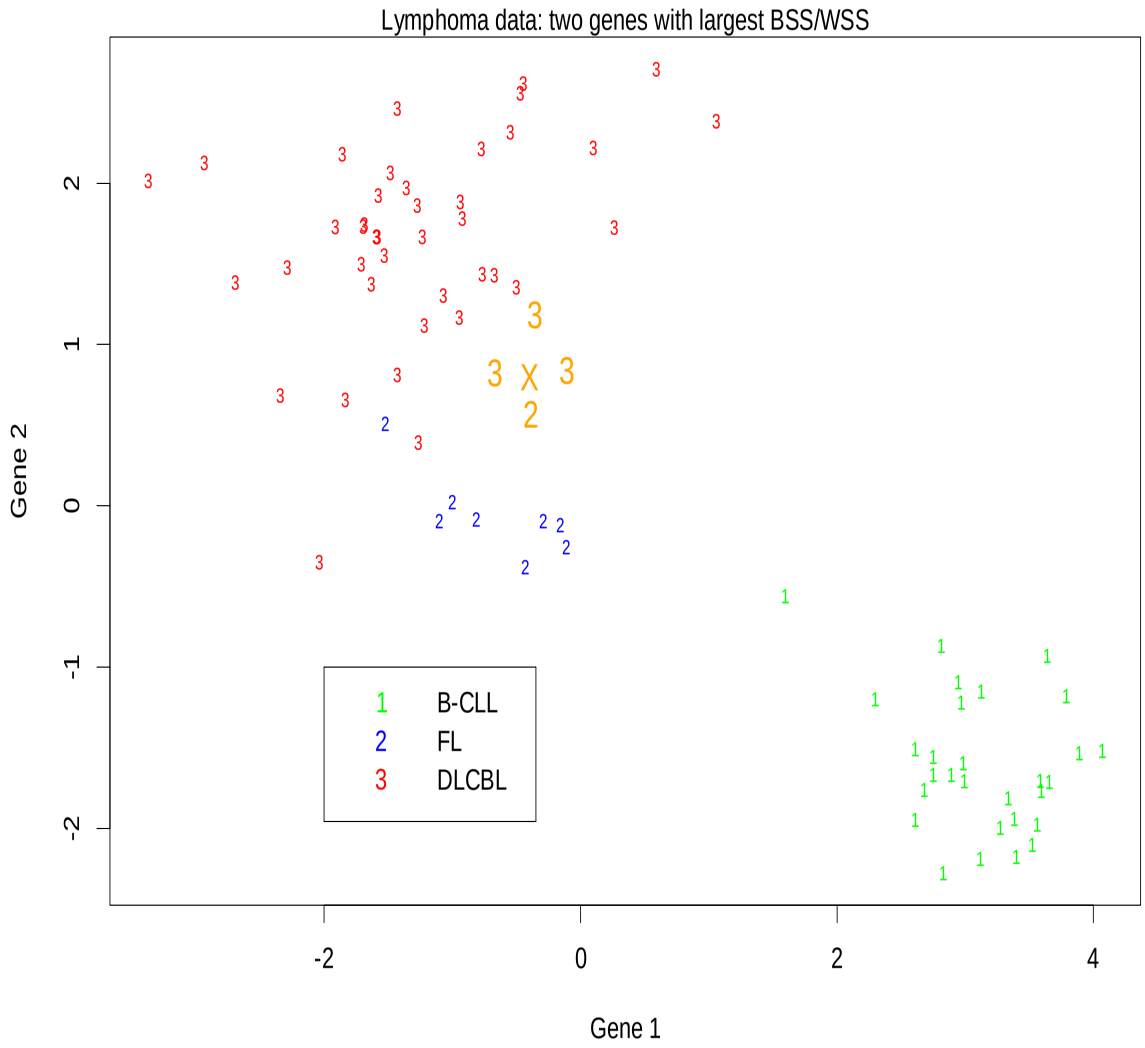
Nearest neighbor methods are based on a measure of distance between observations, such as the Euclidean distance or one minus the correlation between two gene expression profiles.

The k **nearest neighbor** rule, due to Fix and Hodges (1951), classifies an observation $\mathbf{x}$ as follows:

1. find the k observations in the learning set that are closest to $\mathbf{x}$;
2. predict the class of $\mathbf{x}$ by majority vote, *i.e.*, choose the class that is most common among those k neighbors.

The number of neighbors k is chosen by cross-validation.

Nearest neighbor classifier, cont'd



Classification trees

Binary tree structured classifiers are constructed by repeated splits of subsets (nodes) of the measurement space $\mathcal{X}$ into two descendant subsets, starting with $\mathcal{X}$ itself. Each terminal subset is assigned a class label and the resulting partition of $\mathcal{X}$ corresponds to the classifier.

Three main aspects of tree construction: (i) the selection of the splits; (ii) the decision to declare a node terminal or to continue splitting; (iii) the assignment of each terminal node to a class.

Different tree classifiers use different approaches to deal with these three issues. Here, we use **CART - Classification And Regression Trees** - of Breiman *et al.* (1984).

Classification trees, cont'd

Aggregating predictors

Breiman (1996, 1998) found that gains in accuracy could be obtained by **aggregating predictors** built from perturbed versions of the learning set. In classification, the multiple versions of the predictor are aggregated by **voting**.

Let $C(\cdot, \mathcal{L}_b)$ denote the classifier built from the bth perturbed learning set $\mathcal{L}_b$ and let w_b denote the weight given to predictions made by this classifier. The predicted class for an observation $\mathbf{x}$ is given by

$$\text{argmax}_k \sum_b w_b \, I(C(\mathbf{x}, \mathcal{L}_b) = k).$$

Bagging

Breiman (1996).

In the simplest form of **bagging** - **b**ootstrap **agg**regat**ing** - perturbed learning sets of the same size as the original learning set are created by forming non-parametric bootstrap replicates of the learning set, *i.e.* by drawing at random with replacement from the learning set.

Predictors are built for each perturbed dataset and aggregated by plurality voting ($w_b = 1$).

Variants on bagging

Parametric bootstrap. Perturbed learning sets are generated according to a mixture of multivariate normal (MVN) distributions.

Convex pseudo-data. Breiman (1996)
Each perturbed learning set is generated by repeating the following n times:

1. select two instances $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ at random from the learning set;
2. select at random a number v from the interval $[0, d]$, $0 \leq d \leq 1$, and let $u = 1 - v$;
3. define a new instance $(\mathbf{x}'', y'')$ by $y'' = y$ and $\mathbf{x}'' = u\mathbf{x} + v\mathbf{x}'$.

