

# Supervised Clustering of Genes

Marcel Dettling\*

and

Peter Bühlmann

September 30, 2002

## Abstract

**Background:** We focus on microarray data where experiments monitor the gene expression of different tissues and where each experiment is equipped with an additional response variable such as a cancer type. While the number of measured genes is in the thousands, it is assumed that only a few marker components of gene subsets determine the type of a tissue. In this paper, we present a novel method for finding such groups of genes by directly incorporating the response variables into the grouping process, therefore yielding a supervised clustering algorithm for genes.

**Results:** An empirical study on eight publicly available microarray datasets shows that our algorithm identifies gene clusters with excellent predictive potential, often superior to classification with state-of-the-art methods based on single genes. Permutation tests and bootstrapping provide evidence that the output is reasonably stable and more than a noise artifact.

**Conclusions:** In contrast to other methods such as hierarchical clustering, our algorithm identifies several gene clusters whose expression levels clearly discriminate the different tissue types. The identification of such gene clusters is potentially useful for medical diagnostics and may at the same time reveal insights into functional genomics.

**Software:** Is available at <http://stat.ethz.ch/~dettling/supercluster>

**Keywords:** Gene Expression Data, Clustering, Discrimination, Classification

---

\*Corresponding author. Tel.: +41-1-632-5319; fax: +41-1-632-1228; e-mail: [dettling@stat.math.ethz.ch](mailto:dettling@stat.math.ethz.ch)

# 1 Introduction

The recently developed microarray technology allows for measuring expression levels of thousands of genes simultaneously and is expected to contribute significantly to advances in fundamental questions of biology and medicine. We focus on the case where the experiments monitor the gene expression of different tissue samples, and where each experiment is equipped with an additional categorical outcome variable, describing e.g. a cancer type. An important problem in this setting is to study the relation between gene expression and tissue type. While microarrays monitor thousands of genes, it is assumed that only a few underlying marker components of gene subsets account for nearly all of the outcome variation, i.e. determine the type of a tissue. The identification of these functional groups is crucial for tissue classification in medical diagnostics, as well as for understanding how the genome as a whole works.

As a first approach, unsupervised clustering techniques have been widely applied to find groups of co-regulated genes on microarray data. *Hierarchical Clustering* [1, 2] identifies sets of correlated genes with similar behavior across the experiments, but yields thousands of clusters in a tree-like structure. This makes the identification of functional groups very difficult. In contrast, *Self-Organizing-Maps* [3] require a prespecified number and an initial spatial structure of clusters, but this may be hard to come up with in real problems. These drawbacks were improved by a novel graph theoretical clustering algorithm [4], but as all other unsupervised techniques, it usually fails to reveal functional groups of genes that are of special interest in tissue classification. This is because genes are clustered by similarity only, without using any information about the experiment’s response variables.

We focus here on *Supervised Clustering*, defined as grouping of variables (genes), controlled by information about the  $Y$  variables, i.e. the tumor types of the tissues. Previous work in this field encompasses *Tree Harvesting* [5], being a 2-step method which consists of generating numerous candidate groups by unsupervised hierarchical clustering. Second, the average expression profile of each cluster is considered as a potential input variable for a response model and the few gene groups that contain the most useful information for tissue discrimination are identified. Only this second step makes the clustering supervised, since the selection process relies on external information about the tissue types. An interesting supervised clustering approach that directly incorporates the response variables  $Y$  in the grouping process is the *Partial Least Squares* (PLS) procedure [6, 7], an often applied tool in the chemometrics literature, which in a supervised manner constructs weighted linear combinations of genes that have maximal covariance with the outcome. PLS has the drawback that the fitted components involve all (usually thousands of) genes, which makes them very difficult to interpret.

In this paper, we present a promising new method for searching functional groups, each made up of only a few genes whose consensus expression profiles provides useful information for tissue discrimination. As PLS, it is a 1-step approach which directly incorporates the response variables  $Y$  into the grouping process, thus an algorithm for *Supervised Clustering of Genes*. Due to the combinatorial complexity when clustering thousands of genes, we rely on a greedy strategy. It optimizes an empirical objective function that quickly and efficiently measures the cluster’s ability for phenotype discrimination. Inspired by [8], we choose Wilcoxon’s test statistic for two unpaired samples [9], refined by a novel second criterion, the margin function. Our supervised algorithm can be started with or without initial groups of genes, and then clusters genes in a stagewise forward and backward search, as long as their differential expression in terms of our objective function can be improved. This yields clusters typically made up of 3–9 genes, whose coherent average

expression levels allow perfect discrimination of tissue types. In an empirical study, the clusters show excellent out-of-sample predictive potential, and permutation and randomization techniques show that they are reasonably stable and clearly more than just a noise artifact. The output of our algorithm is thus potentially beneficial for cancer type diagnosis. At the same time it is very accessible for interpretation, since the output consists of a very limited number of clusters, each summarizing the information of a few genes. Thus, it may also reveal insights into biological processes and give hints on explaining how the genome works.

The plan for the rest of this paper is as follows. In section 2 the algorithm for supervised clustering of gene expression data is described. In section 3, the new clustering procedure is applied to eight publicly available microarray datasets and the results are tested for their predictive potential, their stability and their relevance. Conclusions and an outlook are given in section 4.

## 2 Supervised Clustering of Genes

This section presents an algorithm for supervised learning of similarities and interactions among predictor variables for classification in very high dimensional spaces, and hence is predestinated for searching functional groups of genes on microarray expression data.

### 2.1 The Partitioning Problem

Our basic stochastic model for microarray data equipped with categorical response is given by a random pair

$$(\mathbf{X}, Y) \text{ with values in } \mathbb{R}^p \times \mathbb{Y},$$

where  $\mathbf{X} \in \mathbb{R}^p$  denotes a log-transformed gene expression profile of a tissue sample, standardized to mean zero and unit variance.  $Y$  is the associated response variable, taking numeric values in  $\mathbb{Y} = \{0, 1, \dots, K - 1\}$ . A usual interpretation is that  $Y$  codes for one of  $K$  cancer types. For simplicity and a concise description of the algorithm, we first assume that  $K = 2$ , so that the response is binary. A generalization of the setting for multi-categorical response ( $K > 2$ ) is given below in section 2.3.

To account for the fact that not all  $p$  genes on the chip individually, but rather a few functional gene subsets determine nearly all of the outcome variation and thus the type of a tissue, we model the conditional probability as

$$P[Y = 1|\mathbf{X}] = f(X_{\mathcal{C}_1}, X_{\mathcal{C}_2}, \dots, X_{\mathcal{C}_q}), \quad (1)$$

where  $f(\cdot)$  is a nonlinear function mapping from  $\mathbb{R}^q$  to  $[0, 1]$ ,  $\{\mathcal{C}_1, \dots, \mathcal{C}_q\}$  with  $q \ll p$  are functional groups or clusters of genes which form a disjoint and usually incomplete partition of the index set:  $\{\cup_{i=1}^q \mathcal{C}_i\} \subset \{1, \dots, p\}$  and  $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ ,  $i \neq j$ . Finally,  $X_{\mathcal{C}_i} \in \mathbb{R}$  denotes a ‘‘representative’’ expression value of gene cluster  $\mathcal{C}_i$ . There are many possibilities to determine such group values  $X_{\mathcal{C}_i}$ , but since we would like to shape clusters that contain similar genes, a simple linear combination is an accurate choice, see [5, 10]:

$$X_{\mathcal{C}_i} = \frac{1}{|\mathcal{C}_i|} \sum_{g \in \mathcal{C}_i} \alpha_g X_g \text{ with } \alpha_g \in \{-1, 1\}. \quad (2)$$

Due to the use of log-transformed, mean centered and standardized expression data, we as a novel extension allow the contribution of a particular gene  $g$  to the group value  $X_{\mathcal{C}_i}$

also to be given by its “sign-flipped” expression value  $-X_g$ . This means that we treat under- and overexpression symmetrically and prevents that the differential expression of genes with different polarity - i.e. one with low expression for class 0 and the other with low expression for class 1 - cancels out when they are averaged. But even by using such simple cluster expression values as in (2), finding a partition of the index set  $\{1, \dots, p\}$  into subsets or clusters  $\{\mathcal{C}_1, \dots, \mathcal{C}_q\}$  that virtually determine the probability structure is still highly nontrivial and the design of a procedure which reveals the exact partition according to (1) is too ambitious. Thus, we develop a computationally intensive procedure that approximately solves (1) and empirically yields good results.

