# Class Discovery in Gene Expression Data: Characterizing Splits by Support Vector Machines

Florian Markowetz and Anja von Heydebreck

Max-Planck-Institute for Molecular Genetics
Computational Molecular Biology
Ihnestrasse 63-73, D-14195 Berlin, Germany
{florian.markowetz, anja.heydebreck}@molgen.mpg.de

**Abstract.** We present a variation of the class discovery method for microarray data described by Heydebreck et al. (2001). The objective is to discover biologically relevant structures in the gene expression profiles of different tissue samples in an unsupervised fashion. Our method searches for binary partitions in the set of samples that show clear separation. Mathematically, each class distinction is characterized according to the size of margin achieved by a support vector machine (SVM) separating the two classes. In three data sets from cancer gene expression studies the SVM margin approach succeeds in detecting relationships between the tissue samples. The known biological classes (cancer subtypes) exhibit an exceptionally large value of the SVM margin.

## 1 Introduction

In microarray experiments different tissue samples are characterized by profiles of gene expression levels. The aim of our work is to find class distinctions among a set of tissue samples which show a clear separation with respect to a subset of genes. In (Heydebreck et al. (2001)) this class discovery problem is approached by a method called ISIS (for "identifying splits with clear separation"). It consists of two steps: First, based on the classification method *Diagonal Linear Discriminant Analysis* a score function (the so called *DLD score*) is proposed to quantify the degree of separability of a given binary class distinction of the set of samples. This score function is defined on the graph of all bipartitions of the set of samples. In the second step all bipartitions are declared as interesting which represent local maxima in this graph, i. e. for which the score does not increase if the class label of a single sample is changed. Since an exhaustive search over all bipartitions is in general not feasible, Heydebreck et al. (2001) propose a fast heuristic to find candidate partitions as starting points for a search of local maxima.

In this paper we will contrast the DLD score with a measure of separability based on *Support Vector Machines*. This results in a variation of the original ISIS algorithm called SVM-ISIS. We show that SVM-ISIS detects the known tumor subtypes present in three example data sets in an unsupervised fashion.

## 2 Methods

### 2.1 Microarray gene expression data

Microarrays allow to quantitate the expression of thousands of genes in parallel and thus to observe gene expression variation in a variety of human tumours (Chipping forecast (1999)). Mathematically, the result of a gene expression study is a matrix $X = (x_{gj})$, whose columns correspond to tissue samples ($j = 1, \ldots, n$) and whose rows correspond to genes ($g = 1, \ldots, k$). Using this matrix, each tissue sample is interpreted as a point in $\mathbb{R}^k$.

### 2.2 The graph of bipartitions

Two subsets $M, \overline{M}$ of the set of samples $\{1, \ldots, n\}$ define a *bipartition* or *split* $\mathcal{B} = \{M, \overline{M}\}$ of this set if $M \cap \overline{M} = \emptyset$ and $M \cup \overline{M} = \{1, \ldots, n\}$. Let $\Gamma$ be the graph whose vertex set is the set of all bipartitions of $\{1, \ldots, n\}$. Two different vertices are joined by an edge (are neighbors) iff they differ only by the class assignment of a single sample.

We will now define a score $S(\mathcal{B})$ on the vertex set of $\Gamma$ that measures how clearly the two classes representing a given bipartition are separated by gene expression levels.

### 2.3 The SVM margin score

Given a bipartition $\mathcal{B} = \{M, \overline{M}\}$ we separate the samples in $M$ from the samples in $\overline{M}$ by a Support Vector Machine with linear kernel, i. e. by a hyperplane yielding a maximal margin of separation (Figure 1).

A hyperplane $\mathcal{H} = \{ x \mid \langle w, x \rangle + b = 0\}$ is defined by its normal vector $w$ and offset $b$. It can be shown that the maximal margin hyperplane is constructed by solving the following constrained optimisation problem (Vapnik (1998)). For $i = 1, \ldots, n$

$$\begin{aligned} &minimize \ \ \frac{1}{2}\|w\|^2 \\ &subject\ to \ \ y_i(\langle w, x_i \rangle + b) - 1 \geq 0, \end{aligned}$$

where $x_i \in \mathbb{R}^k$ are the data points (with $k$ the number of genes) and $y_i = +1$ if $i \in M$ and $y_i = -1$ if $i \in \overline{M}$. The margin $\gamma(\mathcal{H})$ of the separating hyperplane depends on the normal vector $w$ by $\gamma(\mathcal{H}) = 2/\|w\|$.

With this background we define the *SVM margin score* $S(\mathcal{B})$ of a bipartition $\mathcal{B} = \{M, \overline{M}\}$ as the margin achieved by a Support Vector Machine separating the samples in $M$ from the samples in $\overline{M}$.

