

Traces of molecular disease mechanisms on microarrays

Dennis Kostka, Claudio Lottaz and Rainer Spang
Computational Diagnostics Group, Department of Computational Molecular Biology,
MPI for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

May 4, 2005

Abstract

The main roads of statistical microarray data analysis currently include data normalization, multiple testing, classification models, clustering algorithms and graphical models of transcriptional networks. In this paper we describe two side roads. In the first chapter we describe a semi-supervised algorithm for molecular diagnosis, which is designed for the diagnosis of complex molecular phenotypes, first described in (9). And in the second section we describe an algorithm for the detection of differential co-expression patterns in microarray data (7).

1 Structured Analysis of Microarrays

From a machine learning point of view, classification of gene expression patterns is a very particular task. Typically, training data consists of few samples (small number of experiments) but contains many variables (expression levels measured in each experiment). In this context classical machine learning methods may cause various difficulties (6). For instance, statistical models (particularly those with many parameters) may overfit the training data. Thereby, they rather adapt to noise in the data than learn the desired phenomenon. Moreover, common machine learning methods do not provide an intuitive and biologically meaningful explanation of their results. Rather, single signatures typically determined to characterize specific phenotypes contain a set of biologically unfocused genes. It is very questionable, whether a single global signature, optimized for classification power, actually reflects the underlying biological mechanisms. In the context of clinical diagnosis, we expect phenotypically homogeneous groups of patients to carry differing gene expression patterns, since differing biological mechanisms may lead to the same phenotype. Furthermore, we observe much redundancy in gene expression data, since coregulated genes are highly correlated. Thus, genes from biologically unfocused signatures may be replaced by biologically coherent ones with little loss in the classifier performance.

1.1 Molecular Symptoms

In this research, we consider the task to recognize a particular group of patients presenting a specific clinical phenotype. We call this group the disease group, to be separated from the control group. In *Structured Analyses of Microarrays* we suggest to determine several biological classifiers to detect the disease group. We particularly aim for classifiers with excellent specificity and accept classifiers with suboptimal sensitivity. This is in analogy with symptoms in clinical contexts: symptoms are never present in healthy people, they provide evidence for a certain disease, however, some patients do not display them. Therefore, we call these classifiers *molecular symptoms*. They allow for an additional, molecular stratification of patients according to patterns of absence and presence of symptoms.

In order to determine whether a signature is biologically focused, we need functional annotations for the genes present on the microarray in a systematic way. Structuring biological knowledge and systematic collection of gene function annotation are central goals of the Gene Ontology database (1). Biological terms related to molecular functions, biological processes and cellular components are collected into a directed acyclic graph where each node represents a term and child-terms are either members or representatives of their parent-terms. Moreover, genes are attributed to GO-nodes according to their molecular function, involvement into biological processes and localization within the cell.

In Figure 1 we illustrate the Gene Ontology by depicting two small parts of the directed acyclic graph. On the left hand side the root of the graph is shown with its children. These children represent the

major topics distinguished in the Gene Ontology, namely “molecular function”, “biological process” and “cellular component”. On the right hand side a few nodes from the bottom of the graph are shown and illustrate the fact that nodes may have several parents.

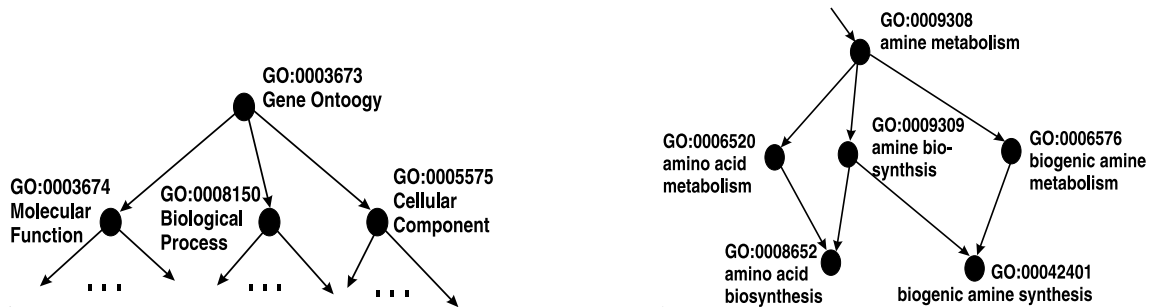


Figure 1: Excerpts of the Gene Ontology, top nodes (left) and some nodes from the bottom in the “biological process”-branch (right).

1.2 Implementation

We suggest to use the Gene Ontology to search for biologically focused classifiers. In order to obtain state-of-the-art performance, we aggregate these focused classifiers representing molecular symptoms as follows.