Boosting

Freund and Schapire (1997), Breiman (1998).

The data are **re-sampled adaptively** so that the weights in the re-sampling are increased for those cases most often misclassified.

The aggregation of predictors is done by **weighted voting**.

Boosting, cont'd

For a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, let $\{p_1, \ldots, p_n\}$ denote the re-sampling probabilities, initialized to be equal.

For bth step of the boosting algorithm (adaptation of `AdaBoost`):

1. generate a perturbed learning set $\mathcal{L}_b$ of size n by sampling with replacement from $\mathcal{L}$ using $\{p_1, \ldots, p_n\}$;
2. build a classifier $C(\cdot, \mathcal{L}_b)$ based on $\mathcal{L}_b$;
3. run the learning set $\mathcal{L}$ through the classifier $C(\cdot, \mathcal{L}_b)$ and let $d_i = 1$ if the ith case is classified incorrectly and $d_i = 0$ o.w.;
4. define

   $$\epsilon_b = \sum_i p_i d_i, \ \beta_b = (1 - \epsilon_b)/\epsilon_b \text{ and } w_b = \log(\beta_b)$$

   and update the re-sampling probabilities for the $(b+1)$st step by

   $$p_i = \frac{p_i \beta_b^{d_i}}{\sum_i p_i \beta_b^{d_i}}.$$

Prediction votes

For aggregated classifiers, prediction votes assessing the strength of a prediction may be defined for each observation.

The **prediction vote** (PV) for an observation $\mathbf{x}$ is defined to be

$$PV(\mathbf{x}) = \frac{\max_k \sum_b w_b \, I(C(\mathbf{x}, \mathcal{L}_b) = k)}{\sum_b w_b}.$$

When the perturbed learning sets are given equal weights, *i.e.*, $w_b = 1$, the prediction vote is simply the proportion of votes for the "winning" class, regardless of whether it is correct or not.

Prediction votes belong to $[0, 1]$.

Datasets - Lymphoma

Study of gene expression in the three most prevalent adult lymphoid malignancies using a specialized cDNA microarray, the Lymphochip (Alizadeh *et al.*, 2000).

- $n = 81$ mRNA samples, three classes:

B-cell chronic lymphocytic leukemia (B-CLL)	29 cases
Follicular lymphoma (FL)	9 cases
Diffuse large B-cell lymphoma (DLBCL)	43 cases

- $p = 4,682$ genes.

Datasets - Leukemia

Study of gene expression in two types of acute leukemias using Affymetrix high-density oligonucleotide arrays (Golub *et al.*, 1999).

- $n = 72$ mRNA samples, three classes:

B-cell acute lymphoblastic leukemia (B-cell ALL)	38 cases
T-cell acute lymphoblastic leukemia (T-cell ALL)	9 cases
Acute myeloid leukemia (AML)	25 cases

- $p = 6,817$ genes.

Datasets - NCI 60

Study of gene expression among the 60 cell lines from the National Cancer Institute's anti-cancer drug screen (NCI 60) using cDNA microarrays (Ross *et al.*, 2000).

- $n = 64$ mRNA samples, 9 classes:

  breast (7, MCF7 $\times$ 3), central nervous system (CNS) (5), colon (7), leukemia (6, K562A $\times$ 3), melanoma (8), non-small-cell-lung-carcinoma (NSCLC) (9), ovarian (6), prostate (2), renal (9), unknown(1).

  We exclude the prostate and unknown classes from our comparison.

- $p = 5,244$ genes.

Data pre-processing

- **Imputation.** k-nearest neighbor imputation, where genes are "neighbors" and the similarity measure between two genes is the correlation in their expression profiles.
- **Standardization.** Standardize observations (arrays) to have mean 0 and variance 1 across variables (genes).

Study design

The original datasets are repeatedly randomly divided into a learning set and a test set, comprising respectively 2/3 and 1/3 of the data. For each of $N = 150$ runs:

- Select a subset of p genes from the learning set based on their ratio of between- to within-groups sum of squares, BSS/WSS. $p = 50$ for lymphoma, $p = 40$ for leukemia, $p = 30$ for NCI 60.
- Build the different predictors using the learning sets with p genes.
- Apply the predictors to the observations in the test set to obtain test set error rates.

