BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 9 2005, pages 1971–1978
doi:10.1093/bioinformatics/bti292

Gene expression

Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data

Claudio Lottaz* and Rainer Spang

Max Planck Institute for Molecular Genetics and Berlin Center for Genome Based Bioinformatics, Ihnestrasse 73, D-14195 Berlin, Germany

Received on September 8, 2004; revised on December 16, 2004; accepted on January 25, 2005
Advance Access publication January 27, 2005

ABSTRACT

Motivation: Today, the characterization of clinical phenotypes by gene-expression patterns is widely used in clinical research. If the investigated phenotype is complex from the molecular point of view, new challanges arise and these have not been adressed systematically. For instance, the same clinical phenotype can be caused by various molecular disorders, such that one observes different characteristic expression patterns in different patients.

Results: In this paper we describe a novel algorithm called Structured Analysis of Microarrays (StAM), which accounts for molecular heterogeneity of complex clinical phenotypes. Our algorithm goes beyond established methodology in several aspects: in addition to the expression data, it exploits functional annotations from the Gene Ontology database to build biologically focussed classifiers. These are used to uncover potential molecular disease subentities and associate them to biological processes without compromising overall prediction accuracy.

Availability: Bioconductor compliant R package

Contact: Claudio.Lottaz@molgen.mpg.de

Supplementary information: Complete analyses are available at http://compdiag.molgen.mpg.de/supplements/lottaz05

1 INTRODUCTION

Supervised tumor classification based on microarray data is among the most promising clinical applications of modern genomics. It opens perspectives for more reliable and efficient diagnosis of established tumor entities (Bhattacharjee *et al*., 2001; Yeoh *et al*., 2002), risk group determination (Huang *et al*., 2003; van't Veer *et al*., 2002), and the prediction of response to treatment (Cheok *et al*., 2003).

Classification in the context of microarray analysis is a well-studied problem in statistics and machine learning. A large number of methods have been suggested, ranging from Fisher's classical linear discrimination to boosting and support vector machines (SVM). Tibshirani *et al*. (2002) investigate a very simple nearest-centroids approach. Classical linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are applied in Dudoit *et al*. (2002). The LDA-like method of Golub *et al*. (1999) is described in more detail in Slonim *et al*. (2000). Artificial neural networks are used in Khan *et al*. (1999). SVMs are suggested in Ben-Dor *et al*. (2000), Furey *et al*. (2000) and Yeoh *et al*. (2002). Nearest-neighbor methods are discussed in Dudoit *et al*. (2002) and classification trees, including boosting, are applied in Ben-Dor *et al*. (2000), Dudoit *et al*. (2002), Schoch *et al*. (2002) and Dettling and Bühlmann (2003). The use of Bayesian binary regression is laid out in West *et al*. (2001) and Spang *et al*. (2002).

In addition to predictive performance, there is also hope that microarray studies uncover molecular disease mechanisms. However, in many cases the molecular signatures discovered by the algorithms are unfocused from a biological point of view. They contain genes attributed to many different biological processes and do not point to particular underlying molecular disease mechanisms. In fact, they often look more like random gene lists than biologically plausible and understandable signatures. This is because standard classification algorithms aim for global signatures. They identify groups of genes whose joint distribution of expression levels is most different between two different clinical phenotypes without considering their biological role.

Moreover, the fact that only one list of genes is determined for all patients reflects the implicit assumption that a single molecular mechanism is responsible for a certain clinical phenotype. This assumption is questionable, e.g. when distinguishing between recurrent and non-recurrent disease, it is quite possible that recurrence has various molecular backgrounds. If this is the case, one will expect different molecular changes in different patients. In order to formalize this idea, we treat the two phenotypical groups (disease and control) in a non-symmetric way. Instead of global expression signatures, we want to identify sets of genes that display characteristic expression patterns in a subset of patients from the disease group but not all of them. We aim for different sets of genes, which possibly identify different subsets of patients in the disease group. We call these patient subset-specific expression patterns molecular symptoms to distinguish them from global molecular signatures.

Another shortcoming of standard classification algorithms is that they treat gene-expression levels as anonymous variables. However, a lot is known about the function and the role of many genes in certain biological processes. This knowledge is stored in databases like the Gene Ontology (GO) (The Gene Ontology Consortium, 2000). Annotations are routinely used today when biologists analyze lists of differentially expressed genes. In contrast to this post analysis exploration of functional annotations, we propose using annotations during the statistical analysis process, i.e. when computing molecular signatures and molecular symptoms. Since gene-expression levels

*To whom correspondence should be addressed.

result from tightly coordinated regulatory processes, they tend to be highly correlated. Consequently there is redundant information in microarray data. We have often observed, that one can replace one gene in a signature by another one with a similar expression profile without significantly changing the predictive power of the signature (data not shown). This opens up the possibility of replacing a non-intuitive, biologically diverse signature by an equally good one with a clear functional focus.

