BIOINFORMATICS

Vol. 1 no. 1 2001
Pages 1–8

Identifying Splits with Clear Separation: A New Class Discovery Method for Gene Expression Data

Anja von Heydebreck[1], Wolfgang Huber[2], Annemarie Poustka[2] and Martin Vingron[1]

[1]Division of Computational Molecular Biology, Max–Planck–Institute for Molecular Genetics, Ihnestr. 73, D–14195 Berlin, Germany and [2]Division of Molecular Genome Analysis, German Cancer Research Center, INF 280, D–69120 Heidelberg, Germany

Received on February 5, 2001; revised and accepted on April 2, 2001

ABSTRACT

We present a new class discovery method for microarray gene expression data. Based on a collection of gene expression profiles from different tissue samples, the method searches for binary class distinctions in the set of samples that show clear separation in the expression levels of specific subsets of genes. Several mutually independent class distinctions may be found, which is difficult to obtain from most commonly used clustering algorithms. Each class distinction can be biologically interpreted in terms of its supporting genes. The mathematical characterization of the favored class distinctions is based on statistical concepts. By analyzing three data sets from cancer gene expression studies, we demonstrate that our method is able to detect biologically relevant structures, for example cancer subtypes, in an unsupervised fashion.

Contact: heydebre@molgen.mpg.de

INTRODUCTION

Microarrays provide a powerful tool to investigate the relationships between phenotypes of cells and their molecular properties, which can lead to a better understanding e. g. of the biology of cancer. Important topics in the analysis of microarray gene expression data are *class prediction* and *class discovery*. Whereas in class prediction the aim is to assign tissue samples to phenotypically characterized categories, in class discovery it is the detection of previously unknown relationships among genes, among tissues, or between genes and tissues.

For several data sets from cancer gene expression studies, the feasibility of discrimination between different types of tumors has been demonstrated. Classification methods such as nearest neighbor classifiers, linear and quadratic discriminant analysis, decision trees and support vector machines have been used for this purpose (Golub *et al.*, 1999; Dudoit *et al.*, 2000; Califano *et al.*, 2000; Slonim *et al.*, 2000; Ben-Dor *et al.*, 2000).

On the other hand, previously unrecognized subtypes of cancer have been discovered through the analysis of microarray gene expression data. In (Alizadeh *et al.*, 2000), two subtypes of diffuse large B-cell lymphoma with significantly different survival rates are detected, and in (Bittner *et al.*, 2000), two subtypes of cutaneous melanoma with differences in cell motility and invasiveness are identified. One can well imagine that clinical categories of tumors could in many cases be refined through the analysis of microarray gene expression data. Standard methods for such *class discovery* tasks include various clustering algorithms (e. g. hierarchical clustering or self–organizing maps; for applications to microarray data see (Eisen *et al.*, 1998; Alon *et al.*, 1999; Ben-Dor *et al.*, 1999; Golub *et al.*, 1999)), as well as dimension reduction techniques such as principal component analysis or multidimensional scaling. When such methods, which are based on global similarity measures, are applied to expression profiles of tissue samples, differences in expression levels of thousands of genes are reduced to a single value that represents the similarity or distance between two samples. However, biologically relevant relationships between samples may be much more complex: A grouping of the samples with respect to one attribute will in general be independent of a grouping with respect to another, with each of them possibly marked by differential expression of a different subset of genes.

In this paper, we develop a new method for finding interesting class distinctions among a set of tissue samples for which expression measurements over a set of typically thousands of genes are available. In cancer gene expression studies, such class distinctions might reflect biological categories like cell type, mutational status, response to a certain drug or tumor progression, but also differences in the experimental protocol.