Our goal is to find bipartitions of the set of samples with high SVM score. Since the total number of bipartitions of $n$ samples equals $2^n$, we can not compute the score function for all of them. Instead, we use the same heuristic as described in Heydebreck et al. (2001). Here we will give a short summary.

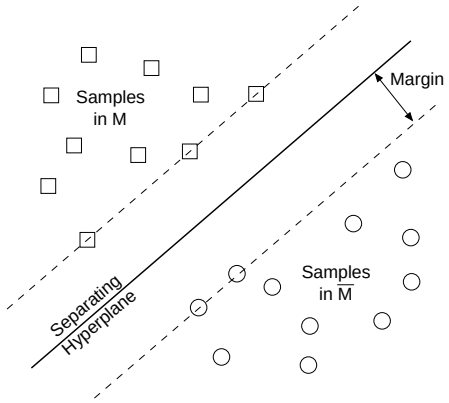


**Fig. 1. The margin of separation.** The two subsets of samples are separated by a hyperplane. The distance between the hyperplane and the nearest sample is the margin. In Support Vector Machines, the separating hyperplane is chosen such that the margin is maximized. The points on the dashed lines are called *support vectors*, because they alone determine the separating hyperplane.

### 2.4 Finding candidate partitions

We begin with the data matrix $X = (x_{gj})$ with $j = 1, \ldots, n$ and $g = 1, \ldots, k$. From this we compute average expression profiles of clusters of genes obtained by a hierarchical clustering. This yields a new data matrix $Y = (y_{ij})$, the rows of which are the cluster average profiles.

For every gene cluster $i$ and every sample $j^* = 1, \ldots n$, the value $y_{ij^*}$ defines a bipartition given by the subsets $M^- = \{\ j \mid y_{ij} \leq y_{ij^*}\}$ and $M^- = \{\ j \mid y_{ij} > y_{ij^*}\}$. Whenever both $M^-$ and $M^+$ have at least two elements, we compute the two-sample $t$-statistic. We argue that a large value of $t$ provides evidence for an interesting bipartition defined by the cut point $y_{ij^*}$, with a strong separation of the two classes by the expression levels of the genes belonging to cluser $i$. We order the splits by the value of the $t$-statistic and take the 50 top-scoring partitions as candidates for the further search.

### 2.5 Feature selection

In highdimensional data sets, classification often benefits from a selection of a subset of variables which show the strongest correlation with the class distinction of interest. We measured this correlation by the two-sample $t$-statistic $t_g(\mathcal{B})$ for each gene $g$:

$$t_g(\mathcal{B}) = \frac{\mu_{gM} - \mu_{g\overline{M}}}{\sqrt{(m-1)\sigma^2_{gM} + (\overline{m}-1)\sigma^2_{g\overline{M}}}} \times \sqrt{\frac{m\overline{m}(n-2)}{n}},$$

where $m = \|M\|$ and $\overline{m} = n - m$. For each split $\mathcal{B}$ we selected the 50 genes with highest absolute value of $t_g(\mathcal{B})$.

### 2.6 Local search for maxima

From each bipartition $\mathcal{B}$ obtained by this procedure, we proceed along a path in $\Gamma$ to a local maximum of the SVM score in a greedy manner: Starting at $\mathcal{B}$, we choose in each step the neighboring vertex with the highest SVM score until a local maximum is reached. The resulting high-scoring bipartitions can then be graphically displayed as in Figures 2–5.

## 3 Example data sets

The performance of the SVM score was tested on three different data sets: leukemia, lymphoma/leukemia and a melanoma data set. We will shortly describe the datasets and sum up the important features in Table 1.

| Dataset | # samples total | # genes used (total) | classes: # samples |
|---|---|---|---|
| **Leukemia** Golub *et al.* (1999) | 72 | 4,000 (6,817) | AML: 25<br>B-cell ALL: 38<br>T-cell ALL: 9 |
| **Lymphoma / leukemia** Alizadeh *et al.* (2000) | 62 | 4,026 (17,856) | CLL: 11<br>FL: 9<br>DLBCL-G: 21<br>DLBCL-A: 21 |
| **Melanoma** Bittner *et al.* (2000) | 31 | 3,613 (6,971) | cluster: 19<br>remaining: 12 |

**Table 1. Overview of example data sets.** Listed are the names of the data sets with reference, the total number of samples, the number of genes used in our computations compared to the total number of genes on the chips and in the last column the name of the subclasses with number of samples therein.