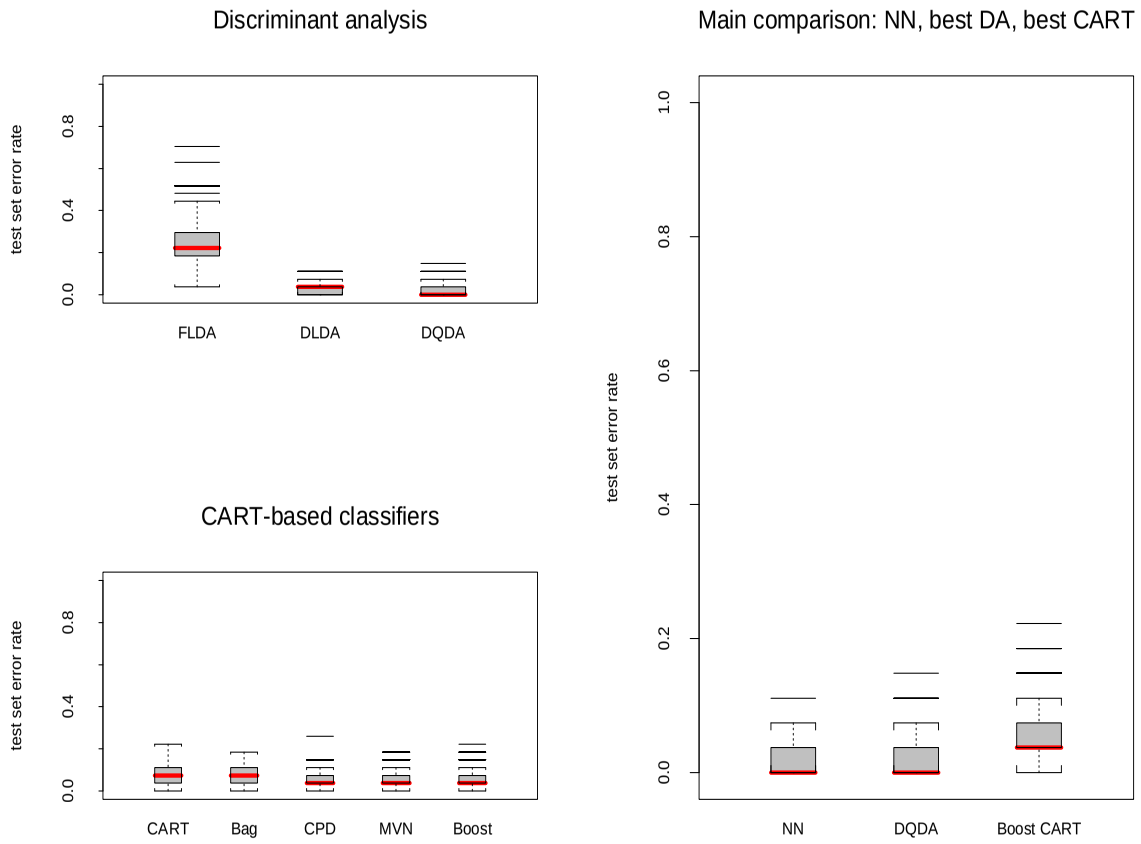


Figure 1: *Lymphoma.* Boxplots of test set error rates, $p = 50$ genes.

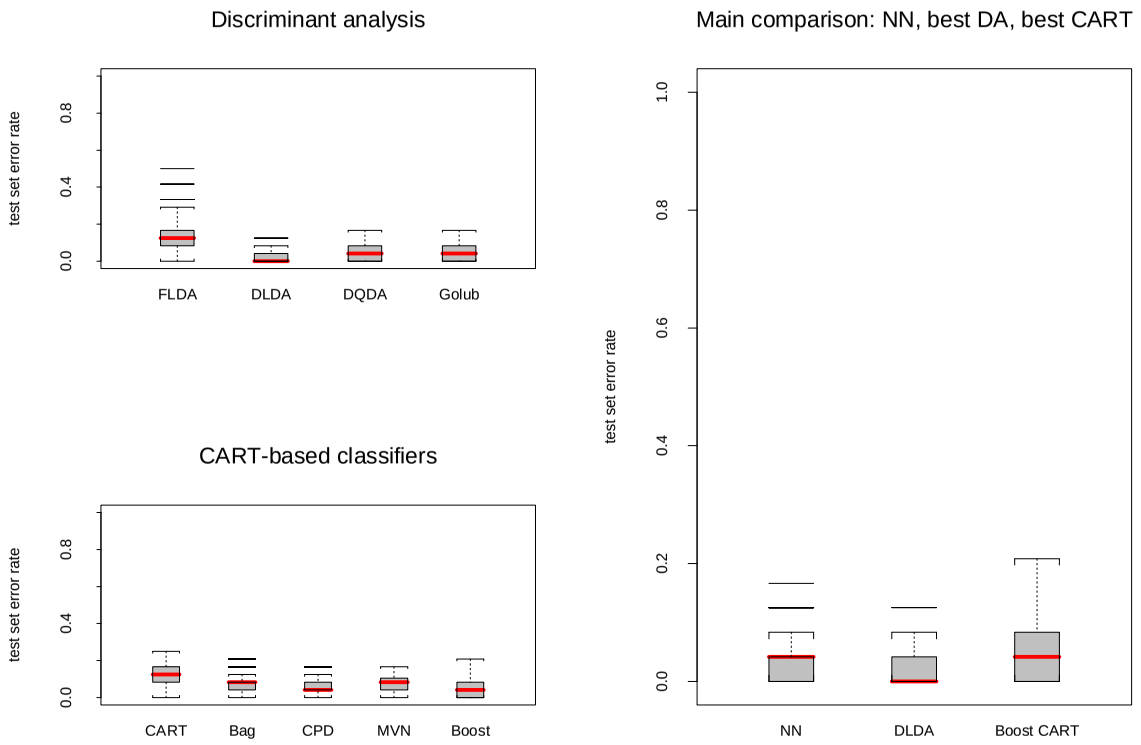


Figure 2: *Leukemia, two classes.* Boxplots of test set error rates, $p = 40$ genes.

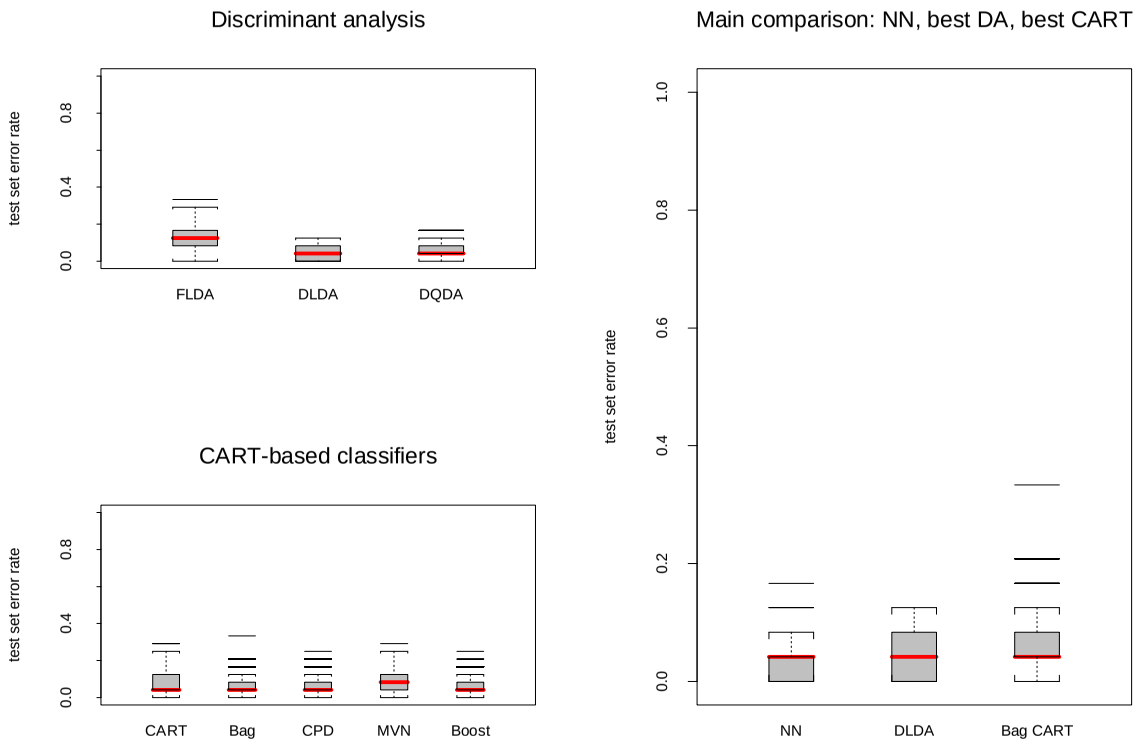


Figure 3: *Leukemia, three classes.* Boxplots of test set error rates, $p = 40$ genes.