Previous work on classification of tumor tissue samples based on gene expression profiles has shown that in many cases, cancer types can be discriminated using only a small subset of genes whose expression levels strongly correlate with the class distinction (Golub *et al.*, 1999; Dudoit *et al.*, 2000). Motivated by this fact, we try to find binary class distinctions among the set of tissue samples that show a clear separation with respect to a subset of genes. As several such bipartitions may exist independently of each other, they can be difficult to detect by usual cluster algorithms, as these either yield a single partition of the set of samples into clusters or a dendrogram.

Our approach to this class discovery problem, which we call ISIS (for “identifying splits with clear separation”), consists of two steps: First, we propose a score function which we call *diagonal linear discriminant (DLD) score*. For each binary class distinction of the set of samples, it quantifies how strongly the two classes are separated by the expression levels of a suitable subset of genes. We focus our attention on bipartitions of the set of samples for which the DLD score does not increase if the class label of a single sample is changed. In other words, these bipartitions represent local maxima in the graph of all bipartitions of the set of samples. We demonstrate that real cancer types in several example data sets are indeed characterized by high values of the DLD score and are close to local maxima.

Second, in order to find high scoring local maxima of the DLD score, we employ a fast heuristic that uses a large set of average expression profiles of clusters of genes as its input (e. g. all clusters produced by a hierarchical clustering algorithm). For each of these average profiles, we check whether it suggests one or more binary class distinctions of the set of samples. The obtained candidate bipartitions are then used as starting points for a search of local maxima of the DLD score in the graph of all bipartitions. We show that ISIS detects in an unsupervised fashion the known cancer subtypes present in three example data sets. Furthermore, several other potentially meaningful class distinctions are found.

METHODS

Microarray gene expression data

In microarray gene expression studies, estimated abundances of thousands of mRNA species in different tissue samples are obtained through hybridization to oligonucleotide or cDNA arrays (Chipping forecast, 1999). In general, the raw data have to be corrected for different experimental conditions by a ’normalization’ procedure, see e. g. (Beißbarth *et al.*, 2000). After this pre–processing step, we apply a logarithmic transformation to the absolute intensities or ratios (in the case of competitive hybridization of two separately labelled mRNA samples). This gives a data matrix $\boldsymbol{X} = (x_{gj})$, whose rows correspond to genes ($g = 1, \dots, k$), and whose columns correspond to tissue samples ($j = 1, \dots, n$). We assume that exactly one value for each gene/sample pair is given, which may be achieved by averaging over repeated measurements for samples or genes.

DLD score of bipartitions

Two subsets $M, \overline{M}$ of the set of samples $\{1, \dots, n\}$ define a *bipartition* or *split* $\mathcal{B} = \{M, \overline{M}\}$ of this set if $M \cap \overline{M} = \varnothing$ and $M \cup \overline{M} = \{1, \dots, n\}$. In the following, we will introduce a score function, which we call *diagonal linear discriminant (DLD) score*, on the set of bipartitions of the samples. This score function measures how clearly the two classes representing a given bipartition are separated by the expression levels of a specific subset of genes. The DLD score is motivated by the classification method of *diagonal linear discriminant analysis* (DLDA, notation following (Dudoit *et al.*, 2000)), which we shall now briefly describe; see also (Mardia *et al.*, 1979). Suppose we would like to classify an additional tissue sample given by its expression profile $\mathbf{y} = (y_1, \dots, y_k)$ with respect to a bipartition $\mathcal{B} = \{M, \overline{M}\}$ of the sample set. DLDA projects $\mathbf{y}$ onto the line generated by the vector

$$\boldsymbol{a} = \boldsymbol{S}^{-1}(\boldsymbol{\mu}_M - \boldsymbol{\mu}_{\overline{M}}), \qquad (1)$$

where $\boldsymbol{\mu}_M$ and $\boldsymbol{\mu}_{\overline{M}}$ denote the average expression profiles of the classes M and $\overline{M}$, and $\boldsymbol{S}$ is the *diagonal* sums–of–squares matrix whose coefficients are the weighted sums

