mdclust—exploratory microarray analysis by multidimensional clustering

M. Dugas1,*, S. Merk1, S. Breit2 and P. Dirschedl1

1Department of Medical Informatics, Marchioninistr. 15, D-81377 Munich, Germany and
2Department of Dermatology, Frauenlobstraße 9-11, D-80337 Munich, Germany

Received on September 24, 2003; accepted on November 3, 2003
Advance Access publication January 29, 2004

ABSTRACT
Motivation: Unsupervised clustering of microarray data may detect potentially important, but not obvious characteristics of samples, for instance subgroups of diagnoses with distinct gene profiles or systematic errors in experimentation.

Results: Multidimensional clustering (mdclust) is a method, which identifies sets of sample clusters and associated genes. It applies iteratively two-means clustering and score-based gene selection.

For any phenotype variable best matching sets of clusters can be selected. This provides a method to identify gene–phenotype associations, suited even for settings with a large number of phenotype variables. An optional model based discriminant step may reduce further the number of selected genes.

Availability: R-code and supplemental information available from http://martin-dugas.de/mdclust/

Contact: dug@ibe.med.uni-muenchen.de

INTRODUCTION
It is difficult to interpret microarray data from oligonucleotide chips (like Affymetrix™) or cDNA chips, mainly because the number of measured genes is much higher than the number of samples. Therefore, it is difficult to distinguish between random associations and ‘true structure’ in the data.

Unsupervised learning techniques (Hastie et al., 2002) aim to detect the relations within tissues or genes and between them. Usually clusters of similar samples and/or genes have to be identified. ‘Similarity’ is commonly determined by methods like Pearson’s correlation coefficient; the similarity within a cluster is maximized while the similarity between clusters is minimized.

This general approach has limitations, because biomedical data are usually multidimensional. For example, when analyzing gene expression profiles of human samples, gender of the patients will influence the level of certain genes and tissue type will influence the level of other genes.

Therefore, the result of a class discovery method may be stratified by gender, by tissue type or by a combination of both. It is probable that there will be a large number of biologically meaningful possibilities to build clusters of, in some respect, similar samples when measuring many thousands of genes simultaneously. When looking at microarray datasets as a whole, the class discovery process may be dominated by certain sample characteristics (e.g. patient gender, tissue type). Other less distinct features could be overlooked.

The general approach of our method is to construct two clusters of samples in the dataset by \( k \)-means clustering \((k = 2)\), and then to select that gene which is—from its marginal view—most differential for this cluster set. This gene is then removed from the dataset and the process of clustering and identification of differential genes is repeated. With this procedure we try to identify sets of clusters and associated genes, which are visualized afterwards.

After this selection process, the association of cluster sets with dependent phenotype variables like diagnosis can be assessed in a second step. Due to the highly selected nature of the cluster sets and the small number of samples, we apply exact non-parametric tests. This enables the investigator to identify phenotype characteristics associated with gene expression patterns. In an optional third step, the full metric information of the remaining few genes can be used to reduce again the set of genes and to assess prediction accuracy.

METHODS
Our approach assumes that distinct clusters of samples exist, defined by different genes within the data matrix of genes and samples. When performing a simple clustering technique like \( k \)-means (Hartigan and Wong, 1979), the predominant structure of the data will determine the allocation of samples to clusters. By identifying and removing these genes, which are most strongly associated with this structure, clusters according to other features of the data may be identified.

In the sequel we denote the number of samples with \( n \), the number of genes with \( m \) and the number of selected genes
with $s$ in the first step of the algorithm. For practical reasons of visualization we propose to choose $s$ in the magnitude of $n$. A general guideline for $s$ is difficult, because the number of differential genes depends on the data. $X = (x_{ij})$ is the $m \times n$-dimensional matrix of gene expression values.

The sets $C_k = \{i|1 \leq i \leq n; \text{sample} \ i \ \text{belongs to cluster} \ k\}$ (for $k = 1, 2$) contain the results from the two-means clustering. $n_k = |C_k|$ is the number of samples in cluster $k$. To select genes we use the $t$-statistic score with correction for unequal variances:

$$\text{score } t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\sqrt{s^2_{i1}/n_1 + s^2_{i2}/n_2}}$$

where

$$\bar{x}_{ik} = \frac{1}{n_k} \sum_{j \in C_k} x_{ij} \quad \text{and}$$

$$s^2_{ik} = \frac{1}{n_k - 1} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (\text{for } k = 1, 2).$$

Note: The $t$-statistic is just a score to measure heterogeneity between two groups, no $p$-values (empirical error probability of a test statistic) are addressed.