We propose a biologically resolved computational diagnosis based on the GO. In GO, biological terms describing biological processes are organized in a directed acyclic graph, where each node represents a biological process and child-terms are either members or representatives of their parent-terms. Genes are attributed to nodes based on the knowledge the biological research community has gathered so far. Our basic idea is as follows: we construct one classifier for each node of the GO graph. Each of these classifiers only depends on expression levels matching the biological aspect the node represents. Similar to global gene selection based classifiers, we shrink the GO graph, getting rid of branches (biological processes) that are most likely unrelated to the investigated phenotype. The remaining nodes represent molecular symptoms. Different biological processes may identify different subsets of patients.

In the next section we describe the StAM algorithm in detail. Section 3 contains an evaluation on three publicly available cancer related datasets. Therefore we illustrate StAM's performance both as a predictive classification method and as an exploratory tool for the molecular stratification of the disease group and for establishing links between complex phenotypes and biological processes. Finally, conclusions are drawn and discussed in Section 4.

2 STRUCTURED MICROARRAY ANALYSIS

In order to provide biologically resolved diagnosis on various levels of granularity, we use GO's hierarchical structure. Based on the GO graph of biological processes, StAM generates a classifier graph holding one classifier for each process. These classifiers only depend on genes annotated to corresponding nodes or their descendants. For instance, the classifier of the node 'apoptosis' only depends on genes involved in apoptosis. Its diagnosis for a patient only reflects altered gene regulation in apoptosis-related pathways. Our approach consists of the following steps:

- generate a rooted, directed classifier graph according to the GO,
- construct leaf node classifiers based on selected expression values using a classical machine learning method,
- propagate these results through inner nodes to the root and
- shrink the classifier graph to determine a concise set of molecular symptoms.

2.1 Classifier structure

StAM's classifiers exclusively predict based either on their children's classification results or on the expression levels of the directly attributed genes. In GO, genes can be annotated to both leaf nodes and inner nodes. Therefore, we augment the GO graph such that genes are annotated to leaf nodes only. If i is an inner node with genes annotated to it, we introduce a novel leaf node i' to the graph, with i as its only parent and move all genes from i to i'.

We generate the graph described above anew for each chip type. Thus, we can choose any GO node as the term of interest and start our procedure with the given node as the root of the graph considering only successors in GO during graph construction. We use two methods to remove non-informative nodes from the result: (1) Nodes are discarded if neither they nor any of their successors have genes or probes annotated to them. (2) A node with a single child is replaced by its child, since results are identical (they depend on the same data). In this manner we generate a classifier graph, specific to the chip type used in the given study and the GO term of interest.

2.2 Leaf node classifiers

Each leaf node contains a set of associated genes. The corresponding classifier is constructed using only expression levels of these genes. It returns a continuous classification output scaled to numbers between 0 and 1, where 0 indicates clear evidence for the control group, one indicates clear evidence for the disease group and intermediate values represent the levels of uncertainty. The number of genes annotated to a leaf node varies strongly from one node to the other. For some nodes it is so high that classification in the leaf nodes still requires regularization to avoid overfitting. In principle, any machine learning method can be used here.

In our current implementation we have chosen the shrunken centroid classification (Tibshirani *et al*., 2002) as the leaf node class prediction method for its simplicity and computational efficiency. Centroid shrinkage is determined by cross-validation node by node, such that the lowest error rate is achieved. In order to regularize the classifier centroid shrinkage excludes genes from the signatures. Thus a classifier associated to the node 'apoptosis' is driven only by genes involved in apoptosis, but not necessarily all of them. The shrunken centroids method defines a continuous classification output by logit transformed discriminant scores. A continuous classification scale smoothes the process of classification propagation to inner nodes.

2.3 Propagating classification results

So far we have classifiers for the leaf nodes, next we combine them with the classifiers in the inner nodes. We do this without breaking the leaf node classifiers apart. There are no novel classifiers built using merged gene sets, since this would lead to the non-intuitive signatures that we want to avoid. Instead we suggest weighted sums of child classification outputs to propagate the results. Thus, the root node naturally displays the overall classification result because it depends on the largest amount of data. Children with good classification performance receive more weight than those with poor performance. StAM chooses weights according to a performance criterion which reflects the properties of molecular symptoms, thus punishing low specificity more severely than lack of sensitivity.

Our performance criterion δ_i for node i is analogus to the deviance used in statistical classification theory. We define a similarity measure d_i, using a calibration parameter β to enforce high specificity for the price of reduced sensitivity. Let S_c and S_d represent the samples of the control and the disease group, respectively, while p_i^s denotes the classifier output of node i and sample s. We define:

$$d_i = \frac{-2(1-\beta)}{|S_d|} \sum_{s \in S_d} \log(p_i^s) + \frac{-2\beta}{|S_c|} \sum_{s \in S_c} \log(1 - p_i^s)$$

for all nodes i. Given that d_is are high for bad classifiers, we flip the scale by subtracting them from the highest d_i observed in leaf nodes.

Finally, in order to eliminate uninformative classifiers, we subtract a shrinkage level Δ and set negative δ_is to zero. With N_L denoting the set of leaf nodes, set:

$$\delta_i = [\max_{j \in N_L} d_j - d_i - \Delta]^+$$

where $[x]^+$ is zero for negative x and x otherwise.