$$s_{gg} = (m-1)\sigma^2_{gM} + (\bar{m}-1)\sigma^2_{g\overline{M}}$$

of within–class variances $\sigma^2_{gM}, \sigma^2_{g\overline{M}}$ for each gene g. Here, $m = |M|$, $\bar{m} = n - m$. The sample $\mathbf{y}$ is then allocated to class M if

$$\boldsymbol{a}(\boldsymbol{y} - \frac{1}{2}(\boldsymbol{\mu}_M + \boldsymbol{\mu}_{\overline{M}})) > 0,$$

and to class $\overline{M}$ otherwise. For high–dimensional data such as microarray data, classification often benefits from selecting a subset of variables that show the strongest correlation with the class distinction of interest. In the context of DLDA, it is natural to measure this correlation by the two–sample t–statistic for each gene g:

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

with

$$c(n,m) = \sqrt{\frac{m\bar{m}(n-2)}{n}}.$$

One may choose e. g. the p variables (genes) with highest absolute value of $t_g(\mathcal{B})$ and discard the other variables from the classification procedure.

In previous studies, DLDA combined with variable selection has proved one of the most successful and robust classification methods for microarray data (see (Dudoit *et al.*, 2000); the method of (Golub *et al.*, 1999) is just slightly different). Dudoit et al. used between 30 and 50 genes for classification and report a range of roughly $p = 10$ to $p = 200$ where the results for the investigated data sets barely changed.

We can now define the *diagonal linear discriminant (DLD) score* $S(\mathcal{B})$ of a bipartition $\mathcal{B} = \{M, \overline{M}\}$ with respect to a parameter p that denotes the number of selected genes: First, all rows but those corresponding to the p genes with highest absolute values of $t_g(\mathcal{B})$ are discarded from the data matrix. Let $\boldsymbol{X}^* = (x^*_{gj})$ be the new data matrix. From $\boldsymbol{X}^*$, we obtain a DLDA discriminant axis $\boldsymbol{a}$ for the bipartition $\mathcal{B}$ according to eqn. (1). The coordinates of the column vectors $\boldsymbol{x}^*_{\cdot j}$ of $\boldsymbol{X}^*$ $(j = 1, \ldots, n)$ projected onto the line generated by the vector $\boldsymbol{a}$ are given by the inner products $\boldsymbol{a} \cdot \boldsymbol{x}^*_{\cdot j}$.

The score $S(\mathcal{B})$ is then defined as the absolute value of the two–sample t–statistic of the values $\boldsymbol{a} \cdot \boldsymbol{x}^*_{\cdot j}$ for the bipartition $\mathcal{B}$:

$$S(\mathcal{B}) = \frac{\mu_{\boldsymbol{a},M} - \mu_{\boldsymbol{a},\overline{M}}}{\sqrt{(m-1)\sigma^2_{\boldsymbol{a},M} + (\bar{m}-1)\sigma^2_{\boldsymbol{a},\overline{M}}}} \cdot c(n,m)$$

Here, $\mu_{\boldsymbol{a},M}$ and $\sigma^2_{\boldsymbol{a},M}$ denote mean and variance of the $\boldsymbol{a} \cdot \boldsymbol{x}^*_{\cdot j}$ with $j \in M$, and $\mu_{\boldsymbol{a},\overline{M}}$ and $\sigma^2_{\boldsymbol{a},\overline{M}}$ correspondingly for $j \in \overline{M}$.

Thus, $S(\mathcal{B})$ measures how clearly the samples of M and $\overline{M}$ are separated after projection onto the one–dimensional subspace which one would use for DLDA classification. Based on the experience with DLDA classification mentioned above, we chose $p = 50$ for all data sets we analyzed. However, we make no assumption on the actual number of differentially expressed genes across the class distinctions we try to detect: A clear difference in the expression levels of only a few genes can be reflected in the DLD score just as strongly as a weaker separation by hundreds of genes.

