Annotation-based Distance Measures for Patient Subgroup Discovery in Clinical Microarray Studies

Claudio Lottaz,* Joern Toedling,† Rainer Spang

Max Planck Institute for Molecular Genetics &
Berlin Center for Genome Based Bioinformatics,
Ihnestr. 73, D-14195 Berlin (Germany)

Abstract:

Background Clustering algorithms are widely used in the analysis of microarray data. In clinical studies, they are often applied to find groups of co-regulated genes. Clustering, however, can also stratify patients by similarity of their gene expression profiles, thereby defining novel disease entities based on molecular characteristics. Several distance-based cluster algorithms have been suggested, but little attention has been given to the choice of the distance measure between patients. Even with the Euclidean metric, including and excluding genes from the analysis leads to different distances between the same objects, and consequently different clustering results.

Methodology We describe a novel clustering algorithm, in which gene selection is used to derive biologically meaningful clusterings of samples. Our method combines expression data and functional annotation data. According to gene annotations, candidate gene sets with specific functional characterizations are generated. Each set defines a different distance measure between patients, and consequently different clusterings. These clusterings are filtered using a novel resampling based significance measure. Significant clusterings are reported together with the underlying gene sets and their functional definition.

Conclusions Our method reports clusterings defined by biologically focused sets of genes. In annotation driven clusterings, we have recovered clinically relevant patient subgroups through biologically plausible sets of genes, as well as novel subgroupings. We conjecture that our method has the potential to reveal so far unknown, clinically relevant classes of patients in an unsupervised manner.

1 Introduction

Gene expression profiling using whole genome microarrays has generated large amounts of data in various clinical contexts. One goal of these studies is the discovery of clinically relevant patient subgroups. Of interest are e.g. groups of patients which require a particular treatment.

*Corresponding author

†Current address: EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD (UK)

An example from lymphoma research Alizadeh et al. [AED+00] define two new subtypes of diffuse large B-cell lymphoma based on a hierarchical clustering analysis using a functionally restricted set of genes. The two disease entities refer to distinct differentiation stages of B-cells. Monti et al. [MSK+05] postulate a different partitioning of diffuse large B-cell lymphomas supported by genes which have been excluded from the first analysis. Their disease entities reflect proliferation properties of the B-cell malignancies. None of the results can be easily proven wrong. In fact, they do not contradict each other. The two research groups had a priori different notions as to which genes are relevant. This led to two dissimilar but relevant clusterings of samples.

Different genes - different distances - different results In the context of class discovery, the objects that are to be clustered are patient samples. For clustering, pairwise distances between these objects are calculated. Using the Euclidean metric to do so, does not yet uniquely define these distances, though. Which genes to include in the analysis is very important. Using all measured genes as such is not a good choice. Several independent molecular characteristics of the patients like age, gender, and disease status will overlap and obscure the result. Gene selection is called for but certainly affects the clustering. Each choice of a gene set to use defines a particular distance between any two samples. Different gene sets lead to different distances between the same objects, although we always use the Euclidean metric to compute them. In many clinical studies, gene selection is used for unsupervised analysis, too. The intention is either to reduce noise in the expression data (e.g. [CSF+05]) or, in addition, to focus on reproducible features (e.g. [BRS+01, MSK+05]). However, little attention on the effect of gene discarding on the resulting disease class definition has been given.

The concept of our algorithm Instead of selecting genes according to purely statistical characteristics, we suggest a systematic approach to gene selection according to functional annotation. We describe an algorithm that produces a list of alternative clusterings using different gene sets for computing distances between samples. We derive candidate gene sets from functional annotation data, and filter the list by a novel significance measure for clustering strength.

Previous work Clustering of gene expression data is routine in bioinformatics. Several methods have been suggested in this field (for a review, see Chapter 4 of [Spe03]). Various approaches to score the quality of clusterings, and to determine the best number of clusters exist [DF02, KC01]. All these methods have in common that the underlying metrics need to be specified beforehand. Several authors also have suggested ways to judge stability and statistical significance of clusters [HBV01, LRBB04, MRF+02, MTMG03, MSS+05]. Semi-supervised clustering approaches include additional clinical information about patients. Bullinger et al. [BDB+04] as well as Bair and Tibshirani [BT04] suggest finding classes of patients using a clustering metric derived from the expression data and additional survival times. In a completely unsupervised setting, biclustering [CC00, TSKS04, MO04] and class-finding algorithms [vHHPV01, RL04, VS04] combine the gene selection process with the clustering. These methods produce alternative clusterings and characterize them by underlying gene sets. Unfortunately, such methods are rarely used in clinical studies. One reason might be that a large set of alternative clusterings is hard to interpret, unless the driving genes have a clear functional focus.