## 2.2 Clustering with Scores and Margins

A practical heuristic for gene clustering is the *Cluster Affinity Search Technique* (CAST) [4]. Our approach is algorithmically similar and also relies on growing the cluster incrementally by adding one gene after the other. Subsequent cleaning steps help us to remove spurious genes that were incorrectly added to the cluster at earlier stages. As in CAST, we repeat growth and removal until the cluster stabilizes and then start a new cluster. The main and very important difference is that we do not augment (shorten) the cluster by the gene that suits best (least) into the current cluster in terms of an unsupervised similarity measure, but base our strategy for supervised clustering of genes on adding (removing) the gene that improves the differential expression of the current cluster most, according to an empirical objective function for the representative group values from (2). To be more explicit, we assume now that we are given  $n$  independent and identically distributed realizations

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n), \text{ with } \mathbf{x}_j \in \mathbb{R}^p \text{ and } y_j \in \{0, 1\}, \quad (3)$$

of the random vector  $(\mathbf{X}, Y)$ , whose expression profiles  $\mathbf{x}_j$  are centered to mean zero and scaled to unit variance. The objective function needs to be a quantitative and efficiently computable measure of a cluster’s ability to discriminate the tissues. Since we aim for subsets of genes with accurate separation in binary problems, we rely on Wilcoxon’s test statistic for two unpaired samples [9], which has been also applied as a nonparametric rank-based score function for genes in [8]. The score of a single gene  $i$  is computed from its  $n$ -dimensional vector of observed values  $\xi_i = (x_{i1}, \dots, x_{in})$ ,

$$\text{Score}(\xi_i) = s(\xi_i) = \sum_{j \in \mathcal{N}_0} \sum_{l \in \mathcal{N}_1} 1_{[x_{ij} \geq x_{il}]}, \quad (4)$$

where  $x_{ij}$  is the expression value of gene  $i$  for tissue  $j$  and  $\mathcal{N}_k$  represents the set of the  $n_k$  tissues  $\in \{1, \dots, n\}$  being of type  $k \in \{0, 1\}$ . The score uses information about the type of the tissues and is thus a criterion for supervised clustering. It can be interpreted as counting, for each experiment having response value 0, the number of tissues from class 1 that have smaller expression values, and summing up these quantities. Computing the score for a gene cluster  $\mathcal{C}_i$  goes likewise via its observed representative values  $\xi_{\mathcal{C}_i} = (x_{\mathcal{C}_i,1}, \dots, x_{\mathcal{C}_i,n})$ . Viewing the score as Wilcoxon’s test statistic, it allows to order genes and clusters according to their potential significance for tissue discrimination. If the expression values of a particular gene or cluster yield exact separation of the classes, the expression values for all tissue samples having response 0 are uniformly lower than the ones belonging

### Supervised Clustering Algorithm

1. Start with the entire  $p \times n$  expression matrix  $X$ . Its rows are genes, and its columns are observations of 2 different tissue types, having zero mean and unit variance.
2. Determine the score of every gene  $i$ , i.e. every  $n$ -dimensional row of observed expression values  $\xi_i = (x_{i1}, \dots, x_{in})$  in  $X$  as in (4). Flip the sign of each gene expression vector  $\xi_i$  that has score  $s(\xi_i) > s_{max}/2$  by multiplying it with  $(-1)$ ,

$$\tilde{\xi}_i = \alpha_i \xi_i = \begin{cases} \xi_i, & \text{if } s(\xi_i) \leq s_{max}/2, \\ -\xi_i, & \text{if } s(\xi_i) > s_{max}/2. \end{cases}$$

This operation changes the score to  $s(\tilde{\xi}_i) = \min(s(\xi_i), s_{max} - s(\xi_i))$ .

3. Composition of the starting values
  - a) If no initial cluster  $\mathcal{C}$  is given, identify the gene  $i^*$  having the lowest score  $s(\tilde{\xi}_i)$ . If more than one is found, the gene  $i^*$  with the largest margin  $m(\tilde{\xi}_i)$  as in (6) is chosen. Set the initial cluster mean  $\xi_{\mathcal{C}}$  equal to the expression vector  $\xi_{i^*}$  of the chosen gene.
  - b) If an initial cluster  $\mathcal{C}$  is given, average the expression of the genes therein,

$$\xi_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{g \in \mathcal{C}} \tilde{\xi}_g = \frac{1}{|\mathcal{C}|} \sum_{g \in \mathcal{C}} \alpha_g \cdot (x_{g1}, \dots, x_{gn})$$

4. Forward Search

Average the current cluster expression profile  $\xi_{\mathcal{C}}$  with each individual gene  $i$ ,

$$\xi_{\mathcal{C}+i} = \frac{1}{|\mathcal{C}| + 1} \left( \tilde{\xi}_i + \sum_{g \in \mathcal{C}} \tilde{\xi}_g \right), \quad i = 1, \dots, p.$$

Identify the winning gene  $i^*$  as  $\arg \min_i s(\xi_{\mathcal{C}+i})$ , i.e. the gene that leads to the lowest score. If not unique, identify the winning gene  $i^*$  as the one that optimizes score *and* margin, i.e.  $i^* = \arg \min_i s(\xi_{\mathcal{C}+i})$  as well as  $i^* = \arg \max_i m(\xi_{\mathcal{C}+i})$ .

5. Repeat step 4 until the identified gene  $i^*$  is no longer accepted to enter the cluster. This is said to happen, if the score of the updated cluster expression vector  $\xi_{\mathcal{C}+i^*}$  worsens, i.e.  $s(\xi_{\mathcal{C}+i^*}) > s(\xi_{\mathcal{C}})$ , or if the score remains unchanged and the margin deteriorates, i.e.  $s(\xi_{\mathcal{C}+i^*}) = s(\xi_{\mathcal{C}})$  as well as  $m(\xi_{\mathcal{C}+i^*}) < m(\xi_{\mathcal{C}})$

6. Backward Search

Exclude each gene  $i$  of the current cluster  $\mathcal{C}$  separately and average the expression vectors of the remaining genes,

$$\xi_{\mathcal{C}-i} = \frac{1}{|\mathcal{C}| - 1} \left( \sum_{g \in \mathcal{C} \setminus \{i\}} \tilde{\xi}_g \right), \quad i \in \mathcal{C}.$$

Compute score and margin of each  $\xi_{\mathcal{C}-i}$ . Identify (as in step 4) that gene  $i^*$  whose exclusion optimizes the score, or if not unique, optimizes score and margin.

7. Repeat step 6 until the exclusion of the identified gene  $i^*$  is (according to the formulation in step 5) no longer accepted.
8. Repeat steps 4 – 7 until the cluster converges and the objective function is optimal.
9. If more than one cluster  $\mathcal{C}$  is desired, discard the genes in the former clusters from  $X$  and restart the algorithm at step 3 with the reduced, sign-flipped expression matrix.

Figure 1: Algorithm for supervised clustering of binary problems with scores and margins.

to class 1, or vice versa. In the former case, the score function returns its minimal value  $s_{min} = 0$ , in the latter case the maximum score  $s_{max} = n_0 n_1$  is assigned.

We rely on the use of log-transformed, mean centered and standardized gene expression data and thus need to prevent that averaging two discriminatory genes with different polarity - i.e. one with low expression for class 0 and the other with low expression for class 1 - cancels out the differential expression of their mean. Therefore, we aim for low expression values pointing to class 0 for all genes, which is achieved by using the sign-flipped expression  $\tilde{\xi}_i$  for all genes  $i \in \{1, \dots, p\}$ ,

$$\tilde{\xi}_i = \alpha_i \xi_i = \begin{cases} (x_{i1}, \dots, x_{in}), & \text{if } s(\xi_i) \leq s_{max}/2, \\ (-x_{i1}, \dots, -x_{in}), & \text{if } s(\xi_i) > s_{max}/2. \end{cases} \quad (5)$$

The sign-flip is equivalent to setting  $\alpha_g = -1$  in equation (2) for all genes that tend to have lower expression values for the tissues of type 1 than for tissues of type 0. After the sign-flip, the scores of all individual genes  $i$  in the expression matrix are equal to

$$s(\tilde{\xi}_i) = \min(s(\xi_i), s_{max} - s(\xi_i)),$$

and since all genes now have the same polarity, we can safely average them to compute group expression values. It is important to notice that the biological interpretation is not impeded by the sign-flips. Nevertheless for interpretational purposes, the information about them should be recorded.

During the clustering process, we typically come across different gene or cluster expression vectors that have equal score (often zero) and hence the same quality according to our objective function. This is due to the discrete range of the score function. To achieve uniqueness in the decisions which gene or cluster is optimal, we need a refinement of our objective function. We thus introduce the margin function, a continuous and real-valued measure for the strength of tissue discrimination of a sign-flipped gene expression vector  $\tilde{\xi}_i$  where low expression values point towards the tissues of class 0,

$$\text{Margin}(\xi_i) = m(\xi_i) = \min_{l \in \mathcal{N}_1} (x_{il}) - \max_{j \in \mathcal{N}_0} (x_{ij}), \quad (6)$$

where  $\mathcal{N}_0, \mathcal{N}_1$  and  $x_{ij}$  are as in (4). The margin function is positive if and only if the score is zero and  $\tilde{\xi}_i$  then perfectly separate the tissues, otherwise it is negative. It measures the size of the gap between the lowest expression value from tissues with response 1, and the highest gene expression corresponding to class 0. The larger this gap and hence the value of the margin function, the easier and clearer the discrimination of the two classes. The computation of the margin is again likewise for clusters via  $\xi_{C_i}$ . Whenever various gene or cluster expression profiles have equal score, their quality is judged by the margin function. Our objective function thus has two components. The score function is regarded with first priority, whereas the margin function serves as a criterion with second priority to achieve uniqueness.