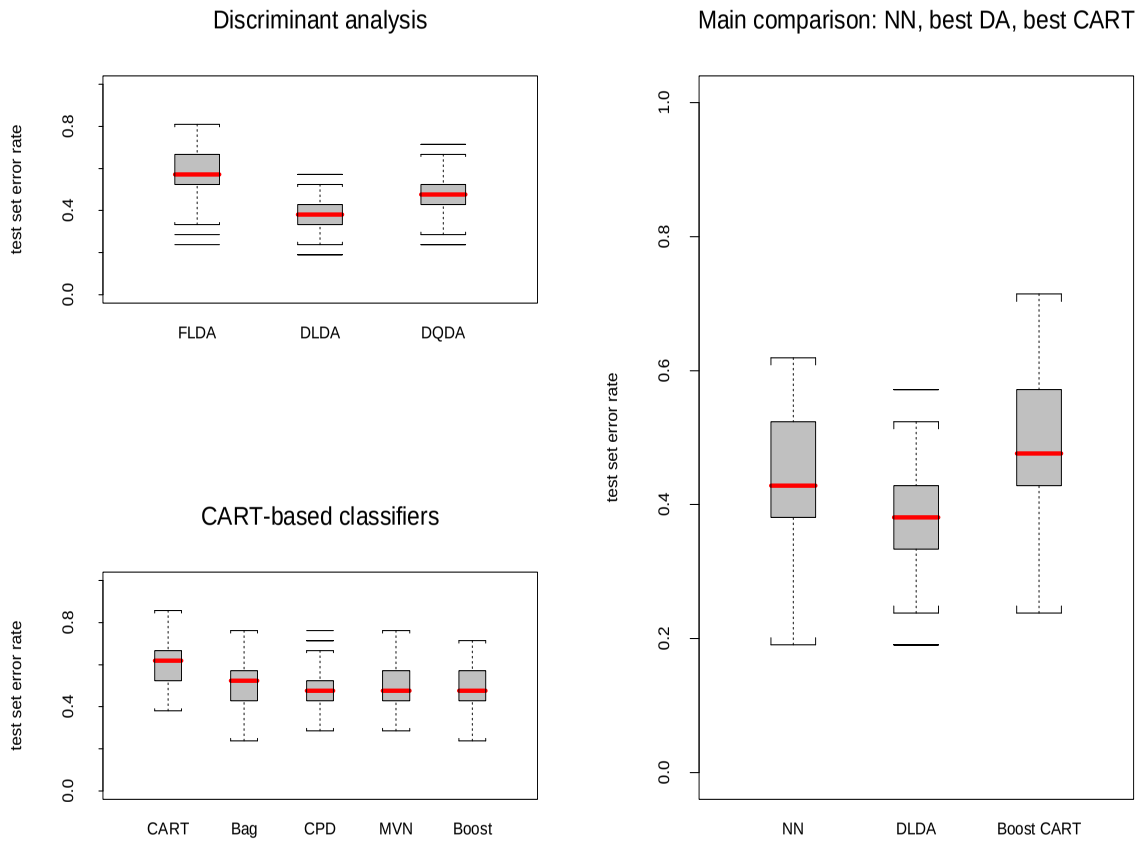
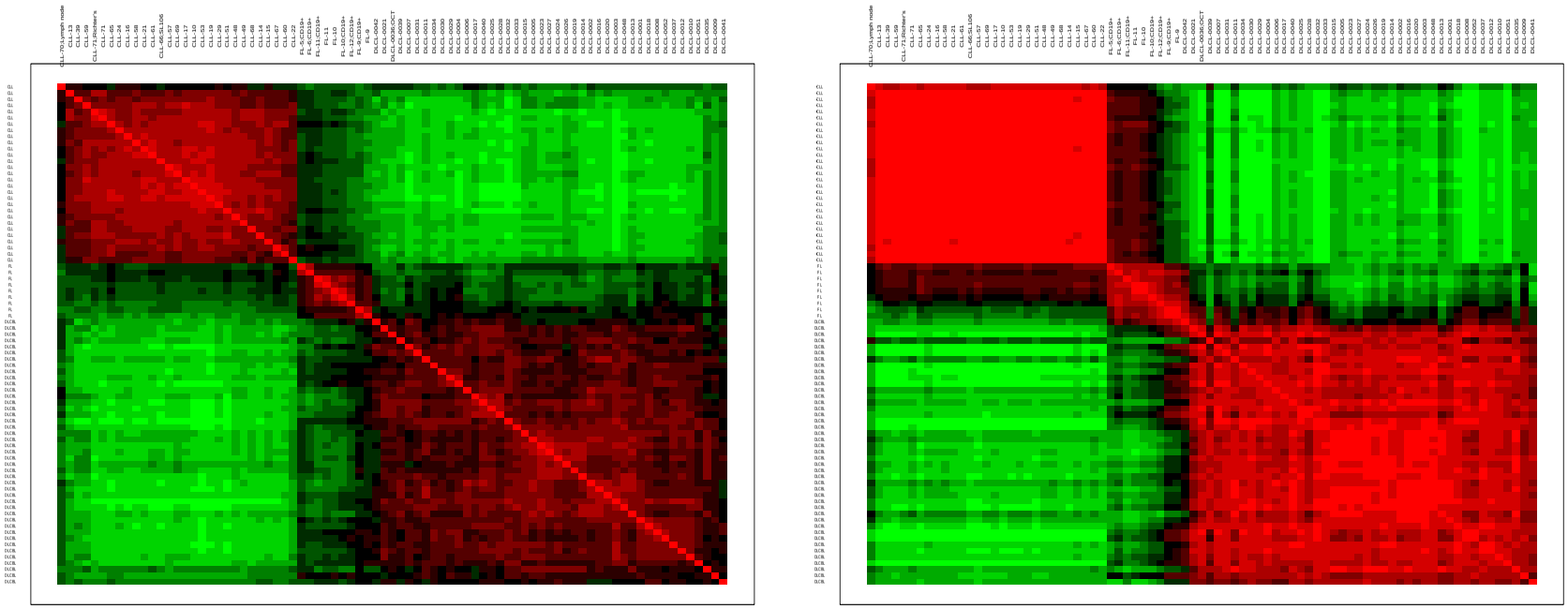