- For each GO-node with annotated genes, one local classifier is implemented using the nearest shrunken centroids method (11) on expression data of the annotated genes.
- According to their classification performance each local classifier obtains a weight. We define a performance criterion analogous to the probability of misclassification.
- Results of children are collected in their parents by weighted sums using these weights, thereby computing probabilities for each class in each node.
- A shrinkage approach in a cross-validation setting is used to regularize weights such that uninformative branches of the classifier vanish.

This procedure generates a graph structured global classifier according to GO’s hierarchical structure. The overall classification result is provided by the root node’s classifier. We describe our method in detail in (9).

We have implemented structured analysis of microarrays as an R package called `stam`. It is compliant to the Bioconductor suite of bioinformatics related R extensions (4). Our implementation uses the R-package `pamr` which provides training, prediction and cross validation for the nearest shrunken centroids method for classification. The computation is performed in a postorder traversal of the Gene Ontology. In a postorder traversal of a graph, all child nodes of a parent are treated before the parent. Thus we ensure that all data needed for training or prediction in a node are actually available. For the associations of probe-sets with GO terms and for the hierarchical structure of GO we rely on Bioconductor data packages.

1.3 Discussion

In structured analysis of microarrays each classifier bases its decision only on information related to the biological aspect it represents. Therefore, when considering an overall classification result, its rationale can be deduced from the various classifier results. Actually, through the identified molecular symptoms associated to subsets of patients in the disease group, we obtain an additional molecular stratification of patients according to patterns of absence and presence of symptoms.

We have evaluated the method and our implementation on a large dataset from a study on acute lymphoblastic leukemia (14). In this study Affymetrix HG-U95Av2 chips have been used to measure the gene expression profiles in bone marrow of 327 patients. We randomly split this data into training and

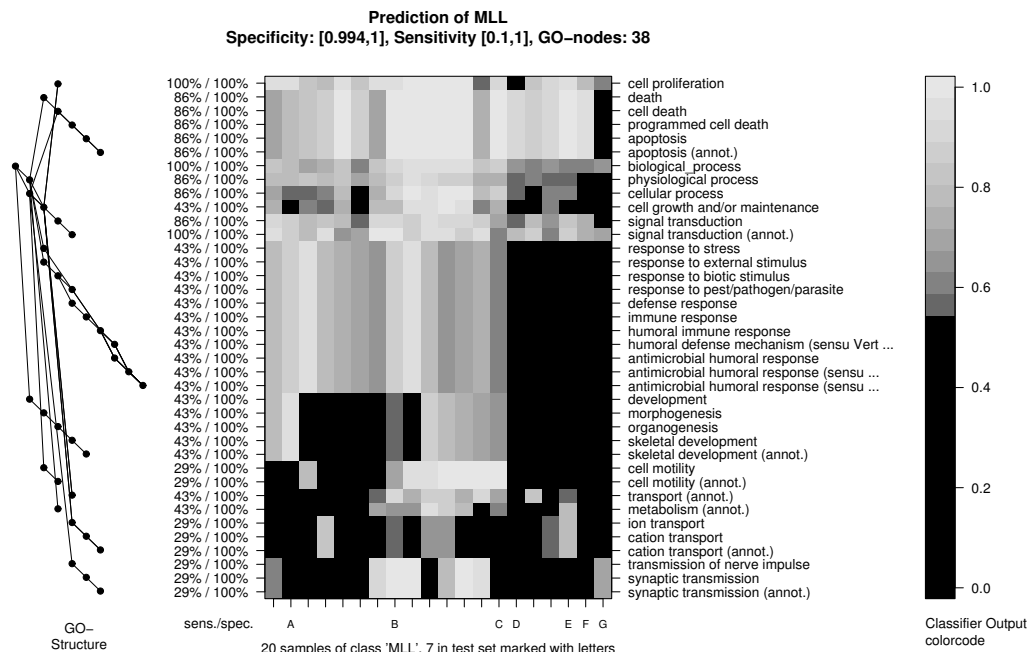


Figure 2: Structured analysis of 327 acute lymphocytic leukemia patients. Molecular symptoms specific for MLL are shown. They are filtered by minimum specificity.

test set. Figure 2 shows an example for molecular symptoms based patient stratification on the MLL sub type of acute lymphocytic leukemia (ALL) investigated in (14). 20 MLL patients have been included in the study. We have trained StAM for detection of MLL on 217 of the available samples including 14 MLL cases. The 110 test samples are classified without error in the root node.