The role of functional annotations We believe that the major shortcoming of class discovery algorithms is that they treat gene expression levels as anonymous variables. For many genes, a lot is known about their function and their role in cellular processes. This knowledge is stored in databases like the Gene Ontology [ABB+00], Transpath [SCG+01], Biocarta (http://www.biocarta.com) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kan96]. Today, such annotations are routinely used to interpret results produced by statistical analysis. Several tools for such a-posteriori analysis are available [BS04, DSH+03, AS04, DSD+03, STM+05, GBRV06].

A-priori use of functional annotations Unlike a-posteriori methods, we propose using annotations *within* the statistical analysis of the expression data. In different contexts this a-priori use of functional annotations has already been investigated. Pavlidis et al. [PLN02] and Zien et al. [ZKZL00] use functional annotations to improve the sensitivity of algorithms for detecting differentially expressed genes. Rahnenführer et al. [RDML04] apply pathway annotations to investigate metabolic pathways. Subclass finding in complex clinical phenotypes using functional annotations is the topic of [LS05]. Here, we apply similar concepts to the problem of molecular class discovery in patients.

Outline of the paper In the next section, we describe the clustering procedure as well as the scoring of clustering results. In Section 3, we illustrate the usefulness of functional gene annotation for producing alternative clusterings of samples on a number of cancer related clinical microarray datasets. Finally, we discuss possible extensions of the method and interpret our observations from a biological perspective in Section 4.

2 Method

We present a novel algorithm for producing a list of alternative patient clusterings in clinical microarray studies. The key idea is to use meaningful gene sets for computing distances between samples. For practical use, it is desirable to have functional rationales characterizing clusterings, such as clusterings related to proliferation or apoptosis. To this end, we define candidate gene sets using functional annotations, and call the resulting clusterings *annotation driven*.

We use the k-means algorithm to generate clusterings based on candidate gene sets. The quality of these clusterings is assessed using the *diagonal linear discriminant* (DLD) score [vHHPV01]. In order to determine the statistical significance of scores, we also compute DLD scores for clusterings driven by randomly chosen gene sets. Empirical p-values are calculated and false discovery rates (FDR) computed according to Benjamini and Hochberg [BH95]. Finally, we filter the list of clusterings for minimal subgroup size and to control the FDR. In a nutshell, the algorithm consists of the following steps:

For each biological term / pathway of interest, denoted B_i:

1. Find all n_{B_i} genes annotated to B_i and discard all others.
2. Perform 2-means clustering on the reduced expression matrix. This yields an anno-

tation-driven clustering C_{B_i}.

3. Compute DLD score $S(C_{B_i})$ for this clustering.
4. Draw 10000 random gene sets of size n_{B_i} from the set of all measured genes. For each of them compute steps 2 and 3. This yields a vector $\mathbf{r}_{n_{B_i}}$ of 10000 scores.
5. Assign an empirical p-value to the original clustering, denoting the proportion of entries of $\mathbf{r}_{n_{B_i}}$ being greater or equal than $S(C_{B_i})$.

In the following, we provide more details on certain steps of the procedure.

2.1 Annotation data

We suggest the use of annotation data to generate candidate gene sets of interest. Genes in a candidate set have common involvement in biological processes or pathways. To generate such gene sets, pathway databases such as KEGG [Kan96] and Gene Ontology [ABB$^+$00] are particularly adequate.

Sets of genes collected for a particular application from literature or a biologist's experience are possible alternatives. Very small gene sets should not be considered, since clusterings supported by very few genes are unlikely to represent a clustering of biological interest. On the other hand, sets containing too many genes are prone to be very unspecific, and thus their results are of little explanatory power.

2.2 Distance metric

K-means clustering is based on pairwise object dissimilarities. Objects in our case are the samples' expression profiles. We obtain dissimilarity measures from the family of restricted Euclidean metrics, which we will define next.

Let $(x_i, x_{i'})$ be any two expression profiles, both containing measured expression values for p genes. Reducing the expression profiles to a limited set of genes before computing the distance, can also be interpreted as computing a Euclidean distance specific for gene set G between the original profiles

$$D_G(x_i, x_{i'}) = \sqrt{\sum_{j=1}^{p} I_{j\in G} \cdot (x_{ij} - x_{i'j})^2}$$

where $I_{j\in G}$ is an indicator variable taking the value 1 if gene j is in set G and 0 otherwise. We call D_G a restricted Euclidean metric on patient space.

By selecting different gene sets before clustering, we choose different measures of distance between any two expression profiles. Since the choice of the distance measure affects the outcome of clustering stronger than the choice of the clustering algorithm (see Chapter 14 in [HTF01], clusterings of the same samples with different metrics disagree substantially.