**Leukemia:** described by Golub et. al. (1999). For a set of 72 acute leukemia mRNA samples, expression levels of 6,817 genes were measured with Affymetrix oligonucleotide arrays. Subclasses are *acute myeloid leukemia* (AML, 25 samples) and *acute lymphoblastic leukemia* (ALL, 47 samples), which is further split into the subtypes B-cell ALL (38 samples) and T-cell ALL (9 samples). Our analysis is based on 4,000 genes with highest median expression levels over the samples.

**Lymphoma/Leukemia:** (Alizadeh et. al. (2000)) Expression profiles of 62 lymphoma and leukemia samples were produced with a "Lymphochip" containing 17,856 cDNA clones. A subset of 4,026 clones was selected by the authors for being "well measured" across the samples. The samples represent the following types of lymphoid malignancies: *chronic lymphocytic leukemia* (CLL, 11 samples); *follicular lymphoma* (FL, 9 samples) and *diffuse large*

*B-cell lymphoma* (DLBCL, 42 samples), which is further subdivided into *germinal center B-cell like DLBCL* (DLBCL-G, 21 samples) and *activated B-cell like DLBCL* (DLBCL-A, 21 samples).

**Melanoma:** described by Bittner et. al. (2000). 31 RNA samples of cutaneous melanoma were investigated by hybridization to a cDNA array representing 6,971 genes. We used the data from 3,613 clones selected by the authors for being "strongly detected" across the samples. By multidimensional scaling and different cluster algorithms, the authors identified a cluster of 19 samples separated from the remaining 12 samples.

## 4 Results

### 4.1 Significance of the SVM score

To investigate the significance of the SVM margin score in characterizing phenotypical class distinctions, we compared its outcome on the biological partitions in the three data sets to the outcome on 10,000 random partitions of the data. The results are shown in Table 2.

For each tumor subtype present in the data set, its SVM margin score, the distance to the next local maximum and the SVM margin at this local maximum is shown. The numbers in the brackets show the percentage of random splits with equal or less distance to the next local maximum or a higher SVM score.

| Dataset | Class | distance to local max. | SVM margin | SVM margin at local max. |
|---|---|---|---|---|
| **Leukemia** | AML | 3 (58.7%) | 1.6 (0%) | 2.2 (0%) |
| | T-cell ALL | 0 (0.4%) | 2.6 (0%) | 2.6 (0%) |
| | B-cell ALL | 1 (14.8%) | 1.9 (0%) | 2.8 (0%) |
| **Lymphoma/ leukemia** | CLL | 2 (13.5%) | 2.0 (0%) | 2.8 (0.1%) |
| | FL | 0 (0.2%) | 2.4 (0%) | 2.4 (0.3%) |
| | DLBCL-G | 3 (24.6%) | 1.7 (0%) | 2.0 (4.7%) |
| | DLBCL-A | 5 (49%) | 1.7 (0%) | 2.9 (0.1%) |
| **Melanoma** | cluster of 19 samples | 1 (8.8%) | 1.2 (0%) | 2.0 (10.2%) |

**Table 2. Significance of the SVM score.** The SVM score of the biological partitions in the example data sets is compared to the performance on 10,000 random partitions. In brackets: percentage of random partitions with a smaller or equal distance to the next local maximum or a higher SVM score.

One can see that proximity to a local maximum is not rare in the random splits, while the scores of all cancer subtypes are very high compared to the sampled random splits. This indicates that a high score is much more significant as a small distance to a local maximum alone. The same was observed for the DLD-score in (Heydebreck er al. (2001)).

### 4.2 Output of SVM-ISIS

The rows of the matrices in Figures 2–5 show the top scoring bipartitions found by SVM-ISIS. They are ordered by their SVM margin score which is displayed to the right of each row. The columns are arranged according to cancer subtypes with no particular order within these types.

**Leukemia dataset.** [Figure 2] All three classes are recovered individually: AML in row 4 (3 errors) and row 10 (2 errors), T-cell ALL in row 2 (without error) and B-cell ALL in row 1 (1 error).

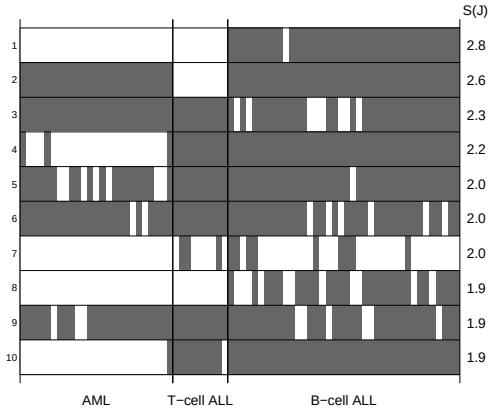


**Fig. 2. Partitions of leukemia samples.**

**Melanoma dataset.** [Figure 3] The class distinction identified by Bittner et al. (2000) as biologically meaningful is found in row 3 of the table of partitions with one error.