Our clustering algorithm is detailed in figure 1. It begins with the sign-flip operation described in (5) to bring all genes to the same polarity. The clustering process can be started with or without initial gene clusters. If none are given, we start the procedure with the single gene that optimizes the objective function. Otherwise, the representative value of the starting cluster is determined. We then proceed by constructing the cluster incrementally. By searching among all genes, we merge and average the current cluster with one single gene, such that the augmented cluster optimizes our objective function, i.e. has the lowest score, or (in case of ‘‘ties’’) the largest margin. The merging process

is repeated until the objective function can no longer be improved. To remove spurious elements out of the current cluster, we then continue with a backward pruning stage, where genes are excluded step by step so that the objective function is optimized by every single removal. This cleaning stage aims to root out genes that were wrongly added to the cluster before. Accordingly, the forward and backward stages are repeated until the cluster converges, i.e. no further improvement of the objective function by adding or removing single genes is possible.

If one wishes to have more than  $q = 1$  cluster for a binary class distinction, the genes forming the first cluster are discarded from the expression matrix, and the clustering process is restarted, again with or without an initial cluster. The algorithm's computations are feasible for dimensions  $p$  and sample sizes  $n$  which are clearly beyond today's common orders and hence also applicable for microarray experiments from the future. The computing time for searching  $q = 5$  clusters in the binary leukemia dataset with  $n = 72$  observations and  $p = 3,571$  genes on a Linux PC with an Intel Pentium IV 1.6 GHz processor is about 5 seconds only. Software for the supervised clustering algorithm is available for free as an *R*-Package at <http://stat.ethz.ch/~dettling/supercluster>.

Summarizing, our cluster algorithm is a combination of variable (gene) selection for cluster membership and forming a new predictor by possible sign-flipping and averaging the gene expressions within a cluster as in (2). The cluster membership is determined with a forward and backward searching technique that optimizes the predictive score and margin criteria in (4) and (6), which both involve the supervised response variables from the data.

### 2.3 Generalization for Multiclass Problems

Here we explain the extension of the supervised clustering algorithm to multi-categorical ( $K > 2$ ) problems, where the response comprises more than two tissue types. We recommend to compare each response class separately against all other classes. This *one-against-all* approach for reduction to  $K$  binary problems is very popular in the machine learning community, since many algorithms are solely designed for binary response. It works by defining

$$Y^{(k)} = \begin{cases} 1, & \text{if } Y = k, \\ 0, & \text{else} \end{cases}$$

and running  $K$  times the supervised clustering algorithm on  $(x_1, y_1^{(k)}), \dots, (x_n, y_n^{(k)})$  as explained above. The interpretation is that we, as in (1), model the conditional probability for discrimination of the  $k$ th class versus all the other response categories as depending on a few gene subsets only,

$$\mathbb{P} [Y^{(k)} = 1 | \mathbf{X}] = f_k (X_{C_1^k}, X_{C_2^k}, \dots, X_{C_q^k}) \text{ for } k = 0, \dots, K - 1,$$

where  $f_k(\cdot)$  are nonlinear functions mapping from  $\mathbb{R}^q$  to  $[0, 1]$ .  $C_1^k, \dots, C_q^k$  are the  $q \ll p$  functional groups of genes and  $X_{C_1^k}, \dots, X_{C_q^k}$  are their representative group values, defined as in (2). When the supervised clustering algorithm is applied to each of the  $K$  binary class distinctions, this results in totally  $K \cdot q$  clusters, which can then be used to model the conditional probability for the  $K$ -class response,

$$\mathbb{P} [Y = k | \mathbf{X}] = f (X_{C_1^0}, \dots, X_{C_q^0}, \dots, X_{C_1^{K-1}}, \dots, X_{C_q^{K-1}}).$$

It is important to notice that instead of considering each class against all the other classes, many more ways to reduce a multiclass problem to multiple binary problems exist, see [11, 12] for a thorough discussion. We assume that problem dependent solutions which utilize deeper knowledge about the biological relation among the tissue types could be even more accurate for reducing multi-categorical to binary problems.

## 3 Numerical Results

### 3.1 Data

- Leukemia dataset

This dataset contains gene expression levels of  $n = 72$  patients either suffering from acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML, 25 cases) and was obtained from Affymetrix oligonucleotide microarrays. For more information see [13]; the data are available at <http://www.genome.wi.mit.edu/MPR>. Following exactly the protocol in [14], we preprocess them by thresholding, filtering, a logarithmic transformation and standardization, so that the data finally comprise the expression values of  $p = 3,571$  genes.

- Breast cancer dataset

This dataset, described in [15], monitors  $p = 7,129$  genes in 49 breast tumor samples. The data were obtained by applying the Affymetrix technology and are available at <http://www.genetics.duke.edu/microarray/Published%20work.htm>. We thresholded the raw data with a floor of 100 and a ceiling of 16,000 before applying a base 10 logarithmic transformation. Finally, each experiment was standardized to zero mean and unit variance. The response variable describes the status of the estrogen receptor (ER). According to [15], two samples failed to hybridize correctly and were excluded from their analysis. In five cases, two different clinical tests for determination of the ER status yielded conflicting results. These five plus another four randomly chosen samples were also separated from the rest of the data, so that a dataset of  $n = 38$  samples remained, of which 18 were ER+ and 20 ER-.

- Colon cancer dataset

In this dataset, expression levels of 40 tumor and 22 normal colon tissues for 6,500 human genes are measured using the Affymetrix technology. A selection of 2,000 genes with highest minimal intensity across the samples has been made in [16]. The data are available at <http://www.molbio.princeton.edu/colondata>. As for all other datasets, we process these data further by carrying out a base 10 logarithmic transformation and standardizing each tissue sample to zero mean and unit variance across the genes.

- Prostate cancer dataset

The raw data are available at <http://www-genome.wi.mit.edu/MPR/prostate> and comprise the expression of 52 prostate tumors and 50 non-tumor prostate samples, obtained from the Affymetrix technology. We use normalized and thresholded data as described in [17]. We also excluded genes whose expression varied less than 5-fold relatively, or less than 500 units absolutely between the samples, leaving us with the expression of  $p = 6,033$  genes. Finally, we applied a base 10 logarithmic transformation and standardized each experiment to zero mean and unit variance across the genes.



- **SRBCT dataset**  
This dataset was described in [18] and contains gene expression profiles for classifying small round blue cell tumors of childhood into 4 classes (Neuroblastoma, Rhabdomyosarcoma, Non-Hodgkin-Lymphoma, Ewing Family of Tumors) and was obtained from cDNA microarrays. A training set comprising 63 SRBCT tissues, as well as a test set consisting of 20 SRBCT and 5 non-SRBCT samples are available at <http://www.nhgri.nih.gov/DIR/Microarray/Supplement>. Each tissue sample is associated with thoroughly preprocessed expression profile of  $p = 2,308$  genes, already standardized to zero mean and unit variance across genes.
- **Lymphoma dataset**  
This dataset is available at <http://llmpp.nih.gov/lymphoma/data/figure1> and contains gene expression levels of the  $K = 3$  most prevalent adult lymphoid malignancies: 42 samples of diffuse large B-cell lymphoma (DLBCL, class 0), 9 observations of follicular lymphoma (FL, class 1), and 11 cases of chronic lymphocytic leukemia (CLL, class 2). The total sample size is  $n = 62$ , and the expression of  $p = 4,026$  well-measured genes, preferentially expressed in lymphoid cells or with known immunological or oncological importance is documented. More information on these data can be found in [19]. We imputed missing values and standardized the data as described in [14].
- **Brain tumor dataset**  
This dataset, presented in [20], contains  $n = 42$  microarray gene expression profiles from  $K = 5$  different tumors of the central nervous system, i.e. 10 medulloblastomas, 10 malignant gliomas, 10 AT/RTs, 8 PNETs and 4 human cerebella. The raw data originated from the Affymetrix technology and are publicly available at <http://www.genome.wi.mit.edu/MPR/CNS>. For data preprocessing, we followed the protocol in the supplementary information to [20]. After thresholding, filtering, a logarithmic transformation and standardization of each experiment to zero mean and unit variance, a dataset comprising  $p = 5,597$  genes remained.
- **NCI dataset**  
This dataset comprises gene expression levels of  $p = 5,244$  genes for  $n = 61$  human tumor cell lines which can be divided in  $K = 8$  classes: 7 breast, 5 central nervous system, 7 colon, 6 leukemia, 8 melanoma, 9 non small cell lung carcinoma, 6 ovarian and 9 renal tumors. A more detailed description of the data can be found on the website <http://genome-www.stanford.edu/nci60> and in [21]. We work with preprocessed data as in [14].