2.3 K-means initialization

K-means clustering critically depends on its initialization step. We derive an initialization based on the first split of a divisive hierarchical clustering (Chapter 6 in [KR90]). Of the resulting two clusters, we compute centroids which provide the starting points for the k-means algorithm [Mac67]. This has been shown to outperform standard k-means with random starting points [MS80]. In other words, k-means is used to refine individual clusters and to correct inappropriate assignments made by the hierarchical method.

2.4 Scoring clusterings

For clustering evaluation, we employ the *diagonal linear discriminant* (DLD) score, adopted from [vHHPV01]. We briefly review it here.

Let $\mathbf{X}$ be the reduced expression matrix with rows containing the genes from the set of interest and columns representing the patient samples. Given a clustering C of samples, i.e. a binary vector of class labels for classes A and B, we are interested in those genes, whose expression levels reflect this class division best. A natural score for this purpose is Student's t-statistic. We discard all genes except those 50 genes with the highest absolute t-statistic. In case there are less than 50 genes in the functional group, all are kept. We avoid clusterings with very few supporting genes by discarding the top m genes with the highest absolute t-statistic to prevent the final DLD score from being strongly influenced by very few genes with extreme expression levels. This also makes results more robust against imprecise annotations. We chose $m = 5$. Discarding the respective rows (genes) from $\mathbf{X}$, yields a shortened expression matrix $\mathbf{X}^*$.

Now, the same projection method, which is used in the classification step of *diagonal linear discriminant analysis* [MKB79], is used to project the samples (columns) of $\mathbf{X}^*$ onto a one-dimensional space. The projection is defined by the vector

$$\mathbf{v} = \mathbf{S}^{-1} \left(\mu_A - \mu_B \right)$$

where μ_K denotes the centroid of all samples of class K and $\mathbf{S}$ is a diagonal matrix containing the weighted sums of within-class variances for each gene g:

$$\mathbf{S}_{gg} = (a-1)\sigma_{gA}^2 + (b-1)\sigma_{gB}^2$$

where A and B denote the two classes with cardinalities a and b respectively. Each patient sample, which is represented as a column of the shortened expression matrix $\mathbf{X}^*_{\bullet j}$, is projected onto the coordinate, given by the inner product $\mathbf{v}^\top \cdot \mathbf{X}^*_{\bullet j}$.

The DLD-Score S of a clustering C is the Student's t-statistic of the projected coordinates:

$$S(C) = \frac{\sqrt{\frac{a \cdot b}{a+b}} \cdot (\mu_{zA} - \mu_{zB})}{\sqrt{\frac{1}{a+b-2} \cdot ((a-1)\sigma_{zA}^2 + (b-1)\sigma_{zB}^2)}}$$

where z denotes the projected coordinates, μ_{zK} and σ^2_{zK} denote the mean and the variance of the projected coordinates of group K, while A and B denote the two groups with cardinalities a and b respectively.

2.5 Assessing clustering significance

We introduce a new approach to address the question whether an annotation-driven clustering is statistically significant. To this aim, we observe clusterings based on randomly drawn gene sets, which have the same size as the set of functionally related genes but otherwise no restrictions on included genes. For each of these random gene sets, we find the optimal clustering and compute its DLD-Score as described above. The score derived from the annotation-driven clustering is compared with these random scores.

The DLD-Scores derived from random gene sets define a null-distribution of scores for gene sets of the given size. For each annotation-driven clustering C, we can compute an empirical p-value $\pi_E(C)$ denoting the proportion of random scores $\mathbf{r}$ being equal to or greater than the annotation-driven clustering's DLD-Score $S(C)$:

$$\pi_E(C) \;=\; \frac{1}{|\mathbf{r}|} \cdot \sum_{r \in \mathbf{r}} I_{r \geq S(C)}$$

where $I_{r \geq S(C)}$ is an indicator variable taking the value 1 if the random score r is bigger or equal than $S(C)$ and 0 otherwise, and $|\mathbf{r}|$ denotes the number of simulated random gene sets. This empirical p-value provides us with a measure of significance for clusterings.

2.6 Multiple testing

The algorithm described so far, determines an empirical p-value for each term we can find associated genes for. Depending on the employed annotation sources and the microarray at hand, hundreds of terms are considered to generate annotation-driven clusterings. Hence, the determination of empirical p-values is subject to multiple testing. A conservative approach to correct for the multiple testing problem is to determine false discovery rates according to Benjamini and Hochberg [BH95]. We employ this correction although its results are to be interpreted with care given the many dependencies between GO and KEGG terms which share commonly associated genes.

2.7 Implementation

We have implemented our clustering method in the statistical programming language R [IG96, R D05]. We employ the divisive hierarchical clustering method from the `cluster` package and the implementation of k-means clustering [HW79] from R's `stats` package. The implementation of the DLD score is taken from the `isis` package [vHHPV01].