A fast heuristic for finding bipartitions with high scores

We consider the graph Γ whose vertex set is the set of all bipartitions of $\{1, \ldots, n\}$, with two different vertices $\mathcal{B} = \{M, \overline{M}\}, \mathcal{B}' = \{L, \bar{L}\}$ joined by an edge if and only if there exists $k \in \{1, \ldots, n\}$ with $M \cup \{k\} \in \mathcal{B}'$ or $\overline{M} \cup \{k\} \in \mathcal{B}'$. In other words, two bipartitions are considered as neighbors if they differ only by the class assignment of a single sample. The DLD score function S is defined on the vertex set of Γ. We would like to find bipartitions with high values of S, and focus on local maxima. Since Γ has 2^{n-1} vertices, an exhaustive search is not feasible for practical sizes of n. Our strategy is to use an efficient heuristic to generate *candidate* partitions which then serve as starting points for a greedy search of local maxima.

As input for the candidate generation step, we take a large collection of average expression profiles of clusters of genes, e. g. the $2k - 1$ clusters (including the single genes) produced by a hierarchical clustering algorithm. This yields an augmented data matrix $\boldsymbol{Y} = (y_{ij})$, 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 $M^- = \{j | y_{ij} \leq y_{ij^*}\}$ and $M^+ = \{j | y_{ij} > y_{ij^*}\}$ of samples with expression levels below or above the cut point y_{ij^*}. Whenever both M^- and M^+ have at least two elements, we compute the two–sample t–statistic $t_{ij^*} = t_i(\{M^-, M^+\})$ (see eqn. (2)). We argue that a large value of t_{ij^*} 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 cluster i.

In this step, we use average expression levels of *clusters* of genes, because these may be more stable indicators of different phenotypes than the values of single genes. See also (Hastie *et al.*, 2001), where such cluster average profiles are used as candidate variables for a regression model.

The values t_{ij^*} are compared to the distribution of the two–sample t–statistic for the $m = |M^-|$ smallest and the $n - m$ largest of n independent identically distributed normal random variables, given by its distribution function F_{nm}. We approximate F_{nm} by Monte Carlo simulation. The bipartition defined by the cut point y_{ij^*} is chosen as a candidate if

$$t_{ij^*} \geq F^{-1}_{nm}(1 - \alpha)$$

for a certain value of α (we used $\alpha = 10^{-4}$). Note that we employ this criterion as a rule of thumb for selecting interesting bipartitions and do not claim statistical significance.

From each candidate bipartition $\mathcal{B}$ obtained by this procedure, we proceed in a greedy manner along a path in Γ to a local maximum of the DLD score: Starting at $\mathcal{B}$, we choose in each step the neighboring vertex with the highest DLD score until a local maximum is reached. The resulting high–scoring bipartitions can then be graphically displayed as in Figures 2–4.

RESULTS

Example data sets

Leukemia data set. For a set of 72 acute leukemia mRNA samples, expression levels of 6,817 genes were measured with Affymetrix oligonucleotide arrays (Golub *et al.*, 1999). 25 of the samples were from *acute myeloid leukemia* (AML), whereas the other 47 samples came from *acute lymphoblastic leukemia* (ALL), which further splits 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 data set. This data set is described in (Alizadeh *et al.*, 2000). Expression profiles of 62 lymphoma and leukemia samples were recorded with a specially designed microarray ("Lymphochip") containing 17,856 cDNA clones. We based our analysis on a subset of 4,026 clones selected by the authors for being "well measured" across the samples. The samples represent the following types of lymphoid malignancies: *diffuse large B–cell lymphoma (DLBCL, 42 samples)*, *follicular lymphoma* (FL, 9 samples), and *chronic lymphocytic leukemia* (CLL, 11 samples). The authors detected a division of the DLBCL samples into two subtypes, which they denoted as *germinal center B–like DLBCL* (21 samples) and *activated B–like DLBCL* (21 samples). We refer to the latter two classes as DLBCL–G and DLBCL–A, respectively. Alizadeh et al. found this distinction by hierarchical clustering of the DLBCL samples with respect to a certain cluster of genes that are highly expressed in germinal center B cells.