### 3.2 Results from the supervised clustering algorithm

In this section we briefly describe the results obtained by applying the supervised clustering algorithm to the above datasets. Generally, the output of the clustering procedure is very promising. In all eight datasets we analyzed, totally comprising 24 binary class distinctions, the average cluster expression  $x_C$  always perfectly discriminates the two response classes (in multiclass problems, this is one class against the rest). Hence, the scores of all clusters are equal to zero. Moreover, the clusters have strongly positive margins, indicating that the different tissue types are clearly separated. As an example, figure 2 impressively shows how well the average cluster expression vectors  $x_{C_1}$  and  $x_{C_2}$  discriminate the three response classes of the lymphoma dataset. It is intuitively clear from figure

2 that our cluster expression vectors  $x_C$  are very suitable as predictor variables for the tissue types and they indeed allow for error-free classification on the training data and also yield good results on independent test datasets, see section 3.4.

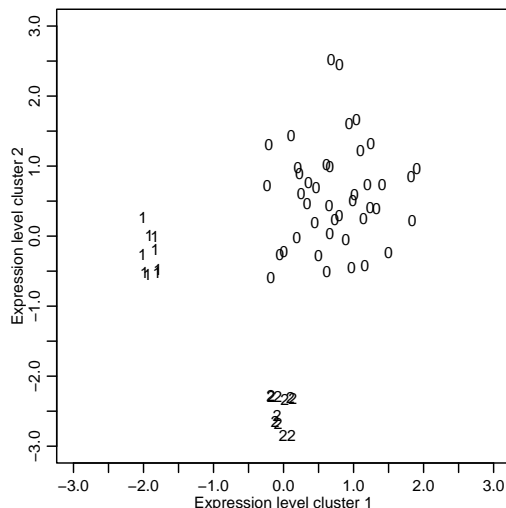


Figure 2: Lymphoma data: Average cluster expression  $x_{C_1}$  shaped for the separation of response class 1 (FL), versus response classes 0 & 2 (DLBCL & CLL) on x-axis, and  $x_{C_2}$  formed for discrimination of class 2 versus classes 0 & 1 on y-axis.

### 3.3 Permutation Test

This section is concerned with assessing relevance and addresses the question whether or not the promising output of the clustering procedure is a noise artifact. For this purpose, we explore quality measures of clusters generated from random noise gene expression data and compare them to the results obtained with the original data. Since the distributions of the score function  $s(\cdot)$  and the margin function  $m(\cdot)$  on noise are not known, we rely on simulations. Let  $(y_1, \dots, y_n)$  be the original set of responses. Then,

$$(y_1^{*(\ell)}, \dots, y_n^{*(\ell)})$$

is a “shuffled” set of responses, constructed from the original response set by a random permutation for each  $\ell = 1, \dots, L$ . We then allocate an element of the permuted response to each of the (fixed) gene expression profiles  $\mathbf{x}_i$ , giving us independent and identically distributed pairs

$$(\mathbf{x}_1, y_1^{*(\ell)}), (\mathbf{x}_2, y_2^{*(\ell)}), \dots, (\mathbf{x}_n, y_n^{*(\ell)}) \text{ for each } \ell = 1, \dots, L$$

as in (3). The supervised clustering procedure is then applied  $L = 1000$  times on such data with randomly permuted responses. For every permuted set of responses, a single cluster ( $q = 1$ ) was formed on the entire dataset and both its final score  $s^{*(\ell)}$  and margin  $m^{*(\ell)}$  were recorded.

<i>Margins</i>	$m^{(0)}$	$\max_{\ell}(m^{*(\ell)})$	$\text{med}_{\ell}(m^{*(\ell)})$	$\min_{\ell}(m^{*(\ell)})$
Leukemia	0.20	0.05	-0.01	-2.41
Breast Cancer	1.29	0.23	0.04	-0.82
Prostate	0.05	0.02	-0.04	-0.90
Colon	0.08	0.05	-0.12	-1.39
SRBCT	1.00	0.11	-0.06	-1.16
Lymphoma	1.65	0.14	0.01	-1.16
Brain	1.03	0.32	0.09	-0.29
NCI	2.52	0.44	0.12	-0.91

Table 1: Margins  $m^{(0)}$  from the original datasets, as well as maximal, median and minimal margins  $m^{*(\ell)}$  from 1000 permuted replicates, for leukemia data (AML/ALL distinction), breast cancer data (ER+/ER- distinction), prostate data (tumor/normal-distinction), colon data (tumor/normal-distinction), SRBCT data (distinction of the Ewing family of tumors versus three other tumor types), lymphoma data (distinction of DLBCL versus FL and CLL), brain tumor data (separation of atypical teratoid/rhabdoid tumors (AT/RTs) against 4 other tumor types) and NCI data (distinction of leukemia against seven other cancers).

<i>Scores</i>	$s^{(0)}$	$\min_{\ell}(s^{*(\ell)})$	$\max_{\ell}(s^{*(\ell)})$	$\#(s^{*(\ell)} = 0)/L$
Leukemia	0	0	279	0.41
Breast Cancer	0	0	43	0.91
Prostate	0	0	566	0.17
Colon	0	0	164	0.11
SRBCT	0	0	148	0.26
Lymphoma	0	0	78	0.67
Brain	0	0	11	0.98
NCI	0	0	13	0.95

Table 2: Scores  $s^{(0)}$  from the original dataset, maximal and minimal scores  $s^{*(\ell)}$  from  $L = 1000$  permuted replicates, and proportion of shuffled bootstrap trials where score 0 was achieved. The selection of data was as in table 1.

We explored the empirical distribution of the scores and margins from permuted data to judge whether the clusters found on the original datasets are of better quality than we would expect by chance. The results given in figure 3 and in tables 1 and 2 for a representative selection of data (see the caption to table 1 for details on data selection) are very satisfactory. As outlined in section 3.2, the scores  $s^{(0)}$  on the original datasets altogether are equal to zero, with clearly positive margins  $m^{(0)}$ . The parameters on the randomly permuted data are worse: The final score  $s^{*(\ell)}$  reached the minimal value of zero in a fraction varying between 11% and 98% in different datasets (e.g. 41% in figure 3). These frequencies represent a nonsignificant result in our permutation test for the score function. However, this is not very troubling, since the final margins  $m^{*(\ell)}$  for the permuted data were at best slightly positive, not indicating a clear separation of the randomly shuffled response classes. Values in the range of the margin in the original data were never achieved with any of the permuted data. This corresponds to a p-value of zero in the permutation test for our entire objective function consisting of score *and* margin.

We thus can “for surely” reject the hypothesis that the clusters found on the original data by our supervised algorithm are irrelevant and just a noise artifact. Moreover, we observed that the clusters from permuted data were much larger in size, exceeding the typical size (see also section 3.5, table 8) on non-permuted data of between 3 to 9 genes clearly, e.g. with a mean of 12.5 and a standard deviation of 3.2 genes for the AML/ALL-distinction on the leukemia dataset.

The fact that the score has highly non-significant p-values is at first sight surprising. The reason for this is that the cluster expression values  $x_{C,j}$  in (2) are highly dependent among the samples  $j = 1, \dots, n$  via the responses  $y_j$  in the supervisedly estimated cluster  $C = C(y_1, \dots, y_n)$  and the sign coefficients  $\alpha_g = \alpha_g(y_1, \dots, y_n)$ . This strong interdependence causes the unusual phenomenon that the null-distribution, assuming no association between the expression values  $X$  and the response  $Y$ , has a substantial probability for score equal to zero. The margin statistics in (6) has much better power properties than the score.

### 3.4 Predictive Potential

In this section, we will evaluate the predictive potential of the supervised clustering algorithm’s output to see if it could successfully reveal functional groups of genes. A predictor or classifier for  $K$  different tissue types is a function  $C(\cdot)$  that assigns a class label  $\hat{y}$ , based on an observed feature vector  $\mathbf{x}$ . More precisely, the classification rule here will be based on average cluster expression values  $\mathbf{x} = (x_{C_1^0}, \dots, x_{C_q^{K-1}})$  as  $K \cdot q$  features,

$$\hat{y} = C(\mathbf{x}) = C\left(x_{C_1^0}, \dots, x_{C_q^0}, \dots, x_{C_1^{K-1}}, \dots, x_{C_q^{K-1}}\right) \in \{0, \dots, K-1\}.$$

In practice, the classifier is built from a learning set of tissues whose class labels are known. Subsequently it can be used to predict the class labels of new tissues with unknown outcome. There are various methods to build classification rules based on past experience and we restrict here on two relatively simple methods which are well suited for our purpose.