We also use Bioconductor's [GCB+04] meta-data packages to retrieve gene annotations for GO and KEGG. Our code is available in the R package `adSplit` [LTS05] from http://compdiag.molgen.mpg.de/software. The package is also part of release 1.8 of the Bioconductor bundle of packages related to the life sciences.

3 Results

We show results of our method on several cancer related datasets from clinical gene expression studies. We focus on the use of Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) for annotations.

3.1 Expression data

We have used 15 clinical microarray studies to investigate the behavior of our clustering procedure. These studies investigate diagnostic and prognostic issues in the context of brain tumors [FCVF+04, NMB+03, PTG+02, RBM+01], breast cancer [HCD+03, WBD+01], leukemia [ASS+02, CYP+03, RMO+04, WJS+04, YRS+02], lung cancer [BKH+02, BRS+01] and prostate tumors [SFR+02].

All 15 microarray studies are based on Affymetrix® GeneChip technology. Eight datasets were generated using the genome wide HG-U95Av2 microarray based on release 95 of UniGene [Sch97]. Four studies are based on the older HU6800 chip, and in [RMO+04] as well as [FCVF+04] the newer HG-U133A chip based on release 133 of UniGene was applied. Finally, Willenbrock et al. have worked with the HG-Focus chip, a microarray holding a subset of the probe-sets of the HG-U133A chip. Table 1 holds further information on the results obtained for these 15 studies.

For each of these datasets, gene expression profiles were background corrected and normalized on probe level using variance stabilization [HvHS+02] before summarizing the probes into probe-set expression levels using median polish [Tuk77] as suggested in the RMA method by Irizarry et al. [IHC+03]. Implementations of these methods were taken from Bioconductor [GCB+04].

3.2 Annotation data

For the systematic exploration of functional gene annotations, we suggest the use of the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). GO holds 17,601 biological terms, while KEGG comprises 231 pathways. For the considered Affymetrix® microarrays, Table 2 states the number of terms and pathways, which have more than 20 probe-sets but less than 10% of all probe-sets on the chip annotated.

Strikingly many GO terms have very few genes attributed: more than 75% of all terms

Author	Cancer type	Study topic	Chip	♯N	♯C	FDR
Freije	brain	survival	U133A	85	71	9.2
Nutt	brain	subtypes	U95Av2	50	8	9.1
Pomeroy	brain	outcome	HU6800	100	23	8.8
Rickman	brain	subtypes	HU6800	51	0	–
Huang	breast (lns)	risk groups	U95Av2	37	40	9.9
Huang	breast (rel)	outcome	U95Av2	52	0	–
West	breast (rel)	outcome	HU6800	49	0	–
Armstrong	leukemia	subtypes	U95Av2	72	18	9.2
Cheok	leukemia	treatment	U95Av2	31	0	–
Ross	leukemia	subtypes	U133A	142	133	10.0
Willenbrock	leukemia	outcome	Focus	45	11	9.6
Yeoh	leukemia	subtypes	U95Av2	327	179	9.6
Beer	lung	outcome	HU6800	96	2	8.3
Bhattacharjee	lung	survival	U95Av2	254	113	9.9
Singh	prostate	subtypes	U95Av2	102	40	8.8

Table 1: Cancer related datasets used for evaluation. In the column '♯C' contains the number of annotation driven clusterings with smallest group size at least 5 when false discovery rate is controlled at 10 %. The column ♯N holds the number of samples. *lns*=lymphnode status, *rel*=relapse.

	Probe-sets	GO	KEGG
HU6800	7129	4534 / 752	130 / 63
HG-U95Av2	12625	5000 / 962	132 / 77
HG-U133A	22283	5417 / 1223	132 / 92

Table 2: Gene sets defined by GO and KEGG per chip. Numbers of gene sets are given before/after filtering for gene sets holding more than 20 and less than 10% of all probe-sets on the chip.

hold less than 20 probe-sets. On the other hand, very few terms are too general holding more than 10% of the genes on the whole-genome microarrays. The KEGG database also defines some very small gene sets, but roughly two thirds hold more than 20 genes.

On commercial Affymetrix® oligonucleotide microarrays, many genes are represented by more than one probe-set, thus several rows in an expression matrix give measurements for the same gene. When extracting probe-sets with a common annotation, either all or none of the probe-sets representing the same gene are included. When drawing random sets of probe-sets, we mimic this fact, by actually drawing Entrezgene-IDs and including all probe-sets mapped to these in our random set. In this manner, we make sure that random scores actually correspond to random gene sets rather than random sets of probe-sets.

All data sets discussed in this article are based on Affymetrix microarrays. Thus, we can use BioConductor's meta-data packages to deduce associations of genes to GO terms and KEGG pathways. Our method, however, is not restricted to this chip technology. For other microarrays the needed annotation data can be extracted from corresponding databases.