Melanoma data set. 31 mRNA samples of cutaneous melanoma were investigated by hybridization to a cDNA array representing 6,971 genes (Bittner *et al.*, 2000). We used the data from 3,613 clones selected in the original study for being "strongly detected" across the samples. Using multidimensional scaling and different cluster algorithms, the authors identified a cluster of 19 samples separated from the remaining 12 samples. By subsequent biological experiments, they could show that this class distinction correlates with differences in cell motility and invasiveness.

For all three data sets, we first selected the 2,000 genes with the highest variance of the log–transformed values across the samples. To obtain the augmented data matrix used for the candidate generation step, we clustered these genes by centroid linkage hierarchical clustering with the correlation coefficient as similarity measure. For the computation of the cluster average profiles, gene vectors were standardized to mean zero and variance one. To save computation time, we took only up to 700 candidate bipartitions with the highest DLD score into account and merged similar ones among them by complete linkage hierarchical clustering with respect to the distance in the graph Γ. This resulted in less than 100 bipartitions per data set which then served as starting points for the search of local maxima. In the output of the algorithm, only bipartitions with each subset containing at least 10 % of the samples are listed. Currently, ISIS is implemented in MATLAB. The running time (without the preparatory clustering of genes) per data set on a SUN Sparc II 400 MHz processor was between 1 and 7 minutes.

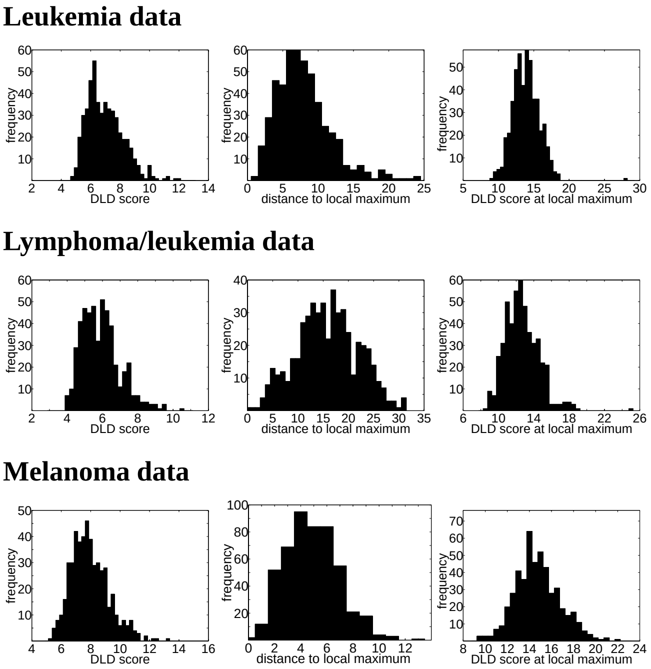

Fig. 1. DLD score, distance to local maximum, and DLD score at local maximum for random splits.

The DLD score of cancer types and its statistical significance

In order to see whether biologically meaningful sample classes can be characterized in terms of the DLD score, we calculated for each known cancer subtype represented in the data sets a) the DLD score of the bipartition of the sample set separating this subtype from its complement, b) the distance in the graph Γ between this bipartition and the local maximum reached by the greedy search starting there, and c) the value of the DLD score at this local maximum (see Table 1).

To explore the landscape imposed by the DLD score on the graph Γ and to assess the statistical significance of the results shown in Table 1, we calculated for each of 500 random bipartitions of the sample sets a) its DLD score, b) the distance from the local maximum reached by the greedy search, and c) the DLD score attained there. Figure 1 shows histograms for these values. One can see that the scores of all cancer subtypes are exceptionally high compared to those of the sampled random splits. On the other hand, proximity to a local maximum is less rare among the random splits. This indicates that a high score is much more statistically significant than a small distance

Table 1. The DLD score characterizes actual phenotypical class distinctions. For each cancer subtype present in the data sets, its DLD score, distance to local maximum, and DLD score at local maximum is shown.