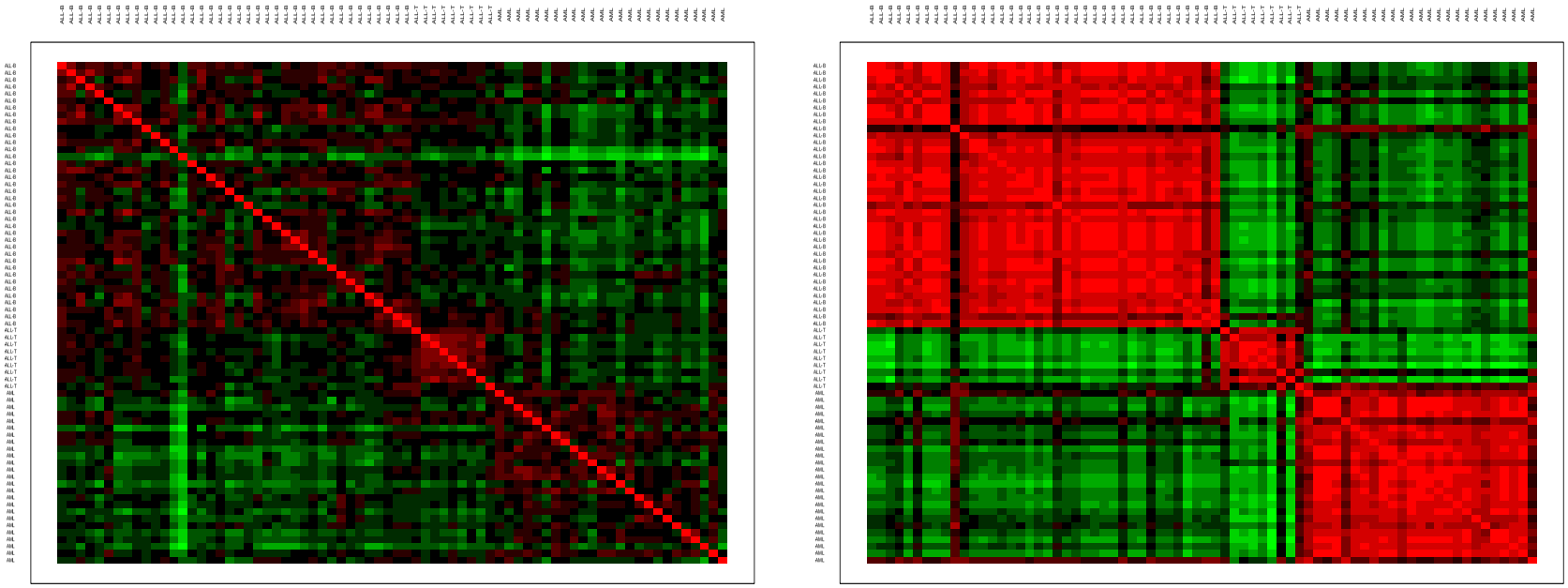
Figure 4: *NCI 60*. Boxplots of test set error rates, $p = 30$ genes.



4,682 genes

50 genes

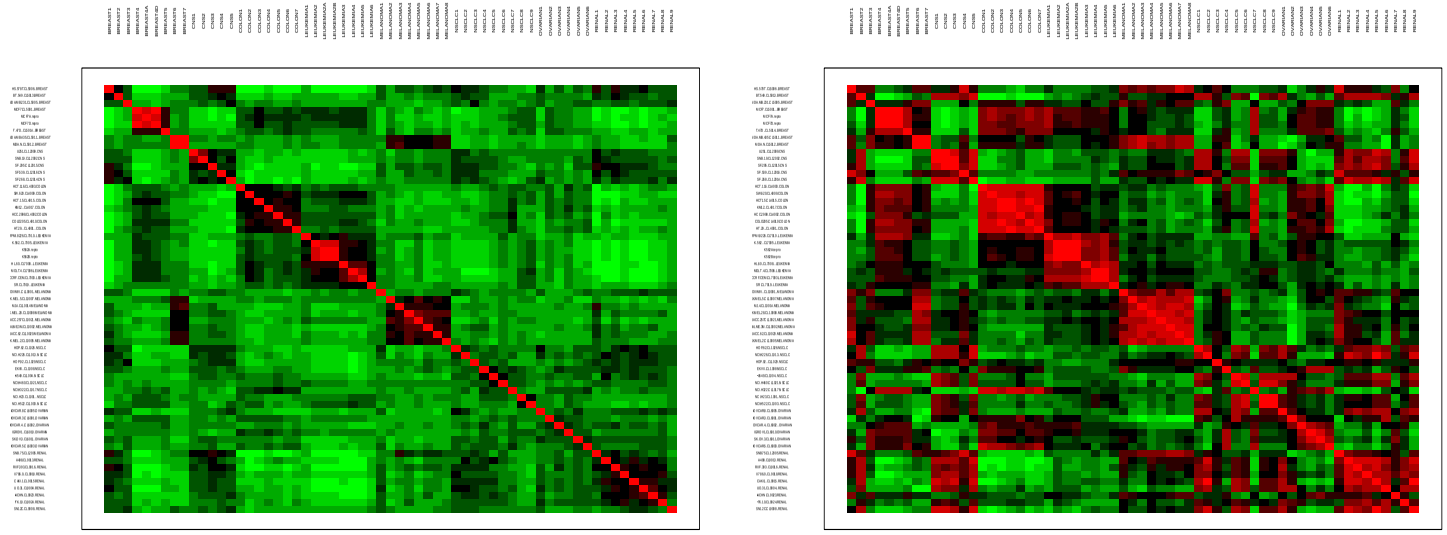
Figure 5: *Lymphoma.* Images of correlation matrix between 81 mRNA samples.



3,571 genes

40 genes

Figure 6: *Leukemia.* Images of correlation matrix between 72 mRNA samples.



5,244 genes

30 genes

Figure 7: *NCI 60.* Images of correlation matrix between 61 mRNA samples.