The prediction results are propagated from the leaf nodes towards the root through the edges E in a postorder traversal of inner nodes. Hence, StAM always computes results for all children of a given node before it computes results for the node itself. Each edge from parent i to child j receives a weight ω_{ij}. The weights reflect the quality measure δ_j of child j and are normalized separately in each inner node. With $\mathrm{Ch}(i)$ denoting the set of children of node i and N_I denoting the set of inner nodes, we can write the propagation of results as follows:

$$\omega_{ij} = \frac{\delta_j}{\sum_{k \in \mathrm{Ch}(i)} \delta_k} \qquad \forall (i,j) \in E$$

$$p_i^s = \sum_{j \in \mathrm{Ch}(i)} \omega_{ij} \cdot p_j^s \quad \forall i \in N_I \wedge s \in S_d \cup S_c.$$

2.4 Classifier graph shrinkage

Most nodes in the classifier graph do not contribute to a good overall classifier. Many biological processes are not involved with the investigated phenotype. In addition, we want to determine a concise set of molecular symptoms. We describe in this section how StAM further simplifies the classifier graph by eliminating irrelevant branches. This is done in analogy to gene selection in the shrunken centroid algorithm. Here we do not shrink weights associated to genes, but the weights associated to edges in the GO graph. If such a weight is shrunken to zero the corresponding edge and the subgraph below it is eliminated from the graph. StAM controls the shrinkage process by choosing the above mentioned graph shrinkage level Δ. We define an objective function for Δ considering two independent goals: good predictive performance in the root and uncovering suboptimally classifying molecular symptoms for patient stratification; and for the second goal, aggressive shrinkage is counterproductive since by focusing on best classifiers it only eliminates too many inherently heterogeneous molecular symptoms.

We propose an objective function composed of the following two measures: the root's performance measure (δ_{root}) and the mean classifier redundancy. While δ_{root} is already defined, we now focus on what we call redundancy. When considering two nodes in the trained classifier graph, we can define a similarity r_{ij} between the two classifiers expressing how different their results are as follows:

$$r_{ij} = \frac{-1}{|S|} \sum_{s \in S} \log\left(p_i^s(1-p_j^s) + (1-p_i^s)p_j^s\right),$$

where $S = S_d \cup S_c$. The mean similarity to all other nodes in the classifier graph is the node's redundancy r_i within that graph. We suggest to use the mean redundancy over all nodes in a shrunken classifier graph as measure $R(\Delta)$ for the heterogeneity of its classifiers. Let $K(\Delta)$ denote the set of nodes remaining in a shrunken classifier graph. Thus:

$$r_i = \frac{1}{|K(\Delta)| - 1} \sum_{j \in K(\Delta) \setminus \{i\}} r_{ij}$$

$$R(\Delta) = \frac{1}{|K(\Delta)|} \sum_{i \in K(\Delta)} r_i$$

Finally, StAM uses a calibration parameter α in the interval [0, 1] to compute a combined score for each shrinkage level. Thereby, α is the weight for the root's performance measure while $(1-\alpha)$ is the weight for the classifier graph's mean redundancy. We scale the root's performance measure and mean redundancy to fit in the interval [0, 1] before computing the compound score $O(\Delta)$:

$$O(\Delta) = \alpha \frac{\delta_{\mathrm{root}}(\Delta) - \delta_{\min}}{\delta_{\max} - \delta_{\min}} + (1-\alpha)\frac{R(\Delta) - R_{\min}}{R_{\max} - R_{\min}}$$

where $R_{\min}$ and $R_{\max}$ are the minimum and maximum mean redundancy over all Δs while $\delta_{\min}$ and $\delta_{\max}$ give the range of root performances over all Δs. StAM chooses the graph shrinkage level Δ to minimize $O(\Delta)$. When several candidates are equivalent, the lowest shrinkage is used in order to provide a more resolved classification result.

2.5 Calibration of parameters

In our method, the user specifies two calibration parameters: the specificity versus sensitivity parameter β and the performance versus redundancy parameter α. Both parameters can be chosen freely within the interval [0, 1]. However, users should bear in mind the following considerations when doing so. The root performance weight α expresses the desired trade-off between prediction accuracy and heterogeneity of molecular symptoms. Setting α to 1 focuses on classification performance only, while setting α to 0 aims to determine a most heterogeneous classifier graph.

Although the parameter β is meant to overstate specificity intentionally, there is a trade-off between classification performance and discovery of molecular symptoms. The specificity weight can be chosen more freely in easy classification tasks, while heavily unbalanced analysis is mostly meaningless in difficult classification tasks. We usually start out with β set to the prevalence of the control group, thus expecting the best prediction results. When the prediction task proves to be simple enough, we attempt a more unbalanced analysis. Our current implementation can compute classifier graphs for several βs in one run. Therefore, we usually compute several variants right from the start, e.g. by setting β to the values 0.75, 0.9, 0.95 and 0.99 in addition to the control group's prevalence.