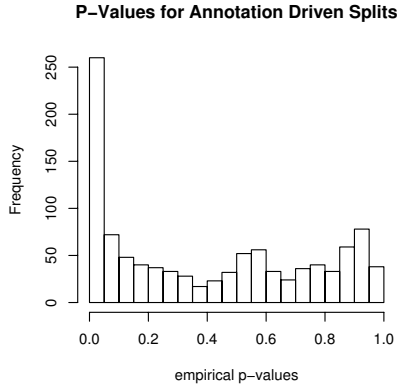


Figure 1: Distribution of empirical p-values of annotation driven clusterings on the gene expression study by Yeoh et al. on leukemia translocations.

3.3 Annotation driven clusterings

We observe that many annotation driven clusterings of patients obtain low empirical p-values. As illustrated in Figure 1 for the leukemia study by Yeoh et al. [YRS+02], the distribution of empirical p-values has a peak close to zero. Apparently, certain gene sets with common functional annotation provide a better basis for clustering samples than random sets of genes. Moreover, the clusterings corresponding to low p-values are of particular interest for the biological focus of their supporting genes.

Our second observation is that many clusterings with small p-values assign only few samples to one of the two clusters. In addition to a stringent p-value, we therefore also require a minimum group size of at least five samples for interesting clusterings. For the datasets analyzed, we thus obtain the number of interesting clusterings shown in the column 'Clusterings' of Table 1.

From the same table, we see that our clustering procedure behaves differently on different datasets. While it finds dozens of annotation-driven clusterings with false discovery rate lower than 10% and size of the small subgroup larger than 5 on most of our evaluation studies, it does not find any clustering in four datasets. In [YRS+02] very heterogeneous expression profiles caused by chromosomal aberrations are included, thus leading to a large number of significant annotation driven clusterings. We observe that our algorithm typically finds fewer annotation driven clusterings in small datasets. This may be caused by our second filtering criteria, which is more stringent on small datasets, given the absolute requirement of 5 samples per group in this criterion.

The set of annotation driven clusterings for one project may be quite heterogeneous. Figure 2 illustrates such a case occurring in the study on embryonic brain tumours by Pomeroy et al. [PTG+02]. Stratifying these tumors by morphological features is controversial. Hence, they present an interesting field of research for diagnosis on a molecular level. The

authors of this study acknowledge that the investigated tumours are very heterogeneous. In accordance with this observation, our method reports clearly differing annotation driven clusterings. Based on terms widely spread over the whole Gene Ontology, we determine 23 different gene sets justifying splits of samples into two groups on significantly better grounds than randomly picked genes.

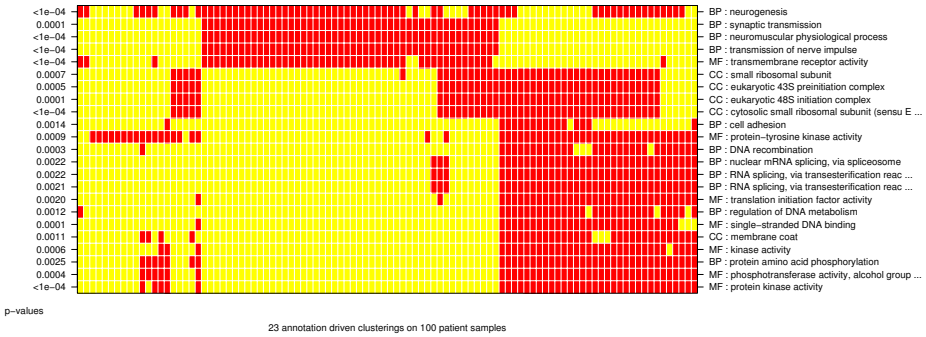
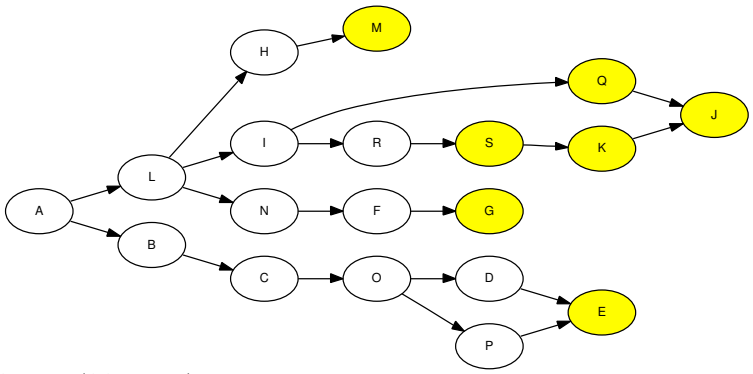


Figure 2: Annotation driven clusterings for the study by Pomeroy et al. Colors code the cluster to which a patient is attributed with respect to the corresponding gene set. In the gene set descriptions to the right of the image, the GO source ontologies of the annotations are indicated by *BP* for biological process *CC* for cellular component and *MF* for molecular function. Columns correspond to samples and rows to gene sets. The image is clustered in both directions in order to bring similar clusterings and similarly attributed samples close together. The depicted set of clusterings achieves a false discovery rate of 8.8%.