	Class	distance to local maximum	DLD score	DLD score at local maximum
Leukemia data	AML	1	20.5	21.3
	T–cell ALL	1	18.7	19.0
	B–cell ALL	1	22.6	28.1
Lymphoma / leukemia data	CLL	0	18.3	18.3
	FL	2	11.8	14.6
	DLBCL–G	1	13.3	13.3
	DLBCL–A	2	11.9	13.6
	DLBCL	0	24.8	24.8
	DLBCL–G + FL	3	17.0	18.4
Melanoma data	cluster of 19 samples	1	16.8	20.8

to a local maximum alone. We also observed that the data shown in Table 1 for the cancer types with high scores (above a value of 15) are stable with respect to the choice of the parameter p denoting the number of selected genes, roughly in a range between 10 and 150.

Output of the class discovery algorithm

Whereas in the last subsection, we provided empirical support for using the DLD score as an objective function, we will now describe the results obtained by applying our class discovery algorithm to the example data sets. The rows of the matrices in Figures 2–4 show the top scoring bipartitions found by ISIS, ordered by their DLD score which is displayed to the right of each row. Columns are arranged according to cancer subtypes, with no specific order within these types.

Leukemia data – Figure 2. The local maxima corresponding to the three acute leukemia subtypes (see Table 1) are found as rows 1, 4 and 10 in the ranked list of top scoring bipartitions.

For 15 patients, data on treatment were published. We investigated the bipartition of this sample set separating 8 patients with failed treatment from the other 7 patients that were successfully treated. Its DLD score is 11.5, with distance 1 to a local maximum with score 18.9. The latter partition ranks as number 11 on the list of top scoring bipartitions found by ISIS for this smaller data set.

Lymphoma/leukemia data – Figure 3. Partition 1 (4) perfectly separates the DLBCL (CLL) samples from the others. Partition 3 groups the FL samples together with most of the DLBCL–G samples (see also the entries in Table 1 for this class distinction). Indeed, in (Alizadeh *et al.*, 2000) it is mentioned that these two types of lymphoma share high expression of genes characteristic for germinal center B cells. Remarkably, although not all

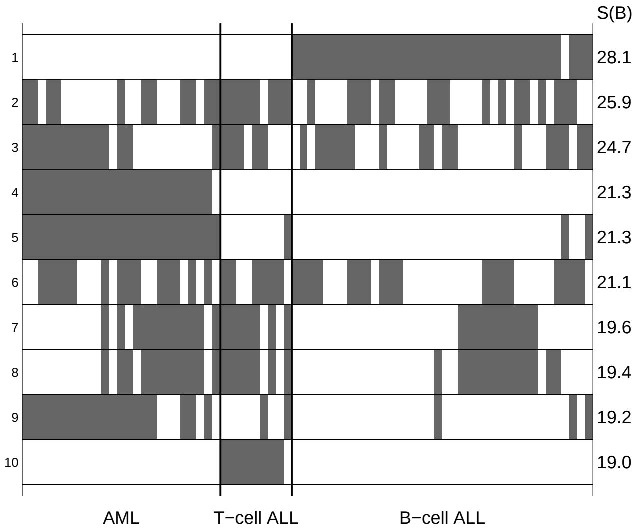

Fig. 2. Partitions of leukemia samples.