2.6 Implementation

StAM is implemented in R (R Development Core Team, 2004) based on Bioconductor packages (Gentleman *et al.*, 2004). We rely on the pamr package (Tibshirani *et al.*, 2002) as implementation of the shrunken centroids classifier. We also use Bioconductor's meta-data packages on chip annotations and the GO. For the layout and illustration of the classifier graph on StAM's result pages, the Graphviz software package is used (Gansner and North, 2000). StAM is itself a part of Bioconductor release 1.5.

The results are written on interlinked HTML pages. The links allow navigation along the edges of the classifier graph. The pages contain classification results and performance evaluation for each

node as well as overall information about cross-validation, the model fit and predictions of test samples. For inner nodes, the weights of the children are provided while in leaf nodes the genes used by the shrunken centroid classifiers are given. The user can further explore term definitions and probeset annotations through external links to the GO and the Affymetrix website.

3 EVALUATION ON CANCER RELATED DATA

We suggest structured analysis of microarrays for different applications. In addition to predictive performance we also aim for making underlying disease mechanisms transparent. We do this by identifying molecular symptoms associated to subsets of patients in the disease group. Molecular symptoms are always restricted to well defined biological processes. Patients who are positive for a molecular symptom display abnormal gene expression in the corresponding process. Not all patients in the disease group are positive for every identified molecular symptom, but some patients can be positive for more than one of them. Using patterns of absence and presence of molecular symptoms, we define an additional molecular stratification of patients.

We have evaluated our approach on three publicly available datasets from cancer-related microarray studies. Here we only discuss a subset of the obtained results. The reader can find a complete collection of our analyses and a detailed description of our data preprocessing protocol on a supplementary website.

3.1 Datasets and GO annotations

The first dataset we used was generated in a breast cancer study (Huang *et al*., 2003). The authors investigate lymph node metastatic status and relapse in 37 and 52 breast cancer patients respectively. The second dataset stems from a lung cancer study (Bhattacharjee *et al*., 2001). Gene-expression profiles from 186 lung cancer and 17 normal biopsies have been analyzed by hierarchical and probabilistic clustering. The authors claim to have discovered distinct groups of adenocarcinomas with corresponding marker genes which are retrospectively correlated to long term outcome. Finally, we also use the dataset on pediatric acute lymphocytic leukemia (ALL) published in Yeoh *et al*. (2002). This study contains gene-expression profiles of 327 patients of various ALL subtypes. Yeoh *et al*. report on an attempt to ease stratification of ALL patients according to relapse risk in order to tailor treatment intensity.

All mentioned studies have been performed using the HG-U95Av2 Affymetrix GeneChip technology. This microarray holds 12625 probesets designed based on the EST clusters from UniGene (Schuler, 1997) version 95. For mapping these probesets to GO nodes we have used the Bioconductor meta-data packages version 1.5.1 built on March 3 and 4 2004. While generating our annotations we focussed on GO's biological process ontology. We have determined 8172 successors of GO:0008150 in the biological process branch of the GO. Of these 1359 have 8679 probesets directly annotated and are held together by 845 inner nodes in our classifier graph. Thus our method has access to 68.7% of the microarray data distributed across 2204 GO terms to achieve the results described below.

3.2 Prediction accuracy

In this section we confirm that StAM's classification performance is comparable to the state-of-the-art classification methods. However, our performance is compromised by the fact that certain probesets are not associated to any GO node and therefore not used in StAM.