Figure 2 is focused on the 20 MLL samples in the dataset. In the center of the figure the probability computed by classifiers in the classifier graph for each sample are shown as color code (see right hand side of the figure). In the image, rows correspond to GO-classifiers and columns reflect samples. The samples from the test set are marked with capital letters on the x-axis. Clustering this image in both directions brings similar classifiers and samples together. The graph to the left of Figure 2 shows the GO relations between the classifiers. The sensitivities and specificities given between the GO structure and the image are computed on the test set only. In Figure 2, bright regions represent presence, black regions absence of molecular symptoms.

We can group patients according to patterns of molecular symptoms. For instance, rows 2 to 6 in Figure 2 represent a molecular symptom related to *apoptosis*, which is present in all test samples except for sample G. Only in test samples A, B and C we observe the symptom driven by genes involved in *antimicrobial humoral response*. Effects in genes usually involved in *skeletal development* are observed in test samples A and B only, while samples B and C show untypical patterns for ALL in *cell motility*. Samples B and G have particular expression in *synaptic transmission*.

With structured analysis of microarrays, we propose an approach to augment microarray gene expression data through functional annotations provided by the Gene Ontology. We use the additional information to compute class predictions for many biological aspects. On various datasets we have found that our approach can deliver classification results of similar accuracy as state-of-the-art methods currently in use. In addition, structured analysis of microarrays points to biological aspects relevant to the recognition of the investigated phenotype. We introduce the notion of molecular symptoms and illustrate their potential to provide an additional molecular stratification of patients.

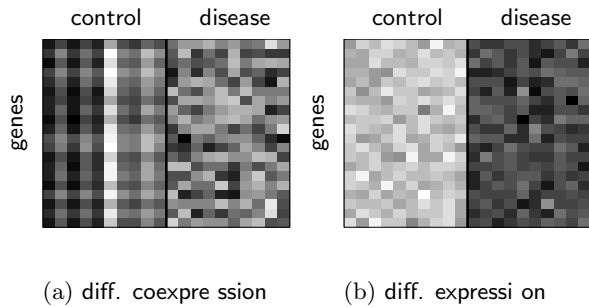


Figure 3: Different types of structure in microarray data. The plots show simulated expression values for two prototypical situations: Plot (a) depicts a group of differentially coexpressed genes, displaying a pattern as a group. Plot (b) shows single differentially expressed genes. Both situations lead to characteristic but distinct patterns.

2 Differential coexpression of genes

Gene expression plays a crucial role in the proper functioning of a cell and is a highly regulated process. Disruption of this regulatory system can lead to degeneration of the cell and cancer. Such changes in the co-regulation of genes are detectable on microarrays (5) and can be utilized for improved diagnosis (13) and prognosis of disease outcome (12).

For detection and interpretation of associated molecular disease mechanisms it is crucial to identify the genes involved. Most analysis strategies either look at up- or down-regulated genes on a gene by gene basis, or focus on genes playing a key role in supervised classification mechanisms. The approach described here is tailored to take into account two things: Firstly, relevant changes need not necessarily be reflected in the data in a univariate (gene-wise) fashion. That is, by looking at up- or down-regulation of individual genes one misses important aspects of the genes' interplay. An illustration can be found in Figure 3. Secondly, highly predictive genes singled out by a possibly sophisticated classification mechanism are not necessarily the biologically relevant ones (8).

The algorithm

Our approach opts to find changes in the co-expression structure of genes. That is, in one group of patients genes are co-expressed and this co-expression is lost in another clinical phenotype. This is interesting, because such a behavior reflects regulatory changes hinting at molecular disease mechanisms. In the following we present a scoring scheme for differential co-expression and present an algorithm that finds groups of *differentially co-expressed* genes.

As a score we utilize:

$$S(I) =: \frac{|J_2| - 1}{|J_1| - 1} \cdot \frac{\sum_{I, J_1} (r_{ij}^{(1)})^2}{\sum_{I, J_2} (r_{ij}^{(2)})^2} \quad (1)$$

where the index set I refers to a group of genes and J_1 as well as J_2 refer to the two groups of patients. The quantities $r_{ij}^{(1,2)}$ are residuals of the expression values of gene i in patient j belonging to group J_1 or J_2 , respectively. The residuals are calculated with respect to an additive model where only genes in the group I contribute. A similar score has been proposed in the context of biclustering by (2).

Now the problem of screening for groups of genes I with optimal scores $S(I)$ remains. This cannot be done exhaustively, since the set of candidate groups is too enormous. Therefore a heuristic is applied. A greedy downhill search with stochastic elements is performed over the huge set of possible gene groups. Random starting points are chosen. This procedure returns a group of genes with a locally optimal differential co-expression score. Also a tuning parameter to influence the size of the finally found groups is implemented.