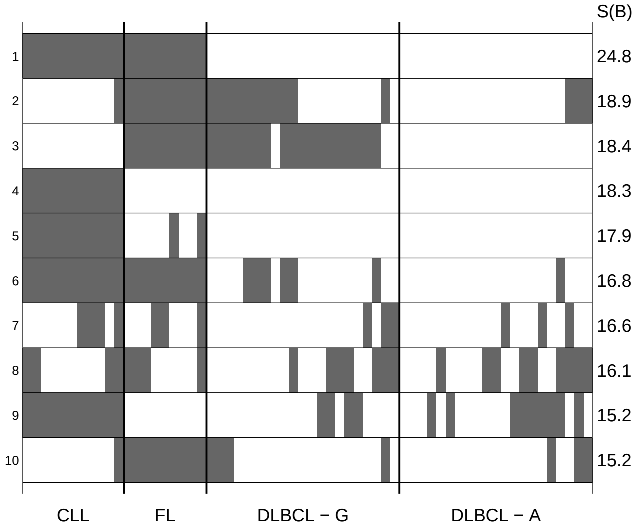

Fig. 3. Partitions of lymphoma/leukemia samples.

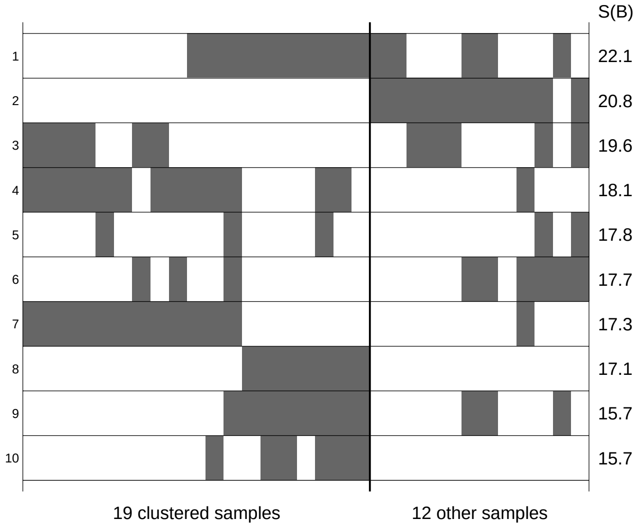

Fig. 4. Partitions of melanoma samples.

of the four cancer subtypes are among the classes with highest scores, the combined information of partition 1 and 3 yields near perfect separation of all four cancer types, with only three samples ’misclassified’. Here we see a structure among the four types which consists of intersecting bipartitions and thus cannot be displayed in a single tree as produced by hierarchical clustering algorithms. Note that the original authors found the distinction of DLBCL-G vs. DLBCL–A samples by focussing on a specific subset of genes and not through a global clustering of the samples.

For 40 DLBCL patients, data on survival times were available. We looked at the restrictions of the 10 top scoring partitions to this subset of samples. For partition 2 (12 vs. 28 samples), survival times differ significantly between the two groups with $p = 0.001$ (unadjusted p–value obtained from a log–rank test).

Melanoma data – Figure 4. Partition 2 in the output of our algorithm coincides with the class distinction that was detected and identified as biologically meaningful in (Bittner *et al.*, 2000), except that our method suggests to reassign one sample, TC–F027, to the cluster of 19 samples. This is consistent with the cluster analysis displayed in (Bittner *et al.*, 2000), where the assignment of this sample looks unclear.

To summarize the above results, we note that most cancer subtypes in the example data sets are not only characterized by a high DLD score and the proximity to a local maximum, but are also among the top scoring bipartitions found by our algorithm. Furthermore, some other class distinctions with similarly high scores and yet unknown biological meaning were found.

DISCUSSION

The high dimensionality of microarray gene expression data creates the need for methods which automatically detect interesting structures in the data. We have introduced a mathematical criterion that characterizes the cancer subtypes represented in several gene expression data sets, and have demonstrated an algorithm that, by employing this criterion, recovers these subtypes without using prior knowledge.