Table 1. Performance comparison of StAM to PAM and SVMs

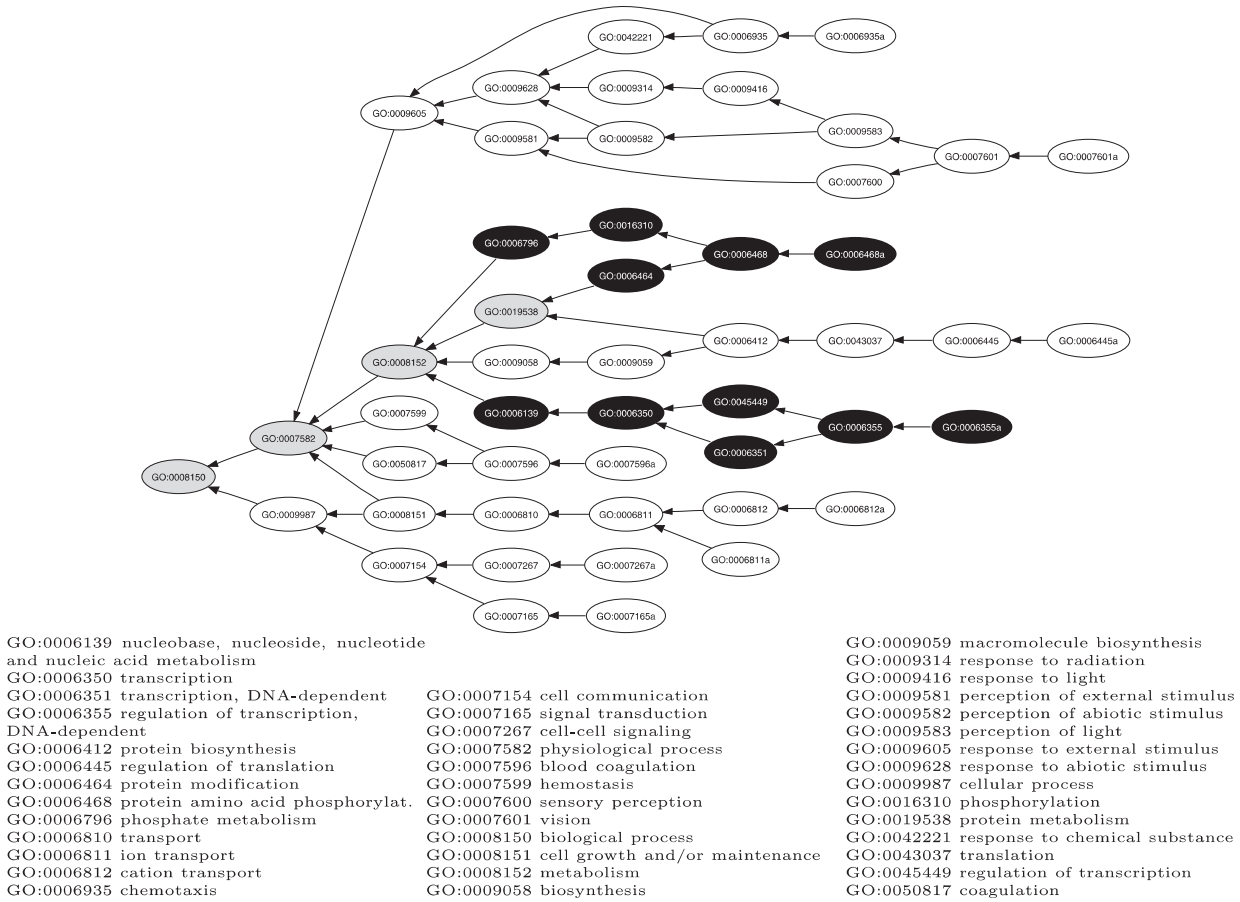
Classification Task		Error Rates		
Study	Groups	SVM	PAM	StAM
St Jude	Hyperdip.: 64	7.6%	7.3%	5.8%
ALL1 Study	Other: 263	(100/53)	(8617/707)	(732/7)
	BCR/ABL: 15	1.2%	4.9%	3.1%
	Other: 312	(100/70)	(15/16)	(383/6)
	E2A/PBX1: 27	0.0%	0.6%	0.6%
	Other: 300	(100/64)	(2/6)	(419/4)
	MLL: 20	0.3%	4.6%	2.1%
	Other: 307	(100/59)	(81/51)	(599/6)
	TEL/AML1: 79	1.2%	2.4%	1.8%
	Other: 2248	(100/61)	(35/38)	(1275/12)
	T-ALL: 43	0.0%	0.0%	0.3%
	B-ALL: 284	(100/48)	(3/6)	(913/21)
Harvard Lung	Adeno: 139	9.9%	9.4%	8.9%
Cancer Study	Other: 64	(100/64)	(3575/498)	(1992/23)
	Adeno: 139	3.8%	1.9%	1.9%
	Normal: 17	(100/58)	(3/4)	(1209/23)
	Carcinoid: 20	0.5%	0.0%	0.5%
	Others: 183	(100/51)	(1/1)	(1730/25)
	Normal: 17	2.0%	1.0%	2.0%
	Other: 186	(100/64)	(3/4)	(504/39)
	Squamous: 21	3.9%	3.0%	4.9%
	Other: 182	(100/48)	(2/3)	(834/18)
	Squamous: 21	0.0%	2.6%	0.0%
	Normal: 17	(100/64)	(4/5)	(583/25)
Duke Breast	Relapse: 18	25.0%	23.1%	21.2%
Cancer Study	Remission: 34	(100/55)	(71/41)	(363/8)
	High risk: 18	32.4%	45.9%	45.9%
	Low risk: 19	(100/78)	(67/53)	(312/2)

Error rates determined in 10-fold cross-validation on public datasets from cancer related microarray studies.

Although we do not claim that our approach outperforms state-of-the-art classifiers, we evaluate its classification power compared to ordinary PAM (shrunken centroids) and SVMs in order to validate its usefulness in clinical diagnosis.

The evaluation is performed in a nested cross-validation scheme where the same cross-validation subsets are used for all three methods. Nested cross-validation means that there is an outer cross-validation loop for model evaluation and an inner loop for model selection. The samples are divided into $k = 10$ sets and one by one each of these sets is left out for evaluation. With the remaining 9/10 of the data we optimize the StAM classifiers, again using cross-validation of only these 9/10 samples. The samples left out in the outer loop are not used at all in any of StAM's model selection steps, including shrinkage determination in the leaf nodes, propagation weight calculation and graph shrinkage. This ensures that the evaluation results do not suffer from the well known overoptimism described in Ambroise and McLachlan (2002).