3.4 Coherence between clusterings and clinical parameters

The cited datasets from clinical microarray studies come with clinical information. For instance, in the lung-cancer study discussed in [BRS$^+$01], histologically defined subtype assignments are provided for the biopsies, while in [RMO$^+$04], cytogenetically determined translocations are given for each patient. In order to assess the clinical relevance of identified significant clusterings, we compare these with clinical parameters. We employ the χ^2-test to search for clusterings which are highly correlated with clinical parameters.

On several datasets, we observed clusterings of striking correlation with clinical parameters, thus supporting previous findings. For instance, on the acute myeloid leukemia (AML) data set of Ross et al. [RMO$^+$04], we found 11 patient splits for which the two groups correspond to some phenotypical separation of the samples. Less than 10 profiles are attributed inconsistently by these splits to the corresponding phenotypical separation and χ^2 contingency table tests yield p-values below 10^{-10}. Seven of these clusterings consistently separate the group of megakaryocytic leukemia profiles plus one other profile described as having an unspecified AML subtype from the other AML subtypes. The 7 clusterings stem from gene sets annotated to *blood coagulation* (GO:0007596) and related



A: Gene Ontology (GO:0003673)
B: cellular component (GO:0005575)
C: cell (GO:0005623)
D: plasma membrane (GO:0005886)
E: integral to plasma membrane (GO:0005887)
F: cell communication (GO:0007154)
G: cell adhesion (GO:0007155)
H: development (GO:0007275)
I: physiological process (GO:0007582)
J: blood coagulation (GO:0007596)
K: hemostasis (GO:0007599)
L: biological process (GO:0008150)
M: morphogenesis (GO:0009653)
N: cellular process (GO:0009987)
O: membrane (GO:0016020)
P: integral to membrane (GO:0016021)
Q: coagulation (GO:0050817)
R: organismal physiological process (GO:0050874)
S: regulation of body fluids (GO:0050878)

Figure 3: Clusterings driven by the gene sets associated to the 7 nodes colored in yellow identify acute megakaryocytic leukemia with just one conflicting class assignment in the dataset by Ross et al. The figure shows the GO subgraph induced by these nodes.

GO-terms. See Figure 3 for a display of the relationships between the 7 GO-terms and their ancestors within the Gene Ontology.

On the lung-cancer dataset by Bhattacharjee et al. [BRS+01], we identified 17 clusterings showing p-values $< 10^{-10}$ in the χ^2-test and differing by not more than 10 cluster assignments from the corresponding morphological classification of the tumors. 9 of these clusterings separate the group of 20 pulmonary *carcinoid tumors* from all other tumors. Five of the 9 clusterings also assign one or two other profiles to the cluster of carcinoid tumors. The 9 clusterings are derived from gene sets annotated to *central nervous system development* (GO:0007417), *ion channel activity* (GO:0005216) and related terms.

4 Discussion

An important goal of clinical microarray studies is the discovery of cohesive subgroups of patients according to molecular criteria. Commonly, unsupervised clustering is employed to this aim, although the evaluation of clustering results is notoriously difficult. One suggestion, to show whether a clustering is biologically meaningful, is to point out that functional annotation of the genes supporting the clustering are coherent or plausible.

In this paper, we propose an algorithm to use functional annotations stored in the Gene Ontology and the KEGG database of pathways directly to search for cohesive groups of

samples. By selecting genes sharing common annotation in GO or KEGG and limiting gene expression profiles to these, we define distinct distances between samples for each term or pathway. Consequently, different clusterings are found for each GO term or KEGG pathway. A notable difference to other approaches to select genes before clustering (e.g., [BDB+04]) is that the selection stems from independent data, which represent biological expert knowledge and are not affected by experimental variations.

The use of curated databases like GO and KEGG to extract functional annotations leads to the inclusion of some unreliable data. These databases, however, are always incomplete and the computationally derived annotations may contain errors. We expect our approach to be robust against such erroneous annotation data as long as the erroneous annotations do not dominate. Robustness is enhanced by the fact that clusterings are always supported by several genes with common annotation. Another characteristic of the Gene Ontology not taken into account by our method is its hierarchy. Genes annotated to a given GO term are also used to find clusterings for all parent terms. However, we do not very often observe that parents of children with significant clustering also have significant clusterings. The dependency between parents and children does not seem to be very strong.