**Step 1: identification of clusters and associated genes**

Repeat (1)–(6) $s$ times.

1. Determine $C_1$ and $C_2$ (two clusters of samples in $X$) using two-means clustering.
2. Calculate score $t_i$ for all genes $i$ ($1 \leq i \leq m$).
3. Select gene $g$ with $|t_g| = \max(|t_i|)$ for $1 \leq i \leq m$.
4. If $t_g > 0$ assign $C_{g, \text{low}} = C_1$, $C_{g, \text{high}} = C_2$; else $C_{g, \text{low}} = C_2$, $C_{g, \text{high}} = C_1$.
5. Store gene $g$, score $t_g$, clusters $C_{g, \text{low}}$ and $C_{g, \text{high}}$.
6. Remove gene $g$ from data matrix $X$.

**Step 2: identification of phenotype–gene associations**

By searching for associations between phenotype variables and the cluster sets from step 1 we can detect phenotype–gene associations indirectly: phenotype variable $V_{\text{pheno}}$ matches a certain cluster set, which is associated to gene $g$; therefore, $V_{\text{pheno}}$ is associated with $g$. Due to the typically small number of samples, exact non-parametric tests are used to assess the degree of matching between the dependent variable $V_{\text{pheno}}$ and the cluster sets.

If $V_{\text{pheno}}$ is a metric variable, an exact Mann–Whitney test for the values of $V_{\text{pheno}}$ grouped by sample clusters $C_{g, \text{low}}$ and $C_{g, \text{high}}$ for each gene $g$ is calculated. If $V_{\text{pheno}}$ is a nominal variable, an exact Fisher test for the values of $V_{\text{pheno}}$ grouped by sample clusters $C_{g, \text{low}}$ and $C_{g, \text{high}}$ for each gene $g$ is calculated. A set of genes with minimal $p$-values is determined (we propose approximately 10 genes). These genes may be associated with $V_{\text{pheno}}$.

**Implementation**

An implementation of mdclust in R Project for Statistical Computing (http://www.r-project.org; Gentleman and Carey, 2002, http://www.bioconductor.org) is available from http://martin-dugas.de/mdclust/. The function mdclust() calculates sets of clusters and associated genes. Function search.mdclust() identifies genes, which are associated with phenotype variables. Plots can be generated with unsorted gene lists as well as sorted by Hamming-distance between cluster sets (starting with the highest-scoring gene), score or gene number.

**RESULTS**

We applied our method to Golub’s leukemia data (Golub et al., 1999), consisting of 27 ALL and 11 AML samples. After preprocessing according to (Dudoit and Ge, 2003, package multtest; http://www.bioconductor.org/repository/release1.1/package/html/multtest.html) 3051 of 7129 genes were analyzed. Figure 1 shows the result after selecting $s = 30$ genes, sorted by Hamming-distance between cluster sets. Two
How can these clusters be interpreted with respect to the sample phenotypes? The function search.mdclust() provides a set of clusters for a given phenotype variable ordered by association with this particular phenotype. Figure 2a presents 10 sets of clusters with best association to the diagnosis (ALL versus AML). Six out of ten genes identified by search.mdclust() belong to the top gene list described in (Golub et al., 1999), for details see Supplementary information.

To verify the qualitative results from our method, we now applied quantitative modeling. By means of a linear logistic discrimination we identified two significant genes (forward selection; \( p < 0.05 \) in the model). Figure 2b presents original data and a discriminant line derived from the linear predictor for these two genes.

Figure 3 illustrates the effect of noise: we generated datasets based on Golub’s leukemia dataset with a noise level from 25 to 100% [noise with a multivariate normal distribution; \( n_{\text{data}} = \text{level} \times \text{noise} + (1 - \text{level}) \times \text{data} \)]. Even at a noise level of 75% the distinction between two classes of samples becomes obvious.

Figure 4 demonstrates how our method can be used to detect systematic errors within the data: we increased the expression levels of each second sample in Golub’s leukemia data to 150%—which could for instance correspond to a different scanner setting and wrong data calibration. Our approach identifies normal and scaled samples as almost distinct clusters.