Our method ISIS is guided by the following two observations that apply to microarray data: First, samples may be grouped in different ways, according to different biological factors. Therefore, we are looking for a non–hierarchical clustering of the samples. Second, on the molecular level these different groupings may correspond to expression patterns in different, relatively small subsets of genes. Therefore, we define classes not with respect to a global measure of similarity, but rather with respect to different selections of gene subsets. In this respect, ISIS is related to some other recent approaches to microarray data analysis (Califano *et al.*, 2000; Chen & Church, 2000; Hastie *et al.*, 2000, 2001; Getz *et al.*, 2000), which also investigate relations between subsets of genes and samples. We specifically focus on binary class distinctions of the set of samples, and rate them by the degree of separation that arises from projecting the sample expression profiles onto a discriminant axis determined by a relatively small subset of genes. This mathematical description by the *diagonal linear discriminant* (DLD) score allows to assess the statistical significance of class distinctions.

ISIS is related to ideas underlying projection pursuit methods (Huber, 1985), where one tries to find low–dimensional projections of a cloud of data points that maximize a suitable projection index. Here, the directions in the space of sample expression profiles given by the discriminant axes used for *diagonal linear discriminant analysis* (DLDA) are essentially selected for multimodal distributions of the sample expression vectors projected onto them.

To further analyze the class distinctions obtained by our method for a set of samples, the first and most obvious step is to identify the genes that are differentially expressed across the classes and to evaluate their functional annotations. On the other hand, one may examine the distribution of the sample expression profiles projected onto the DLDA discriminant axis (see eqn. (1)) in order to see whether the class assignment of some samples is unclear. Also the changes in the DLD score due to reassignment of single samples may provide information on the stability of a class distinction.

An important topic in the analysis of microarray data is variable selection. Often, the majority of the genes represented on an array are not related to the investigated phenotypes and contribute only noise to the data. For the computation of the DLD score of bipartitions, we found it both intuitive and useful to first discard genes with low intensity or low variation across the samples before selecting the best discriminating genes for each bipartition by the *t*–statistic. This procedure was consistently applied to all three data sets. As the *t*–statistic is scale–invariant, the different selection criteria are somewhat complementary to each other. Generally, we believe that the influence of variable selection on classification and class discovery methods for microarray data remains to be more systematically studied.

ACKNOWLEDGEMENTS

This work was supported by grants from the German Ministry of Science (BMBF) within the German Human Genome Project (DHGP). We thank Dirk Buschmann for helpful discussions and Johanna Holbrook for proofreading the manuscript.

REFERENCES

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. (2000). Distinct types of diffuse large B–cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. & Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745–6750.

Beißbarth, T., Fellenberg, K., Brors, B., Arribas-Prat, R., Boer, J., Hauser, N., Scheideler, M., Hoheisel, J., Schütz, G., Poustka, A. & Vingron, M. (2000). Processing and quality control of DNA array hybridization data. *Bioinformatics*, **16**, 1014–1022.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology*, **7**, 559–583.

Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, **6**, 281–297.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N. & Trent, J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Califano, A., Stolovitzky, G. & Tu, Y. (2000). Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*. AAAI Press, pp. 75–85.

Chen, Y. & Church, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*. AAAI Press, pp. 93–103.

Chipping forecast (1999). The chipping forecast. Special supplement to Nature Genetics, volume 21.

Dudoit, S., Fridlyand, J. & Speed, T. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576, University of Berkeley, Dep. of Statistics, http://www.stat.Berkeley.EDU/users/terry/zarray/Html/index.html.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998). Cluster analysis and display of genome–wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Getz, G., Levine, E. & Domany, E. (2000). Coupled two–way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, **97**, 12079–12084.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hastie, T., Tibshirani, R., Botstein, D. & Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology*, **2**, research0003.1–12.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D. & Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**, research0003.1–21.

Huber, P. (1985). Projection pursuit. *Annals of Statistics*, **13**, 435–475.

Mardia, K., Kent, J. & Bibby, J. (1979). Multivariate Analysis. Academic Press, San Diego.

Slonim, D., Tamayo, P., Mesirov, J., Golub, T. & Lander, E. (2000). Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)*. ACM Press, pp. 263–272.