For StAM we have performed classification using the method described in Section 2 with annotations from the biological process branch of GO as described in Section 3.1. For PAM we have used the default parameters of the corresponding R package (version 1.12.1). The evaluation for the SVM has been performed using the implementation from the e1071 R package (version 1.3-16,



GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism
GO:0006350 transcription
GO:0006351 transcription, DNA-dependent
GO:0006355 regulation of transcription, DNA-dependent
GO:0006412 protein biosynthesis
GO:0006445 regulation of translation
GO:0006464 protein modification
GO:0006468 protein amino acid phosphorylat.
GO:0006796 phosphate metabolism
GO:0006810 transport
GO:0006811 ion transport
GO:0006812 cation transport
GO:0006935 chemotaxis
GO:0007154 cell communication
GO:0007165 signal transduction
GO:0007267 cell-cell signaling
GO:0007582 physiological process
GO:0007596 blood coagulation
GO:0007599 hemostasis
GO:0007600 sensory perception
GO:0007601 vision
GO:0008150 biological process
GO:0008151 cell growth and/or maintenance
GO:0008152 metabolism
GO:0009058 biosynthesis
GO:0009059 macromolecule biosynthesis
GO:0009314 response to radiation
GO:0009416 response to light
GO:0009581 perception of external stimulus
GO:0009582 perception of abiotic stimulus
GO:0009583 perception of light
GO:0009605 response to external stimulus
GO:0009628 response to abiotic stimulus
GO:0009987 cellular process
GO:0016310 phosphorylation
GO:0019538 protein metabolism
GO:0042221 response to chemical substance
GO:0043037 translation
GO:0045449 regulation of transcription
GO:0050817 coagulation

Fig. 1. Union of classifier graphs for relapse prediction (white nodes) and detection of metastatic lymph nodes (black nodes) in breast cancer patients. Gray nodes occur in both classifier graphs.

a libsvm interface available at http://www.cise.ntu.edu.tw/cjlin/libsvm (Chang and Lin, 2004)) selecting 100 features based on the SAM score (Tusher *et al*., 2001). Table 1 summarizes the results for a series of classification tasks on the three datasets mentioned above.

From Table 1 we see that StAM outperforms both reference methods in three of the tasks, while it is worse than these in only two cases. StAM yields either an equal result as one of the other methods or its performance is between the reference methods for nine more classification problems. This result confirms that StAM delivers competitive classification results compared with the state-of-the-art classification methodology.

The numbers in parentheses in Table 1 give the number of features selected for PAM and SVM as well as the number of GO categories to which these features are annotated (biological process only). The figures confirm our observation that many signatures are composed of genes from many biological processes with little biological focus. The corresponding figures given for StAM give the number of leaf nodes in the structured classifier and the number of features selected. StAM often uses many genes, but explicitly attributes them to few molecular symptoms. For the vast majority of cases even the global StAM signatures are more focused than in non-structured methods. However, even when the global signature is not focused, the molecular symptoms are by definition.

3.3 Uncovering disease mechanisms

In addition to predictive performance, we also aim to make underlying disease mechanisms transparent. We do this by identifying molecular symptoms involved in the investigated disease. In this section we illustrate the explorative detection of disease mechanisms; classification performance is of secondary interest.

Huang *et al*. (2003) claim that although the lymph node metastatic status in breast cancer is a commonly accepted risk indicator for relapse, different biological mechanisms appear to be involved. To provide evidence for this claim, two classifiers are trained on separate datasets: one to predict disease outcome in terms of recurrence and the other one to characterize metastatic lymph node status. Huang *et al*. observe that the signatures of the two classifiers overlap in very few genes and thus conjecture that different biological mechanisms are involved in metastasis development and breast cancer relapse. We confirm and further characterize these findings using structured analysis.

When using StAM exploratively, we do not split the dataset into test and training set. For the two classification tasks relapse prediction and lymph node metastasis detection, we train classifier graphs using all available samples. The models generated for the two tasks are shown in Figure 1. Only gray nodes occur in both classifiers and may, therefore, point to common biological mechanisms.