#### 3.4.1 Nearest Neighbor Classification

An easy to implement and, compared to more sophisticated methods, impressively competitive classifier for microarray data is the  $k$ -nearest neighbor rule [22]. It is based on a distance function  $d(\cdot, \cdot)$  for pairs  $\mathbf{x}$  and  $\mathbf{x}'$  of feature vectors. Since we consider here standardized gene expression data, the Euclidean distance function

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{K \cdot q} (x_i - x'_i)^2}$$

is a reasonable choice. Then, for each new feature vector, the  $k$  closest feature vectors from the tissues in the learning data are identified and the predicted class is given by majority vote of the associated responses of these  $k$  closest neighbors. We found a choice of  $k = 1$  neighbors to be appropriate, but more data driven approaches via cross validation for the determination of  $k$  would be possible.

#### 3.4.2 Aggregated Trees

Another approach which proved to be very fruitful in our setting is as follows: When knowing conditional probabilities  $p_k(\mathbf{x}) = P[Y^{(k)} = 1 | \mathbf{X} = \mathbf{x}]$ , which specify how likely a

tissue with feature vector  $\mathbf{x}$  belongs to the  $k$ th or one of the other classes, the classifier function is

$$\hat{y} = C(\mathbf{x}) = \arg \max_{k \in \{0, \dots, K-1\}} p_k(x_{C_1^k}, \dots, x_{C_q^k}), \quad (7)$$

meaning that a tissue is assigned to the class with highest probability. In practice of course, we have to rely on estimated probabilities  $\hat{p}_k(\mathbf{x})$ . An often applied method for this task is the CART-algorithm for fitting classification trees [23]. Its drawback when using it with our supervised clusters as input is that in case of perfect separation of the tissues in the training data, it only uses one (the first) component  $x_{C_1^k}$  of the feature vector  $\mathbf{x}$  to determine conditional probabilities  $\hat{p}_k(\mathbf{x})$ , and does not take into account any of the useful information about the remaining  $(q - 1)$  input variables  $x_{C_2^k}, \dots, x_{C_q^k}$ . To improve the tree-based probability estimates, we design a novel technique based on plurality voting with classification trees, called *aggregated trees*. The idea is to fit  $q$  trees, one each with the  $q$  cluster expression profiles (components of the feature vector  $\mathbf{x}$ ) that have been found by our supervised algorithm for a particular binary class distinction. Each tree casts a weighted vote  $\hat{p}_{ki}(x_{C_i^k}), i = 1, \dots, q$ , for response class  $k$  against the rest. Averaging then yields

$$\hat{p}_k(\mathbf{x}) = \hat{p}_k(x_{C_1^k}, \dots, x_{C_q^k}) = \frac{1}{q} \cdot \sum_{i=1}^q \hat{p}_{ki}(x_{C_i^k}).$$

as estimated conditional probabilities, which can be plugged into (7) for maximum likelihood classification.

### 3.4.3 Empirical Study

Since except for the leukemia and SRBCT data, no genuine test sets are available, our empirical study for exploring the classification potential is based on random divisions into learning and test set as well as leave-one-out cross validation. For the latter, we set aside the  $i$ th tissue and carry out cluster identification and classifier fitting by considering only the remaining  $(n - 1)$  data points. We then honestly predict  $\hat{y}_i$ , the class label of the  $i$ th tissue sample and repeat this process for all data we have. Each observation is held out and predicted exactly once. We can determine the test set error by calculating the fraction of predicted class labels which differ from the true class labels. Results for the nearest neighbor and the aggregated tree classifier and varying number of clusters  $q$  are given in table 3.

It is known from theory (e.g. [24], p.71) that error rates from leave-one-out cross validation have low bias but large variance. Estimating error rates by repeated random splitting of the data into training and (larger) test sets may be better in terms of mean squared error. In table 4 we report misclassification rates which are based on  $N = 100$  random divisions into a learning set comprising two thirds, and a test set containing the remaining third of all  $n$  data. We took care that the class proportions were roughly identical in learning and test set. Also here in every run, both cluster identification and classifier construction are carried out on the training data, followed by honestly predicting the class labels  $\hat{y}_i$  for the test data with the two classifiers and various number of clusters  $q$ . The misclassification rate is then calculated as the averaged fraction of predicted class labels which differ from the true one.

We observe that the error estimates obtained from random splitting are on a slightly higher level than the ones from leave-one-out cross validation. We also see that introducing

<i>Leukemia</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	5.56%	5.56%	4.17%	2.78%	2.78%	2.78%	2.78%
Aggregated Trees	5.56%	5.56%	1.39%	1.39%	2.78%	2.78%	2.78%
<i>Breast</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aggregated Trees	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
<i>Prostate</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	13.73%	7.84%	4.90%	6.86%	4.90%	4.90%	5.88%
Aggregated Trees	13.73%	13.73%	6.86%	8.82%	6.86%	5.88%	5.88%
<i>Colon</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	27.42%	22.58%	22.58%	19.35%	16.13%	17.74%	19.35%
Aggregated Trees	27.42%	29.03%	19.35%	19.35%	16.13%	17.74%	17.74%
<i>SRBCT</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.59%
Aggregated Trees	3.17%	0.00%	0.00%	0.00%	1.59%	1.59%	1.59%
<i>Lymphoma</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	3.23%	1.61%	1.61%	1.61%	0.00%	0.00%	0.00%
Aggregated Trees	3.23%	1.61%	1.61%	1.61%	0.00%	0.00%	0.00%
<i>Brain</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	30.95%	23.81%	19.05%	16.67%	19.05%	16.67%	16.67%
Aggregated Trees	42.86%	23.81%	21.43%	19.05%	14.29%	11.90%	11.90%
<i>NCI</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	40.98%	40.98%	36.07%	29.51%	24.59%	27.87%	26.23%
Aggregated Trees	49.18%	47.54%	39.34%	29.51%	21.31%	21.31%	19.67%

Table 3: Misclassification rates for out-of-sample classification with  $q$  gene clusters as features, based on leave one out cross validation.

some redundancy for the discrimination process by using additional clusters, i.e. increasing  $q$ , yields better performance; but of course, a too large value of  $q$  would exhibit overfitting.

### 3.4.4 Comparison to Classification with Single Genes

Does the use of averaged cluster expression profiles from our supervised algorithm improve the classification results compared to non-averaged, individual genes? To answer this important question, we also classified our datasets with exactly the same genes that were contained in the clusters, but did not average them. Instead of  $q$  average expression profiles we then have roughly five times as many single genes as predictor variables. Misclassification rates from repeated random splitting are given in table 5. We observe that the aggregated tree classifier yields in 54 of 56 cases better results with cluster averages than with individual genes as input. Also the nearest neighbor classifier is in 43 of 56 cases better when used in conjunction with clusters than with single genes. Note that since the events are not independent, we cannot employ a binomial test for the null hypothesis of equal performance between clusters and single genes. An analysis of score and margin of the individual genes that were used in the clusters shows that most of them are not the strongest individually for predicting the tissue types, i.e. individually often only have mediocre scores and margins, but have very good predictive power as a group. So far,

<i>Leukemia</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	6.58%	4.62%	4.21%	3.75%	3.33%	3.38%	3.25%
Aggregated Trees	6.58%	6.12%	3.71%	3.54%	2.79%	2.71%	2.62%
<i>Breast</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	1.00%	0.75%	0.75%	1.00%	0.83%	1.00%	1.00%
Aggregated Trees	1.00%	1.58%	1.67%	2.33%	2.58%	2.42%	3.00%
<i>Prostate</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	14.47%	11.68%	9.62%	7.97%	7.26%	6.94%	6.91%
Aggregated Trees	14.47%	16.47%	10.32%	8.79%	8.12%	8.00%	7.79%
<i>Colon</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	23.35%	20.35%	19.10%	16.95%	16.45%	16.05%	15.95%
Aggregated Trees	23.35%	21.80%	19.70%	18.10%	16.95%	16.20%	16.45%
<i>SRBCT</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	1.33%	0.48%	0.43%	0.48%	0.76%	0.95%	1.05%
Aggregated Trees	5.76%	0.95%	0.71%	1.10%	1.76%	1.90%	2.14%
<i>Lymphoma</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	2.15%	2.20%	1.50%	0.85%	0.65%	0.50%	0.50%
Aggregated Trees	3.45%	2.45%	1.40%	0.80%	0.25%	0.20%	0.30%
<i>Brain</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	31.21%	27.50%	26.36%	24.71%	23.86%	23.71%	23.36%
Aggregated Trees	35.43%	28.43%	24.43%	22.14%	19.64%	18.29%	16.86%
<i>NCI</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	45.25%	40.25%	37.90%	34.80%	32.10%	30.50%	29.65%
Aggregated Trees	51.85%	42.35%	38.05%	34.05%	29.30%	27.75%	26.50%

Table 4: Misclassification rates for out-of-sample classification with  $q$  gene clusters as features, based on  $N = 100$  random divisions into learning set (two thirds of the data) and test set (one third of the data).

we gained evidence that our algorithm really identifies functional groups of genes whose average expression level has high explanatory power for the response classes.