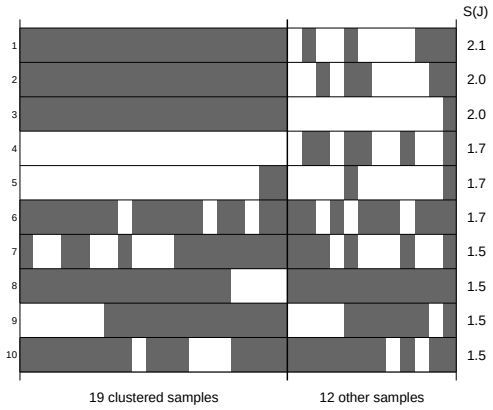


**Fig. 3. Partitions of melanoma samples**

**Lymphoma/leukemia dataset.** [Figure 4, Figure 5] First, we tried SVM-ISIS on the dataset by selecting the 50 genes with highest $t$-statistic for each split. This does not achieve satisfying results: only the split between CLL+FL and DLBCL is found (see in Figure 4: rows 1,2 and 3 with few misclassifications), but none of the classes is detected individually.

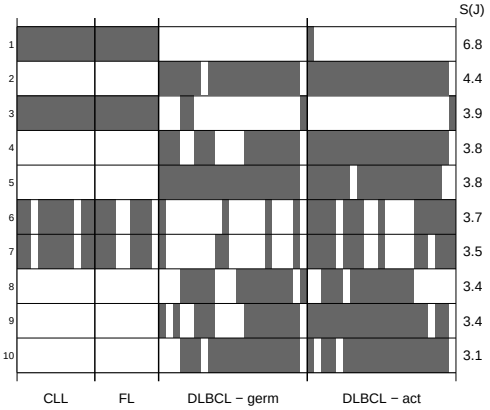


**Fig. 4. Partitions of lymphoma/leukemia samples** on 50 selected genes

In this situation we tried the performance of SVMs on high-dimensional data. Figure 5 shows the results of SVM-ISIS on the data set without feature selection. Row 4 detects CLL and row 8 FL. Even if germinal center B-cell like DLBCL is still not clearly separated from activated B-cell like DLBCL, this is an serious improvement to Figure 4. In the other two data sets doing witut feature selection for each split did not improve the results.

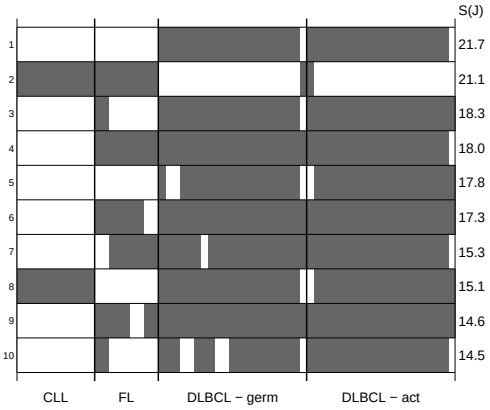


**Fig. 5. Partitions of lymphoma/leukemia samples** without feature selection

## 5 Discussion

We have introduced the SVM margin as a criterion to characterize the cancer subtypes represented in three gene expression data sets. Applying this criterion, we recovered these subtypes in an unsupervised fashion without using prior knowledge.

The behavior of classification methods on high-dimensional data sets, such as microarray data, is far from being understood. There is the danger of overfitting to training data, if the variance of the parameter estimates determining the classification rule is high. Our results provide empirical evidence that linear SVMs are not very prone to overfitting at least on the microarray data sets studied: The true biological class distinctions are clearly among those with the biggest SVM margin, or, in other words, it is practically impossible for a random bipartition to achieve a similarly large, but meaningless margin.

Further steps in the analysis could be to identify the genes that are differentially expressed across the samples, or to examine the geometry of the separating hyperplane: which samples are support vectors, which violate the margin or are misclassified? By this one can see whether in some samples the class assignment is unclear. The stability of the class distinction may be assessed by observing the change in the SVM margin due to reassignments of single samples. Thus, our method provides an auspicious beginning for a more detailed analysis of the data.

## References

ALIZADEH et al. (2000): Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403, 503-511.*

BITTNER et al. (2000): Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature, 406, 536-540.*

CHIPPING FORECAST (1999): The chipping forecast. *Special supplement to Nature Genetics, volume 21.*

DUDOIT, S., FRIDLYAND, J., and SPEED, T. (2002): Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97, 77–87.*

GOLUB et al. (1999): Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286, 531-537.*

VON HEYDEBRECK,A., HUBER,W., POUSTKA,A. and VINGRON,M. (2001): Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics, Vol. 17 Suppl.1, S107–S114.*

VAPNIK, V. (1998): *Statistical Learning Theory* Wiley, New York.