We applied our method to a number of gene expression data sets (see Table 1) and found several significant annotation driven clusterings, which strongly correlate to patient stratifications based on clinical criteria and agree with previous reports on the biology behind tumor development. For instance, on the acute myeloid leukemia (AML) data set of [RMO+04], we found a large number of significant clusterings. AML is a heterogeneous disease, comprising abnormal proliferation of the precursors of granulocytes, monocytes, and thrombocytes [JHSV01]. Thus, it is not surprising to find many significant clusterings dividing one type of AML from the rest. For example, 7 clusterings that separate AML of the FAB-M7 type, i.e. acute megakaryocytic leukemia, from the other AML types, are based on gene sets attributed to *blood coagulation* (GO:0007596), *cell adhesion* (GO:0007155) and five related terms. Since megakaryocytes give rise to thrombocytes, whose primary function is to mediate cell adhesion to damaged endothelium and blood coagulation, they are bound to excel in the expression of genes involved in these processes. Remarkably, one patient profile that was clinically described as having an unspecified AML subtype is consistently assigned to the cluster of FAB-M7 samples. This sample seems to display molecular characteristics of the FAB-M7 subtype, although it would not be assigned to this subtype based on clinical criteria.

In accordance with other studies, Bhattacharjee et al. [BRS+01] have described lung cancer to be a general concept comprising very different tumor subtypes. We as well observe large biological differences between these subtypes in form of significant annotation driven clusterings. For example, 9 clusterings clearly separate pulmonary *carcinoid tumors* from all other types of lung cancer. These 9 clusterings are derived from gene sets annotated to *central nervous system development* (GO:0007417), *ion channel activity* (GO:0005216) and 7 related terms. Pulmonary carcinoid tumors have been previously reported to be of neuroendocrine origin and to be closely related to brain tumors [ATB+99]. Our finding of remarkable expression of nerve-cell associated genes by these tumors supports such reports.

In summary, the method presented in this paper has the potential to uncover clinically

relevant clusterings in gene expression studies. Moreover, such clusterings may be of particular interest due to the biological focus of their supporting genes.

Acknowledgments

The authors are grateful to Jochen Jäger, Dennis Kostka, 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 fruitful discussions. This research has been supported by BMBF grants 01GS0445 and 01GR0455 of the German Federal Ministry of Education and the National Genome Research Network.

References

[ABB+00] 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):25–29, May 2000.

[AED+00] AA Alizadeh, MB Eisen, RE Davis, C Ma, IS Lossos, A Rosenwald, JC Boldrick, H Sabet, T Tran, X Yu, JI Powell, L Yang, GE Marti, T Moore, J Hudson, L Lu, DB Lewis, R Tibshirani, G Sherlock, WC Chan, TC Greiner, DD Weisenburger, JO Armitage, R Warnke, and LM Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000.

[AS04] B Adryan and R Schuh. Gene-Ontology-based clustering of gene expression data. *Bioinformatics*, 20(16):2851–2, 2004.

[ASS+02] SA Armstrong, JE Staunton, LB Silverman, R Pieters, ML den Boer, MD Minden, SE Sallan, ES Lander, TR Golub, and SJ Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–47, Jan 2002.

[ATB+99] R Anbazhagan, T Tihan, DM Bornman, JC Johnston, JH Saltz, A Weigering, S Piantadosi, and E Gabrielson. Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Res*, 59(20):5119–5122, Oct 1999.

[BDB+04] L Bullinger, K Döhner, E Bair, S Fröhling, RF Schlenk, R Tibshirani, H Döhner, and JR Pollack. Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. *N Engl J Med*, 350(16):1605–16, 2004.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[BKH+02] David G Beer, Sharon LR Kardia, Chiang-Ching Huang, Thomas J Giordano, Albert M Levin, David E Misek, Lin Lin, Guoan Chen, Tarek G Gharib, Dafydd G Thomas, Michelle L Lizyness, Rork Kuick, Satoru Hayasaka, Jeremy MG Taylor, Mark D Iannettoni, Mark B Orringer, and Samir Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, 8(8):816–24, Aug 2002.

[BRS+01] A Bhattacharjee, WG Richards, J Staunton, C Li, S Monti, P Vasa, C Ladd, J Beheshti, R Bueno, M Gillette, M Loda, G Weber, EJ Mark, ES Lander, W Wong, BE Johnson, TR Golub, DJ Sugarbaker, and M Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, 98(24):13790–5, Nov 2001.

[BS04] T Beissbarth and TP Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–5, jun 2004.

[BT04] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):E108, Apr 2004.

[CC00] Y Cheng and G Church. Biclustering of expression data. In *Intelligent System in Molecular Biology*, pages 93–103, aug 2000.