### 3.4.5 Comparison to other Studies

We now classify the breast cancer validation sample of [15], which contains four randomly chosen tissues plus five instances where two different clinical tests for determination of the ER status yielded conflicting results. We choose the nearest neighbor method with  $q = 3$  clusters to be our classifier for the validation sample, since it had the best predictive potential on the  $n = 38$  training data. Our predictions shown in table 6 always agree with the class label provided on the PNAS supporting information website, which corresponds to the outcome of the immunoblot assay method.

Not only the results on the validation sample are very convincing, but also the cross validation on the  $n = 38$  training tissues is error free. This is different from the results in [15] with precedent feature selection, singular value decomposition and Bayesian binary regression, where 7 of 9 tissues in the validation sample and 36 of 38 tissues in the training sample were accurately predicted. Moreover, our result confirms that the breast cancer expression matrix contains a strong signal for discriminating the ER status.

<i>Leukemia</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	6.33%	4.79%	4.50%	4.08%	3.67%	3.75%	3.79%
Aggregated Trees	8.50%	6.04%	4.54%	3.92%	4.83%	6.79%	8.46%
<i>Breast</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	1.08%	0.83%	0.92%	1.17%	1.33%	1.50%	1.58%
Aggregated Trees	5.42%	2.50%	1.83%	2.42%	4.17%	5.42%	8.33%
<i>Prostate</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	13.24%	10.68%	9.15%	8.44%	7.76%	8.18%	7.85%
Aggregated Trees	25.47%	21.29%	18.56%	17.44%	16.65%	17.65%	18.94%
<i>Colon</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	23.40%	21.95%	20.15%	18.90%	16.65%	16.25%	15.70%
Aggregated Trees	30.95%	29.70%	30.20%	31.20%	33.55%	34.15%	34.90%
<i>SRBCT</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	1.76%	0.86%	0.81%	1.05%	1.19%	1.43%	1.48%
Aggregated Trees	4.38%	2.00%	2.62%	3.95%	6.48%	6.95%	8.43%
<i>Lymphoma</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	2.43%	2.29%	1.76%	1.05%	0.81%	0.81%	0.86%
Aggregated Trees	4.38%	2.81%	2.10%	1.00%	0.81%	1.05%	1.24%
<i>Brain</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	30.79%	29.07%	29.50%	27.57%	28.50%	28.00%	27.50%
Aggregated Trees	40.14%	35.29%	34.64%	33.50%	34.36%	34.79%	35.29%
<i>NCI</i>	$q = 1$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 15$	$q = 20$
Nearest Neighbor	39.63%	34.89%	32.84%	31.95%	30.68%	29.74%	28.95%
Aggregated Trees	56.58%	49.53%	44.84%	42.42%	39.21%	39.05%	37.79%

Table 5: Benchmark misclassification rates for out-of-sample classification with the very same but non-averaged genes from  $q$  clusters as features, based on  $N = 100$  random divisions into learning set (two thirds of the data) and test set (one third of the data).

Tumor	14	31	33	44	45	46	47	48	49
Status	Neg?	Neg?	Neg?	Neg	Pos?	Pos?	Pos	Pos	Neg
Prediction	Neg	Neg	Neg	Neg	Pos	Pos	Pos	Pos	Neg

Table 6: Classification of the validation sample with  $q = 3$  cluster expression profiles based on the training sample with 38 tumors as features and aggregated trees as predictor. The status of the tumors is according to the information provided on the PNAS-website. The question mark means that two clinical tests yielded conflicting results. Displayed here is the outcome of the immunoblot assay method.

Next, we use our method to classify the original 34 test samples in the leukemia dataset. We applied the supervised clustering algorithm on the  $n = 38$  training data, where we also fit the best predictor from our random splitting study (aggregated trees with  $q = 20$  clusters as input features) as classifier for the independent sample. Our predictions turned out to be error-free, a result which can be directly compared to [13], where 29 of 34 observations were classified correctly by a weighted voting scheme. With support vector machines, results ranging between 30 to 32 correct classifications were reported



[25]. Moreover, a full leave-one-out cross validation on the  $n = 38$  training data (results not shown) resulted in perfect classification for various  $q$ 's and also the performance for cross validation on the entire dataset with  $n = 72$  observations is competitive, compared e.g. to [26].

The SRBCT data contains an additional test set of 20 SRBCT and 5 non-SRBCT samples. We first classified the 20 SRBCT tissues with the best classifier from the random splitting study on the  $n = 63$  training samples, the nearest neighbor method with  $q = 3$  clusters as input. The predictions turned out to be error-free, approving the perfect classification with artificial neural networks and principal components as in [18], as well as the correct diagnosis obtained with multi-category support vector machines in [27]. Since aggregated trees and the 1-nearest-neighbor classifier with  $q = 3$  clusters as input are not well suited for assessing prediction strengths on the 5 non-SRBCT samples, we applied logistic discrimination and rejected every classification that was done with a probability lower than 0.95. All 5 non-SRBCT's did not exceed this threshold and were thus correctly rejected, whereas 3 of the 20 SRBCT tissues did not exceed it and could not confidently be classified either, though they were predicted correctly. Also this result, as well as our error-rate from leave-one-out cross validation on the training data which achieves the benchmark error-rate of 0% are consistent with [18, 27]. This provides more evidence that our method can at least keep up with state-of-the-art classifiers such as neural networks or support vector machines.

The five remaining microarray studies don't contain genuine test sets and we thus compare our error-rates from cross validation and random splitting against the literature. The classification of tumor versus normal prostate tissue has been evaluated with leave-one-out cross validation [17]. After precedent feature selection, an accuracy of "greater than 90%" was obtained, a results which may be beaten by our error-rate of 4.90%, which corresponds to 5 misclassifications in totally 102 samples. The colon cancer datasets has already been considered by various authors, e.g. in [26] with classifiers based on single genes, such as nearest neighbors, boosting and support vector machines in a cross validation study. Our method does not clearly improve their results, although it seems to have an edge over them. However, we could not accomplish a cross validation error-rate of 9.68%, as reported in [25] with support vector machines. The error-rates on the lymphoma, brain tumor and NCI data provide evidence that our method based on a one-against-all approach does a good job in multiclass problems, too. On the lymphoma data we observe perfect classification, thus achieve the non-to-improve benchmark. On the brain tumor data, our minimal cross validation error-rate of 11.90% is superior to the 16.67% obtained in [20] with a weighted voting algorithm. Many more misclassifications occur on the NCI than on the other datasets, due to the large number of classes and their heterogeneity. However, when comparing our predictions to the results in a broad evaluation of classifiers on the NCI data [14], they prove to be very valuable. We consistently obtained mean error-rates of less than 30% with random splitting, the optimum is 26.50% using aggregated trees with  $q = 20$  clusters, whereas the best median error-rates reported in [14] are in a range around 35% and higher.

In summary, our predictions from simple classifiers based on the supervised clustering's output can easily keep up with sophisticated methods that are based on single genes, and as table 7 shows, our supervised clusters beat the best reported results from the literature in four out of eight datasets. On three further datasets, we achieve the benchmark of perfect classification. The success of our method may be because the averaging of genes according to (2) has a variance reducing effect and yields more stable and accurate features for classification. Besides its good predictive potential, the cluster structure provided by our

	Leuk.	Breast	Prost.	Colon	SRBCT	Lym.	Brain	NCI*
Sup. Cl.	1.39%	0.00%	4.90%	16.13%	0.00%	0.00%	11.90%	26.50%
Literature	1.39%	5.26%	9.80%	9.68%	0.00%	?	16.67%	≈35%

Table 7: Best leave-one-out cross validation error-rates from our supervised clustering procedure compared to best reported results from the literature where directly comparable, references are given in the main text. \*The mean error-rate on the NCI data is based on random divisions into training and test set, and compared against the median error-rate obtained under the same framework in [14].

method is very accessible for biological interpretation and can be beneficial for functional genomics.

### 3.5 Stability

The stability of the gene clusters detected by our supervised clustering algorithm is a critical issue. The output is much more useful for functional genomics if it remains unchanged for “similar” input data. We use the bootstrap as a tool for assigning statistical significance, see [28]. We assume to have  $n$  pairs of observations  $(\mathbf{x}_i, y_i)$  with binary response  $y_i \in \{0, 1\}$ , from which we form a resampled gene expression dataset

$$(\mathbf{x}_1, y_1)^*, \dots, (\mathbf{x}_n, y_n)^*$$

of length  $n$  by drawing with replacement from the original data pairs. We can then apply our supervised algorithm to extract clusters  $\mathcal{C}_1^*, \dots, \mathcal{C}_q^*$  out of these resampled data. For an empirical study, we generated  $L = 1000$  resampled gene expression datasets of size  $n$  to explore the compositional variability of the first cluster  $\mathcal{C}_1^*$  in eight binary problems as detailed in the caption of table 8.