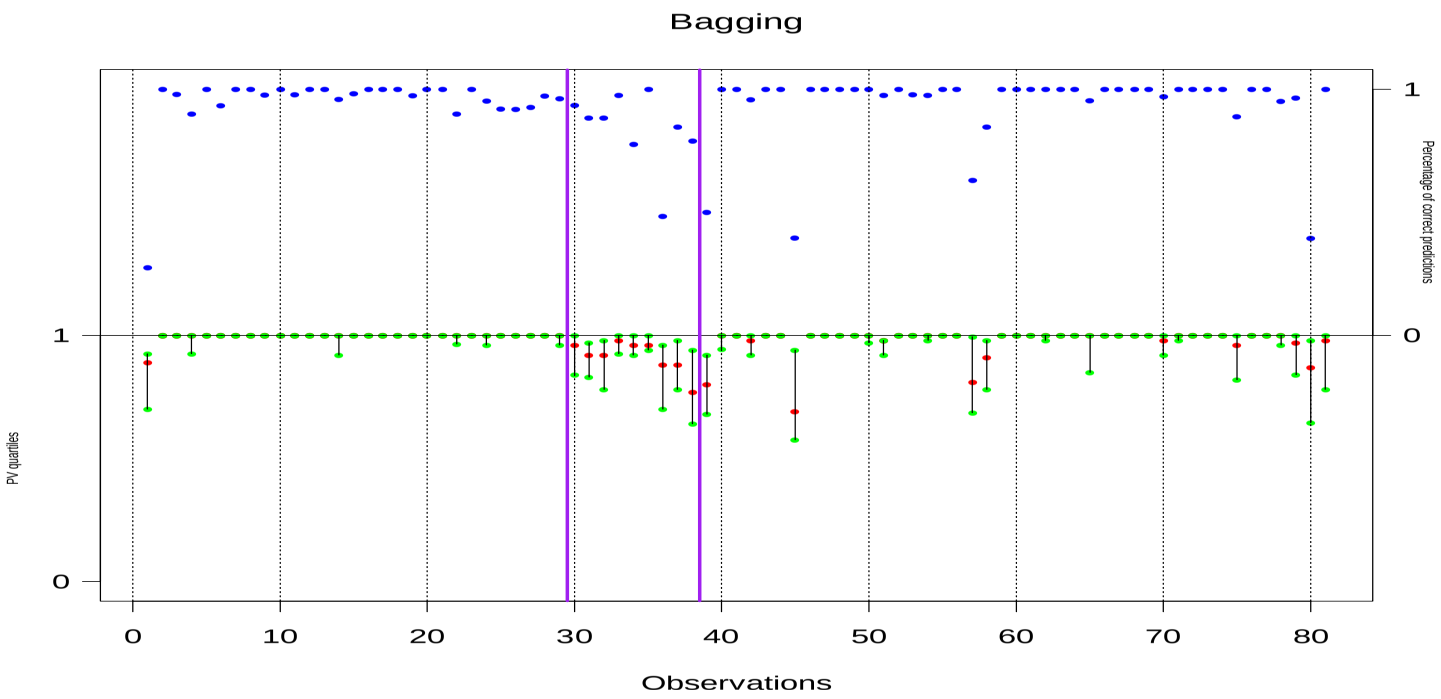


Figure 8: *Lymphoma.* Proportion of correct predictions (upper panel) and quartiles of prediction votes (lower panel) for bagged CART classifiers, $p = 50$ genes.

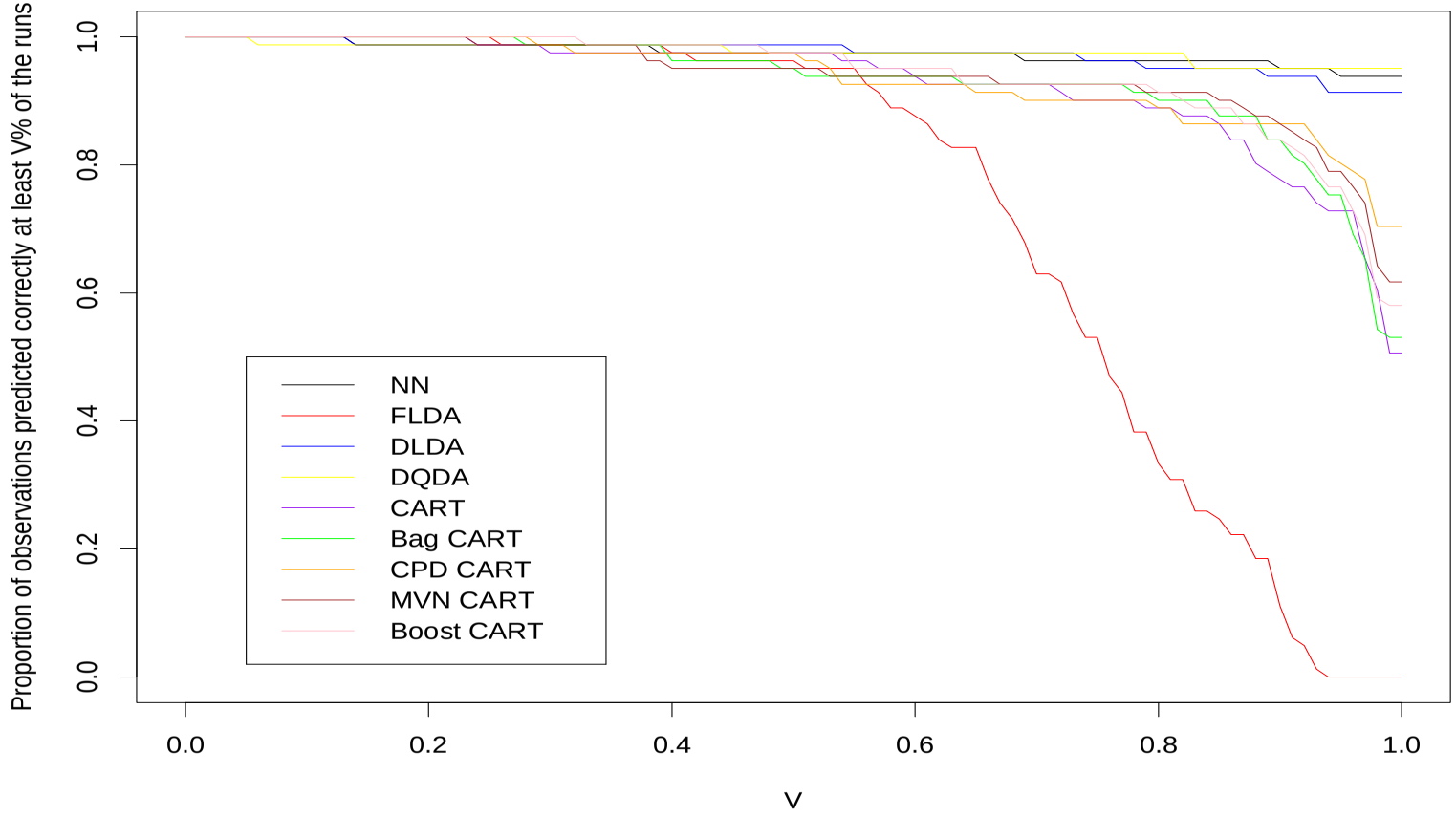


Figure 9: *Lymphoma*. Fraction of observations predicted correctly at least $V\%$ of the time (out of the runs for which a given observation belonged to the test set) *vs.* $V\%$, for classifiers built using $p = 50$ genes.