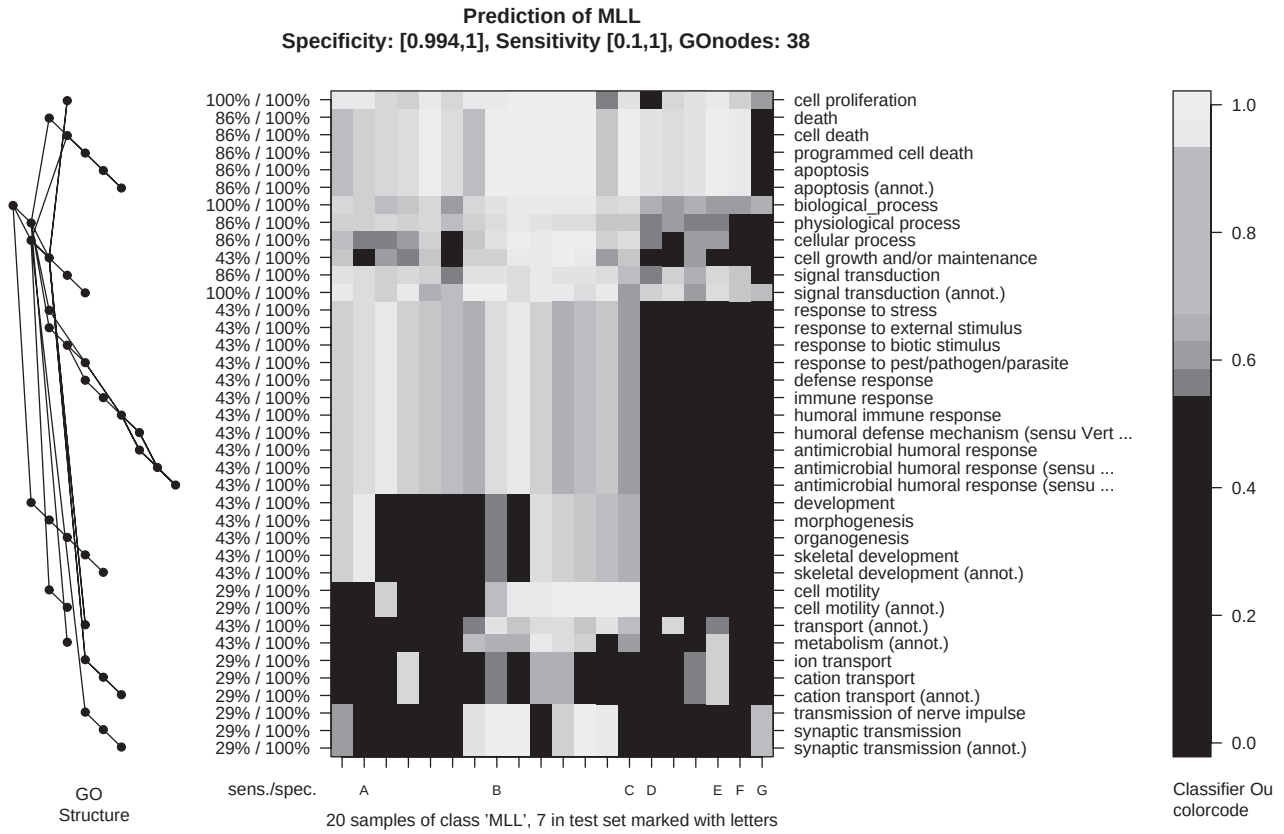


Fig. 2. Structured analysis of 327 acute lymphocytic leukemia patients. Molecular symptoms specific for MLL are shown. They are filtered by minimum specificity.

The structured classifier to detect metastatic lymph nodes is based on 15 GO nodes while the classifier graph for the relapse predictor holds 39 nodes ($\alpha = 0.7, \beta = 0.9$). They share only four nodes, most of which are high level nodes and no leaf nodes are shared. The risk-assessment classifier relies on data annotated to the terms 'protein amino acid phosphorylation' and 'DNA dependent regulation of transcription'. On the other hand prediction of relapse is based on data from 'blood coagulation', 'cation transport', 'regulation of translation', 'vision' and 'chemotaxis'. This illustrates an advantage of structured analysis of microarrays. While classical methods for detecting global signatures return an unstructured list of genes, StAM directly determines biological aspects involved in addition to the corresponding genes. For each of the leaf nodes remaining in the shrunken classifier graph, StAM provides a list of genes relevant for classification and potentially useful for further investigation.

3.4 Patient stratification

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 such symptoms. To illustrate this use of StAM we randomly split our data into training and test set. Figure 2 shows an example for StAM-based patient stratification on the MLL subtype of acute lymphocytic leukemia (ALL) investigated in Yeoh *et al*. (2002). A group of 20 MLL patients has been included in the study. We have trained StAM for detection of MLL on 217 of the available samples including 14 MLL cases with β set to 0.94, the control prevalence. We set the shrinkage level manually to obtain a reasonable number of nodes in the classifier graph. 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 2–6 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.

Table 2. Stability of classifier graphs

Classification Task		Nodes in CV-graphs			
Study	Task	CV_{all}	$M_{\geq 1}$	$M_{\geq 7}$	$M_{=10}$
Leukemia	Hyperdip.	27	24	22	18
	BCR/ABL	44	24	19	15
	E2A/PBX1	28	18	10	10
	MLL	46	33	17	17
	TEL/AML1	74	45	32	31
	T-ALL	89	78	75	75
Lung cancer	Adeno:	126	80	78	70
	Adeno/normal	105	83	77	76
	Carcinoid	132	89	81	77
	Normal	220	136	119	110
	Squamous	94	72	66	37
	squamous/normal	124	85	74	74
Breast Cancer	Relapse	123	40	30	11
	High risk:	28	15	9	9

For tasks described in Table 1, count nodes occuring in any cross-validation run (CV_{all}). Further columns indicate how many nodes of the final graph occur in any ($M_{\geq 1}$), at least in 7 ($M_{\geq 7}$) and in all ($M_{=10}$) cross-validation runs.