We first analyze the variability in cluster size. The results, summarized in table 8, show surprising stability across the eight different datasets. We observe that quite small clusters, typically made up of 3-9 genes, were found. The standard deviation in cluster size was fairly low in all eight datasets. As a next and more difficult step, we try to explore the compositional variability of the clusters. To give a rough overview which proportion of genes is actively present in the clustering process, we assess a confidence level to each individual gene  $i$ , which measures how likely it is to be clustered,

$$\pi_i = \frac{N_i}{L} = \frac{1}{L} \cdot \sum_{\ell=1}^L 1_{[gene\ i \in \mathcal{C}_1^{*(\ell)}]}, \quad i = 1, \dots, p, \quad (8)$$

where  $N_i$  is the number of the  $L$  clusters that contain gene  $i$ . The numerical results given in table 9 show that except for the colon tumor data, only a minority of genes ever entered a cluster. For the leukemia data this proportion was somewhat bigger, but still most of the genes never took part in the clustering process. More importantly, only a very small part of the genes is used frequently, i.e. more than fifty times in the thousand clusters. We conjecture that our supervised algorithm discriminates phenotypes with a small core of genes only, and in this sense it is reasonably stable.

We continue by assessing confidence levels to pairs of genes which gives a clue about pairwise interactions. We count the number  $N_{ij}$  of clusters  $\mathcal{C}_1^*$  found with our bootstrapped

<i>Clustersize</i>	Mean	Stdev	Min	Max
Leukemia	5.855	2.910	1	23
Breast Cancer	4.344	2.062	1	13
Prostate	6.327	2.373	2	17
Colon	6.642	2.733	2	20
SRBCT	4.739	1.816	1	14
Lymphoma	5.485	2.679	1	16
Brain	6.094	2.751	1	19
NCI	6.174	2.930	1	20

Table 8: Variability in size of clusters that have been shaped with the supervised algorithm, based on 1000 bootstrap replicates. Leukemia stands for distinction between AML and ALL; in the breast cancer data, the separation of the ER receptor status has been analyzed; prostate and colon stand for discrimination of normal versus tumorous tissue; in the SRBCT dataset, the Ewing family of tumors was separated against three other phenotypes; for the lymphoma dataset discrimination of DLBCL against FL & CLL was considered; in the brain tumor dataset AT/RTs were discriminated from four further malignancies; and in the NCI dataset, leukemia was separated against seven other cancers. The presented figures for the two multiclass datasets are representative for all their binary distinctions between a tumor type against all others.

<i>Active Genes</i>	$\sum_i 1_{[\pi_i > 0]}$	$\sum_i 1_{[\pi_i > 0]}/p$	$\sum_i 1_{[\pi_i > 0.05]}$	$\sum_i 1_{[\pi_i > 0.05]}/p$
Leukemia	624	17.474%	18	0.504%
Breast Cancer	128	1.803%	9	0.130%
Prostate	949	15.730%	16	0.265%
Colon	1028	51.400%	12	0.600%
SRBCT	68	2.946%	11	0.477%
Lymphoma	279	6.930%	19	0.472%
Brain	345	6.164%	21	0.375%
NCI	227	4.329%	23	0.439%

Table 9: Number and proportion of genes that ever have been used in the first cluster  $\mathcal{C}_1^*$  (first two columns), as well as number and proportion of genes that have been used for cluster  $\mathcal{C}_1^*$  in more than 50 out of the 1000 bootstrap trials (last two columns). The selection of data is identical to table 8.

gene expression datasets that both contain the genes  $i$  and  $j$ , and then divide by the number of replicates  $L$ ,

$$\pi_{ij} = \frac{N_{ij}}{L} = \frac{1}{L} \cdot \sum_{\ell=1}^L 1_{[gene\ i \in \mathcal{C}_1^{*(\ell)}]} \cdot 1_{[gene\ j \in \mathcal{C}_1^{*(\ell)}]}, \quad i, j \in \{1, \dots, p\}. \quad (9)$$

These confidence levels not only give an idea how likely the pairs are, but also provide information for functional genomics, since we can now analyze whether pairs of genes preferentially enter clusters simultaneously or not. The number of hits  $N_i$  for individual genes  $i$  follows a Binomial( $L, \pi_i$ ) distribution (given the data), and for pairs  $(i, j)$  we have that  $N_{ij}$  is Binomial( $L, \pi_{ij}$ ) (we ignore here the fact that  $\pi_i$  in (8) and  $\pi_{ij}$  in (9) are computed with  $L = 1000$  replicates instead of the theoretical  $L = \infty$ ). If there were

no attraction or repulsion between genes, the joint probability  $\pi_{ij}$  would be given by the product  $\pi_i\pi_j$  of the marginal probabilities. By calibrating the observed number of hits  $N_{ij}$  with the Binomial( $L, N_iN_j/L$ ) distribution under independence, we can test the hypothesis

$$H_0 : \pi_{ij} = \pi_i\pi_j,$$

and compute the associated p-values. Low p-values indicate significant pairs of genes. Moreover, we also distinguish between two genes which are attracting (with  $N_{ij}$  larger than expected under the null hypothesis), and which are repelling (with  $N_{ij}$  lower than expected under  $H_0$ ). We implemented an empirical analysis based on  $L = 1000$  bootstrap trials, for pairs made up of the five genes with the highest confidence levels  $\pi_i$  in the discrimination of lymphoma's class 0 (DLBCL) against the other two phenotypes. Numerical results are summarized in table 10, clone numbers and function of the genes are given in table 11. Among the 10 pairs, several significant gene pairs which are strongly attracting or repelling are present, e.g. the genes 3786 and 3804 strongly attract each other. Moreover, 78% of the clusters that contained gene 3804 also included gene 3786, again signaling a special relation among these two. An interpretation of such facts in the framework of functional genomics is beyond the scope of this paper.

<i>Numbers</i>	Gene 3786	Gene 3804	Gene 761	Gene 780
Gene 3763	184 (301)	68 (220)	144 (155)	173 (133)
Gene 3786		289 (187)	153 (132)	72 (113)
Gene 3804			136 (96)	60 (83)
Gene 761				40 (58)
<i>p-values</i>	Gene 3786	Gene 3804	Gene 761	Gene 780
Gene 3763	(-) 0.000	(-) 0.000	(-) 0.359	(+) 0.001
Gene 3786		(+) 0.000	(+) 0.055	(-) 0.000
Gene 3804			(+) 0.000	(-) 0.007
Gene 761				(-) 0.015

Table 10: Top: numbers of observed and expected (under hypothesis of independence, numbers in parentheses) gene pairs of the five most frequently clustered genes in the distinction of DLBC-lymphoma against the other two phenotypes, based on 1000 bootstrap replicates. Bottom: p-values for attraction (+) and repulsion (-) of gene pairs from two-sided binomial tests that compare the joint probability against the product of the marginals.

It is now tempting to extend this kind of analysis from pairs to tuples of third and higher orders. But estimating higher-order interactions will become very unreliable due to the limited amount of sample size  $n$ .

### 3.6 Additional Modifications

Our supervised clustering procedure can be understood as a generic method and allows alteration of various details according to the users choice and specific demands. We also tried to improve the supervised clustering procedure ourselves with additional modifications, the most important of which are described here in the following. The averaging of the gene expression in (2) is specified by the arithmetic mean plus sign-flips, a very simple

Sign	Gene	Clone	Function
-	3763	769861	CD63 antigen (melanoma 1 antigen)
-	3786	345538	Cathepsin L
-	3804	343867	Allograft-inflammatory factor-1 <i>or</i> Interferon gamma induced macrophage protein <i>or</i> Ionized calcium binding adapter molecule 1
+	761	1341294	Unknown
+	780	1334411	Unknown UG Hs.32553 ESTs

Table 11: Clone numbers and function description of the five genes that have been clustered most frequently in the discrimination of DLBC-lymphoma against the other two phenotypes in the lymphoma dataset.

linear combination of genes, since it is impractical to repeatedly optimize a general linear combination such as

$$X_{\mathcal{C}_i} = \sum_{g \in \mathcal{C}_i} \beta_g X_g \quad \text{with} \quad \sum_g |\beta_g| = 1$$

during the clustering process. But theoretically, once the cluster algorithm has done its work, we could try to improve the discriminatory power of the actual cluster by numerically optimizing a weighted linear combination as above with respect to score and margin. In practice, we recognized that the numerical optimization was very difficult. If we started it with equal weights, they only changed slightly and the objective function, i.e. the margin did not improve much. Because of this we favor the more simple method.