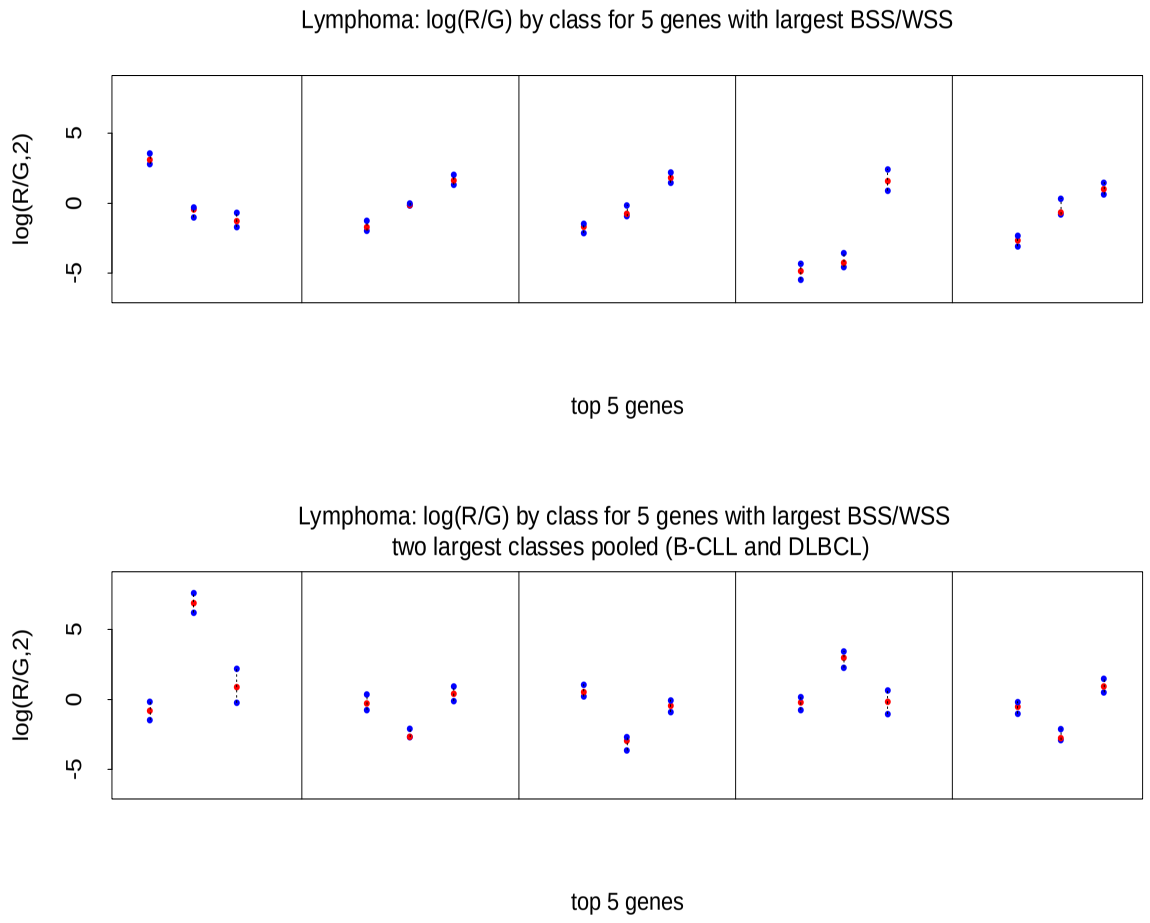


Figure 10: *Lymphoma.* Summary statistics of within class expression levels for 5 genes.