Results

We applied the algorithm to a data set from a clinical study focusing on childhood leukemia (14). Part of the data are expression profiles of children without cytogenetic abnormalities, in contrast to a group of children which carry the Philadelphia chromosome. Philadelphia positive patients display significantly worse survival. We took the cytogenetically normal samples as one phenotype, the Philadelphia positive ones as another. Application of our algorithm led to the result, summarized in figure 4:

The most prominent pattern of differential coexpression contained 55 genes, five of which were coding for proteasome genes. This corresponds to a p-value of less than $3E-7$ when employing a hypergeometric

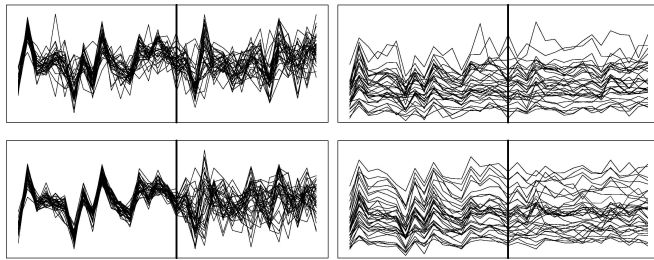


Figure 4: Comparison of two sets of expression profiles. On the left side the profiles are scaled and centered. The upper half presents genes on the chip that code for proteins associated with the proteasome. The lower half represents a pattern of differentially co-expressed genes, only five of which are proteasome associated. While reflecting the mean expression of the proteasome well in the normal samples, this group of genes loses its coherence in the Philadelphia positive samples.

test. Comparing the mean expression value of all the proteasome associated genes with the ones found to be differentially co-expressed, we observed (see figure 4): In the normal tissue samples, where the genes are co-expressed, the mean value of the proteasome genes is reflected well. This changes, when the co-expression is lost in the Philadelphia positive samples. Since the proteasome is known to play a role in oncogenesis (10; 3) this sets the stage for generating hypothesis about the role of the other genes present in the differential co-expression pattern.

References

- [1] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, May 2000.
- [2] Y Cheng and GM Church. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:93–103, 2000.
- [3] R Chiarle, LM Budel, J Skolnik, G Frizzera, M Chilosi, A Corato, G Pizzolo, J Magidson, A Montagnoli, M Pagano, B Maes, C De Wolf-Peeters, and G Inghirami. Increased proteasome degradation of cyclin-dependent kinase inhibitor p27 is associated with a decreased overall survival in mantle cell lymphoma. *Blood*, 95(2):619–26, Jan 2000.
- [4] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.
- [5] TR Golub, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, Oct 1999.
- [6] T Hastie, R Tibshirani, and J Friedman. *The elements of statistical learning*. Springer, New York, 2001.
- [7] Dennis Kostka and Rainer Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20 Suppl 1:I194–I199, Aug 2004.
- [8] B Krishnapuram, L Carin, and A Hartemink. *Kernel methods in computational biology*, chapter Joint feature selection and classifier design. MIT Press, 2004.
- [9] Claudio Lottaz and Rainer Spang. Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics*, 21(9):1971–8, May 2005.

- [10] P Masdehors, S Omura, H Merle-Bral, F Mentz, JM Cosset, J Dumont, H Magdelnat, and J Delic. Increased sensitivity of CLL-derived lymphocytes to apoptotic death activation by the proteasome-specific inhibitor lactacystin. *Br J Haematol*, 105(3):752–7, Jun 1999.
- [11] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, May 2002.
- [12] L J van 't Veer, H Dai, MJ van de Vijver, YD He, AAM Hart, M Mao, HL Peterse, K van der Kooy, MJ Marton, AT Witteveen, GJ Schreiber, RM Kerkhoven, C Roberts, PS Linsley, R Bernards, and SH Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, Jan 2002.
- [13] M West, C Blanchette, H Dressman, E Huang, S Ishida, R Spang, H Zuzan, JA Olson, JR Marks, and JR Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA*, 98(20):11462–7, Sep 2001.
- [14] Eng-Juh Yeoh, Mary E Ross, Sheila A Shurtleff, WK Williams, D Patel, R Mahfouz, FG Behm, SC Raimondi, MV Relling, A Patel, C Cheng, D Campana, D Wilkins, X Zhou, J Li, H Liu, CH Pui, WE Evans, C Naeve, L Wong, and JR Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–43, Mar 2002.