Since the margin function in (6) is not scale-invariant, we also considered clustering with an adjusted margin. This means that we optimized the quotient of margin and within group variation for a gene expression vector  $\xi_i = (x_{i1}, \dots, x_{in})$ ,

$$\text{Adjusted margin}(\xi_i) = \frac{\text{Margin}(\xi_i)}{\sqrt{s_0^2/n_0 + s_1^2/n_1}}.$$

Here,  $n_k$  is the size and  $s_k^2$  is the sample variance of class  $k \in \{0, 1\}$ . While theoretically, the size of the gap between the two response classes is meaningful only in relation to the within-group variance, the adjustment of the margin proved not to be very important in practice, due to the use of standardized gene expression data. It did not improve the predictive performance of the clusters and slightly deteriorated their stability. Since it is common practice to standardize expression data, we recommend to work with the non-adjusted margin.

Our algorithm, as described in section 2, yields disjoint clusters of genes. To account for the fact that genes may function in multiple pathways one could modify as follows. (i) run the clustering algorithm on the data, producing a first cluster; (ii) compute a probability estimate for  $P[Y = 1|X]$  for a two-class problem, e.g. with probability based classification methods or in a logistic model; (iii) reweight the data with weights as in the Real AdaBoost algorithm ([29]; algorithm 2, p.340); (iii) return to (i) but now with reweighted data. Doing the loop  $q$  times produces  $q$  clusters which are allowed to be non-disjoint.

We also explored the improvement of the supervised clustering algorithm by biasing it towards larger clusters. Specifically, we did not stop the forward search when score and/or

margin first worsened, but continued as long as the objective function remained within a factor of the best. Our intention was that the objective function could improve again and reach even better values. As soon as the objective function once dropped below the tolerance (a factor times the best ever achieved value), we stopped the forward search and continued the algorithm with the cluster that yielded the best parameters ever. While our first guess was that the biasing could result in larger clusters with clearer separation, it rarely ever had any effect in practice.

## 4 Conclusions

We have proposed an algorithm for supervised clustering of genes from microarray experiments. Our procedure is potentially useful in the context of medical diagnostics, since it identifies groups of interacting genes that have high explanatory power for given tissue types, and which in turn can be used to accurately predict the class labels of new samples. At the same time, such gene clusters may reveal insights into biological processes and may be valuable for functional genomics.

In summary, our algorithm tries to cluster genes such that the discrimination of different tissue types is as simple as possible. It builds the clusters incrementally and relies on a fast, stepwise strategy that allows exhaustive search among thousands of genes. More specifically, the aim is to identify sparse linear combinations of genes, whose average expression level is uniformly low for one response class and uniformly high for the other class(es).

In empirical studies, the average cluster expression profiles showed superior classification potential compared to other techniques where unclustered genes had been used. The clusters showed reasonable stability and there are several reasons that point towards their biological significance: a) they do not only contain the genes which are individually good, but groups of genes whose consensus expression profile is best with respect to the objective function; b) the predictive potential of the very same, non-averaged genes cannot keep up with the one of the corresponding cluster means; and c) an application of our algorithm to randomly permuted data points out that the identified structure is more than just a noise artifact.

An important task which remains to be addressed in future research is the generalization of the supervised clustering algorithm to quantitative response variables and to censored survival data. The fundamental idea of supervised clustering can be pursued again, but needs alternative objective functions that rank individual genes and gene clusters based on their explanatory power for non-categorical response variables.

**Acknowledgments.** We thank Jane Fridlyand for providing the preprocessed NCI data.

## References

- [1] Weinstein J, Myers T, O'Connor P, Friend H, Fornace Jr A, Kohn K, Fojo T, Bates S, Rubinstein L, Anderson N, Buolamwini J, van Osdol W, Monks A, Scudiero D, Sausville E, Zaharevitz D, Bunow B, Viswanadhan V, Johnson G, Wittes R, Paull K: **An Information-Intensive Approach to the Molecular Pharmacology of Cancer**. Science 1997, **275**: 343–349.

- [2] Eisen M, Spellman P, Brown P, Botstein D: **Cluster Analysis and Display of Genome-Wide Expression Patterns**. Proceedings of the National Academy of Science 1998, **95**: 14863–14868.
- [3] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T: **Interpreting Patterns of Gene Expression with Self-Organizing-Maps: Methods and Application to Hematopoietic Differentiation**. Proceedings of the National Academy of Science 1999, **95**: 2907–2912.
- [4] Ben-Dor A, Shamir R, Yakhini Z: **Clustering Gene Expression Patterns**. Journal of Computational Biology 1999, **6**: 281–297.
- [5] Hastie T, Tibshirani R, Botstein D, Brown P: **Supervised Harvesting of Expression Trees**. Genome Biology 2001, **1**: 1–12.
- [6] Nguyen D, Rocke D: **Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data**. Bioinformatics 2002, **18**: 39–50.
- [7] Geladi P, Kowalski B: **Partial Least Squares Regression: A Tutorial**. Analytica Chimica Acta 1986, **185**: 1–17.
- [8] Park P, Pagano M, Bonetti M: **A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data**. Pacific Symposium on Biocomputing 2001, **6**: 52–63.
- [9] Wilcoxon F: **Individual Comparisons by Ranking Methods**. Biometrics 1945, **1**: 80–83.
- [10] Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Botstein D, Brown P: **Gene Shaving as a Method of Identifying Distinct Sets of Genes with Similar Expression Patterns**. Genome Biology 2000, **1**: 1–21
- [11] Hastie T, Tibshirani R: **Classification by Pairwise Coupling**. Annals of Statistics 1998, **26**: 451–471
- [12] Allwein E, Schapire R, Singer Y: **Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers**. Journal of Machine Learning Research 2000, **1**: 113–141.
- [13] Golub T, Slonim D, Tamayo P, Huard C, Gassenbeek M, Coller H, Loh M, Downing J, Caliguri M, Bloomfield C, Lander E: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**. Science 1999, **286**: 531–538
- [14] Dudoit S, Fridlyand J, Speed T: **Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data**. Journal of the American Statistical Association 2002, **97**: 77–87.
- [15] West M, Blanchette C, Dressman H Huang, E Ishida, S Spang, R Zuzan, H Olson, J Marks, J, Nevins J: **Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles**. Proceedings of the National Academy of Science 2001, **98**: 11462–11467.

- [16] Alon U, Barkai N, Notterdam D, Gish K, Ybarra S, Mack D, Levine A: **Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays**. Proceedings of the National Academy of Science 1999, **96**: 6745–6750.
- [17] Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D’Amico A, Richie J, Lander E, Loda M, Kantoff P, Golub T, Sellers W: **Gene Expression Correlates of Clinical Prostate Cancer Behavior**. Cancer Cell 2002, **1**: 203–209.
- [18] Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C, Meltzer P: **Classification and Diagnostic Prediction of Cancer Using Gene Expression Profiling and Artificial Neural Networks**. Nature Medicine 2001, **6**: 673–679.
- [19] Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson J, Lu L, Lewis D, Tibshirani R, Sherlock G, Chan W, Greiner T, Weisenburger D, Armitage J, Warnke R, Levy R, Wilson W, Grever M, Byrd J, Botstein D, Brown P, Staudt L: **Distinct types of diffuse large B-Cell-Lymphoma Identified by Gene Expression Profiling**. Nature 2000, **403**: 503–511.
- [20] Pomeroy S, Tamayo P, Gaasenbeek M, Sturla L, Angelo M, McLaughlin M, Kim J, Goumnerova L, Black P, Lau C, Allen J, Zagzag D, Olson J, Curran T, Wetmore C, Biegel J, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis D, Mesirov J, Lander E, Golub T: **Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression**. Nature 2002, **415**: 436–442.
- [21] Ross D, Scherf U, Eisen M, Perou C, Rees C, Spellman P, Iyer V, Jeffrey S, Van de Rijn M, Waltham M, Pergamenschikov A, Lee J, Lashkari D, Shalon D, Myers T, Weinstein J, Botstein D, Brown P: **Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines**. Nature Genetics 1999, **3**: 227–235.
- [22] Fix E, Hodges J: **Discriminatory Analysis – Nonparametric Discrimination: Consistency Properties**. Report 1951, No. 4, US Air Force School of Aviation Medicine, Random Field, Texas. Published in Agrawala A: *Machine Recognition of Patterns*. New York: IEEE Press, 1977.
- [23] Breiman L, Friedman J, Olshen R, Stone C: *Classification and Regression Trees*. Wadsworth: Belmont, 1984.
- [24] Ripley B: *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.
- [25] Furey T, Cristianini N, Duffy N, Bednarski D, Schummer M, Haussler D: **Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data**. Bioinformatics 2000, **16**: 906–914.
- [26] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue Classification with Gene Expression Profiles**. Journal of Computational Biology 2000, **7**: 559–583



- [27] Lee Y, Lee C: **Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data**. Technical Report **1051**: Department of Statistics, University of Wisconsin, Madison, 2002.
- [28] Efron B, Tibshirani R: **The Problem of Regions**. Annals of Statistics 1998, **26**: 1687–1718.
- [29] Friedman J, Hastie T, Tibshirani R: **Additive Logistic Regression: A Statistical View of Boosting**. Annals of Statistics 2000, **28**: 337–407 (with discussion).