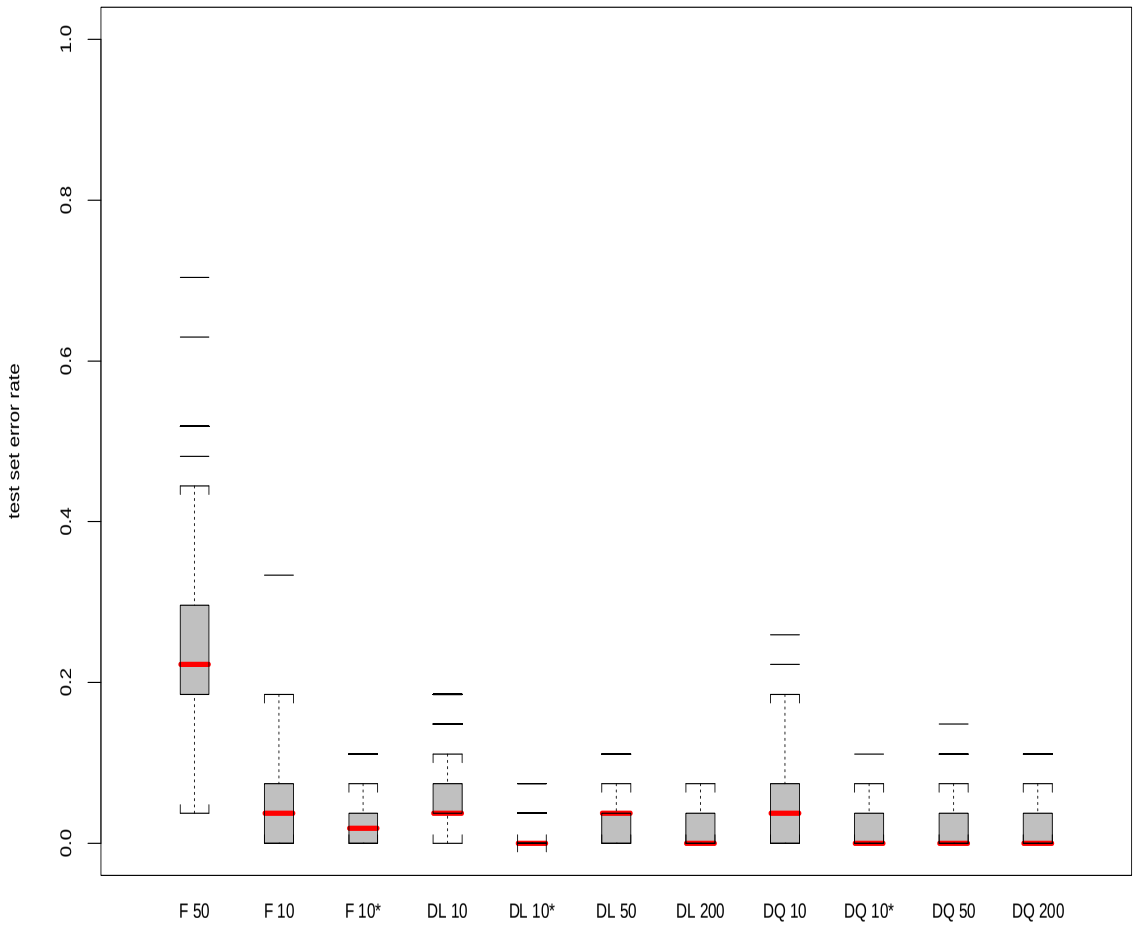


Figure 11: *Lymphoma.* Boxplots of test set error rates for discriminant analysis based on different gene sets.

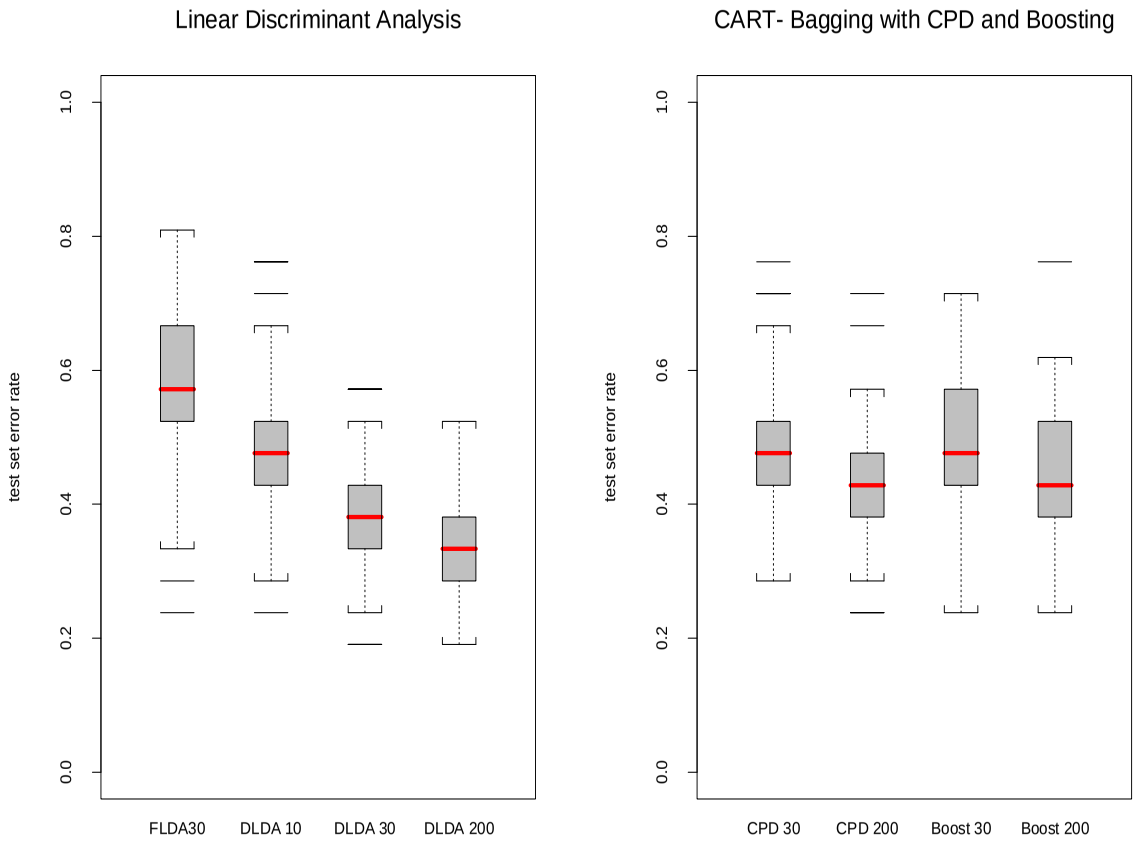


Figure 12: *NCI 60*. Boxplots of test set error rates for different gene sets, discriminant analysis and aggregated CART.

Results

- In the main comparison, the nearest neighbor classifier and DLDA had the smallest error rates, while FLDA had the highest error rates.
- Aggregation improved the performance of CART classifiers, the largest gains being with boosting and bagging with convex pseudo-data.
- For the lymphoma and leukemia datasets, increasing the number of variables to $p = 200$ didn't affect much the performance of the various classifiers. There was an improvement for the NCI 60 dataset.
- A more careful selection of a small number of genes ($p = 10$) improved the performance of FLDA dramatically.

Discussion

- "Diagonal" LDA *vs.* "correlated" LDA: ignoring correlation between genes helps here.
- Unlike classification trees and nearest neighbors, "diagonal" or "correlated" LDA are unable to take into account gene interactions.
- Although nearest neighbors are simple and intuitive classifiers, their main limitation is that they give little insight into mechanisms underlying the class distinctions.
- Classification trees are capable of handling and revealing interactions between variables.
- Useful by-product of aggregated classifiers: prediction votes.
- The relative performance of the different classifiers may vary with variable selection.

Open questions

- **Variable selection**. A crude criterion such as BSS/WSS may not identify the genes that discriminate between all the classes and may not reveal interactions between genes.

  Statistical *vs.* biological significance.

- **Cluster analysis**. Identification and validation of new tumor classes.

Acknowledgments

Biologists: Ash Alizadeh, Pat Brown, Mike Eisen.

Statisticians: Leo Breiman, Sam Buttrey, Yoram Gat,
David Nelson, Mark van der Laan, Jean Yang.

`http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html`