We have checked the stability of the classifier graph discussed here by comparing the results of its 10 cross-validation runs. Of these 127 nodes are in the classifier graph for at least one cross-validation run and 78 of these are present in all classifier graphs. From the overall model shown in Figure 2, 84 nodes occur in at least seven cross-validation models. In Table 2 we show results of similar analyses on all classification tasks discussed in Section 3.2. From these results, we conjecture that the molecular symptoms identified in StAM are fairly stable. They allow to resolve a patient's diagnosis according to their presence or absence and thus characterize patient subgroups which may be of clinical relevance.

4 DISCUSSION

In this paper, we present an approach to integrate biological annotation into statistical class prediction analysis of microarray data in an apriori fashion. We use the functional annotation collected in the GO database to construct structured classifiers. Class predictions are computed for each term in the GO which is related to the disease. Our method allows for biologically resolved diagnosis of patients. It is thus able to diagnose complex clinical phenotypes, where different patients who show the phenotype may display different molecular characteristics. Our method can be generalized easily beyond the common two class problem, although the interpretation of molecular symptoms may be difficult in the multiple class context.

A simpler approach to use GO annotations a priori in class prediction would be to collect genes for each term including genes of successor terms and generate a classifier for each of these just as we do for leaf nodes. We did not further develop this idea for three reasons: First, this approach generates increasingly unfocused signatures for high level terms. Second, our approach is computationally more efficient, since no training for inner nodes is needed. And finally, we show that the weighting of terms has the potential to improve classification accuracy. In seven cases considered in Section 3.2, StAM outperforms the ordinary shrunken centroids approach (PAM), while in only four cases PAM achieves lower error rates than StAM.

We evaluate our method using three cancer related publicly available datasets. Thereby, we show that StAM achieves competitive prediction performance compared to state-of-the-art classification methods like SVMs. In addition StAM provides molecular disease group stratification according to biologically focused gene expression patterns, molecular symptoms, by exploiting functional annotations during the statistical analysis. This is in contrast to most of the previous approaches, which use functional annotation only in an aposteriori manner to interpret gene lists. We also see our work orthogonal to the approach suggested in Pavlidis *et al*. (2002) who use functional annotations for testing groups of genes and not in a diagnostic setting like we do.

In summary, structured analysis of microarrays has the potential to uncover previously unknown molecular disease subentities. Moreover, the novel notion of molecular symptom may allow us to characterize new subgroups of patients. We see perspectives for further development in the exploitation of additional or alternative sources of biological annotation such as KEGG (Kanehisa, 1996) or Transpath (Schacherer *et al*., 2001). Improvement in prediction accuracy may also be achieved by using other classifiers in leaf nodes or by a novel method to generate the overall prediction.

ACKNOWLEDGEMENTS

The authors are grateful to Florian Markowetz, Jörn Tödling, Jochen Jäger, Stefanie Scheid and Stefan Bentink from our work group as well as to our partners Renate Kirschner-Schwabe, Christian Hagemeier and Karl Seeger from the Charité Medical Center for the fruitful discussions. This research has been supported by BMBF grant 03U117/031U217 of the German Federal Ministry of Education and the National Genome Research Network.

REFERENCES

Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. *Proc. Natl Acad. Sci., USA*, **99**, 6562–6566.

Ben-Dor,A. *et al*. (2000) Tissue classification with gene expression profiles. *J. Comp. Biol.*, **7**, 559–583.

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci., USA*, **98**, 13790–13795.

Chang,C.-C. and Lin,C.-J. (2004) Libsvm: a library for support vector machines.

Cheok,M.H. *et al*. (2003) Treatment-specific changes in gene expression discriminate *in vivo* drug response in human leukemia cells. *Nat. Genet.*, **34**, 85–90.

Dettling,M. and Bühlmann,P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics*, **19**, 1061–1069.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Furey,T.S. *et al*. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Gansner,E.R. and North,S.C. (2000) An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, **30**, 1203–1233.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Huang,E. *et al*. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.

Kanehisa,M. (1996) Toward pathway engineering: a new database of genetic and molecular pathways. *Sci. Tech. Japan*, **59**, 34–38.

Khan,J. *et al*. (1999) Expression profiling in cancer using cdna microarrays. *Electrophoresis*, **20**, 223–229.

Pavlidis,P., Lewis,D.P. and Noble,W.S. (2002) Exploring gene expression data with class scores. *Proc. Pac. Symp. Biocomp.*, 474–485.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-00-3.

Schacherer,F. *et al*. (2001) The transpath signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, **17**, 1053–1057.

Schoch,C. *et al*. (2002) Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. *Proc. Natl Acad. Sci., USA*, **99**, 10008–10013.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Slonim,D.K., Tamayo,T., Mesirov,J.P., Golub,T.R. and Lander,E.S. (2000) Class prediction and discovery using gene expression data. In *Proceedings of the International Conference on Computer Biology*, pp. 263–272.

Spang,R. *et al*. (2002) Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol.*, **2**, 369–381.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types using shrunken centroids of gene expression. *Proc. Natl Acad. Sci., USA*, **99**, 6567–6572.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J.A., Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci., USA*, **98**, 11462–11467.

Yeoh,E.-J. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–145.