Figure 5 shows that our method can identify more than two clusters of samples. We analyzed log-transformed data from Karaman et al. (2003), consisting of 46 samples from three species (human, bonobo and gorilla) and 12625 genes each. However, results with simulated data (five groups, 10000 genes, 100 differential genes) indicate that identification of more than two clusters only works for very clear structures in the data (see Supplementary information).

In addition, we tested the method with lymphoma data (Alizadeh et al., 2000) and colon cancer data (Alon et al., 1999) (see Supplementary information).

**DISCUSSION**

Microarray experiments are often applied to get an overview, whether several biological states can be characterized by a different ‘gene expression signature’. Microarray datasets are too large to be inspected manually; therefore, systematic computer-based techniques for exploratory data analysis are necessary to generate hypotheses which are supported by the data.

Unsupervised analysis may detect potentially important, but not apparent characteristics of the data matrix, e.g. subgroups of diagnoses with distinct gene profiles as well as systematic errors like mix-up of samples or different scanner settings for a subset of samples.

Different cluster techniques like hierarchical clustering (Eisen et al., 1998), \( k \)-means (Hartigan and Wong, 1979), self-organizing maps (SOM) (Kohonen, 1990; Tamayo et al., 1999; Hastie et al., 2002) and others (Datta and Datta, 2003)
Fig. 3. Influence of noise. Clustering of Golub’s leukemia data with different noise levels: 25, 50, 75 and 100% noise. Even at 75% noise level the two clusters of samples can be identified.

can be employed to identify expression patterns. Hierarchical clustering is commonly used but not ideal, because due to several discussible parameters (like distance metrics, gene selection techniques) different output can be generated from the same dataset leading to potentially inconsistent interpretations (Fig. 6). Interestingly, the overlap between Figures 6b and 2a is quite small, only three genes (2851, 2124, 1995). This discrepancy is caused by the different allocation of sample no. 12 (Fig. 2a). SOM can be viewed as a constrained version of $\kappa$-means clustering. It needs a set
of heuristic input parameters, which complicates its use in a standardized manner.

In contrast to conventional $k$-means clustering, our method removes in each iteration step one gene from the data matrix and generates a sequence of possibly discriminating gene clusters. To select the most differential gene we apply a score analogous to a $t$-test statistic. We also tested non-parametric scores with similar, but slightly worse results. According to empirical simulations (Fig. 3), our approach appears to be sufficiently robust against noise.

However, there are limitations to the method: it is not guaranteed that the most differential genes are identified first (see Supplementary information, gene scores by iteration step). Single differential genes might be overlooked because of the noise in other genes. Due to numerical instabilities very similar but not identical results can be obtained when repeating the analysis (see Supplementary information), and the method is sensitive to the mean intensity of each sample. Therefore, careful data preprocessing and calibration are important.

The genes selected by our method are highly correlated. An optional additional step which uses quantitative gene expression values and an appropriate discriminant model, reduces the number of extracted genes even more. For a binary phenotype variable, logistic regression with forward selection would be a reasonable choice. Our stepwise approach results in a small set of genes. Available biomedical knowledge should be used to guide the selection process and to interpret results.

Identifying splits with clear separation (von Heydebreck et al., 2001; Markowetz and von Heydebreck, 2002, http://www.molgen.mpg.de/~markowetz/docs/markowetz@gfkl2002.pdf) provides a similar output of distinct sample clusters like our approach, but applies a different method to find these structures based on diagonal linear discriminant scores of bipartitions.

Particularly in the field of medicine it is very common that a phenotype can be characterized by many (in the range of 50–1000) different variables (Dugas et al., 2001). To identify genes associated with phenotype variables, a common approach would be to define clusters of samples according to phenotype characteristics (e.g. diagnosis groups) and then to search for differentially expressed genes. Using the function search.mdclust(), this can be done more efficiently the other way around: by searching for associations between phenotype variables and sets of clusters, potentially relevant gene–phenotype associations can be found.

However, unsupervised analysis is an exploratory technique and its results therefore need validation in prospective, hypothesis-driven experiments. Given the abundance of microarray data and analysis methods, further research is needed to improve microarray analysis, to make the results more transparent, reproducible and comparable.
Fig. 6. Expression pattern of 42 genes from Golub’s leukemia dataset, generated by hierarchical clustering. (a) Genes selected according to largest SD. (b) Top 42 differential genes (AML versus ALL) selected according to t-test statistic. The three genes found in Figure 2a as well are marked by an asterisk.

REFERENCES