[CSF+05] G Cario, M Stanulla, BM Fine, O Teuffel, NV Neuhoff, A Schrauder, T Flohr, BW Schafer, CR Bartram, K Welte, B Schlegelberger, and M Schrappe. Distinct gene expression profiles determine molecular treatment response in childhood acute lymphoblastic leukemia. *Blood*, 105(2):821–826, Jan 2005.

[CYP+03] MH Cheok, W Yang, CH Pui, JR Downing, C Cheng, CW Naeve, MV Relling, and WE Evans. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet*, 34(1):85–90, May 2003.

[DF02] S Dudoit and J Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3:R36, jun 2002.

[DSD+03] SW Doniger, N Salomonis, KD Dahlquist, K Vranizan, Lawlor SC, and B R Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, jan 2003.

[DSH+03] G Jr. Dennis, BT Sherman, DA Hosack, J Yang, W Gao, HC Lane, and R A Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5):P3, 2003.

[FCVF+04] WA Freije, FE Castro-Vargas, Z Fang, S Horvath, T Cloughesy, LM Liau, PS Mischel, and SF Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*, 64(18):6503–6510, Sep 2004.

[GBRV06] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron. An Improved Statistic for Detecting Over-representated Gene Ontology Annotations in Gene Sets. In A Apostolico, C Guerra, S Istrail, P Pevzner, and M Waterman, editors, *Research in Computational Molecular Biology: 10th Annual International Conference, Proceedings of RECOMB 2006, Venice, Italy, April 2-5, 2006*, volume 3909 of *Lecture Notes in Computer Science*, pages 85–98. Springer, Heidelberg, 2006.

[GCB+04] RC Gentleman, VJ Carey, DM 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, AJ Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, JY Yang, and J Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[HBV01] M Halkidi, Y Batistakis, and M Vazirgiannis. On Clustering Validation Techniques. *J. of Intell. Inform. Systems*, 17(2-3):107–45, 2001.

[HCD+03] E Huang, SH Cheng, H Dressman, J Pittman, MH Tsou, CF Horng, A Bild, ES Iversen, M Liao, CM Chen, M West, JR Nevins, and AT Huang. Gene expression predictors of breast cancer outcomes. *Lancet*, 361(9369):1590–1596, May 2003.

[HTF01] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, 2001.

[HvHS+02] W Huber, A von Heydebreck, H Sültmann, A Poustka, and M Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(Suppl 1):96–104, 2002.

[HW79] JA Hartigan and M A Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–4, 1979.

[IG96] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[IHC+03] RA Irizarry, B Hobbs, F Collin, YD Beazer-Barclay, KJ Antonellis, U Scherf, and TP Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, Apr 2003.

[JHSV01] ES Jaffe, NL Harris, H Stein, and JW Vardiman, editors. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues*, chapter 4. IARC Press, Lyon, France, 2001.

[Kan96] M Kanehisa. Toward pathway engineering: a new database of genetic and molecular pathways. *Sci & Tech Japan*, 59:34–8, 1996.

[KC01] MK Kerr and G A Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci.*, 98(16):8961–5, jul 2001.

[KR90] L Kaufman and P J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[LRBB04] Tilman Lange, Volker Roth, Mikio L Braun, and Joachim M Buhmann. Stability-based validation of clustering solutions. *Neural Comput*, 16(6):1299–323, Jun 2004.

[LS05] 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.

[LTS05] Claudio Lottaz, Joern Toedling, and Rainer Spang. Annotation-Driven Class Discovery. Technical Report 2005/02, Max Planck Institute for Molecular Genetics, Berlin (Germany), 2005.

[Mac67] J B MacQueen. Some Methods for classification and analysis of multivariate observations. In *Symposium on Math, Statistics, and Probability*, volume 1, pages 281–97, 1967.

[MKB79] K Mardia, J Kent, and J Bibby. *Multivariate Analysis*. Academic Press, San Diego, 1979.

[MO04] SC Madeira and A L Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), Jan-Mar 2004.

[MRF+02] LM McShane, MD Radmacher, B Freidlin, R Yu, MC Li, and R Simon. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, Nov 2002.

[MS80] G Milligan and L Sokol. A Two Stage Clustering Algorithm with Robust Recovery Characteristics. *Educational and Psychological Measurement*, 40:755–9, 1980.

[MSK+05] S Monti, KJ Savage, JL Kutok, F Feuerhake, P Kurtin, M Mihm, B Wu, L Pasqualucci, D Neuberg, RC Aguiar, P Dal Cin, C Ladd, GS Pinkus, G Salles, NL Harris, R Dalla-Favera, TM Habermann, JC Aster, TR Golub, and MA Shipp. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105(5):1851–1861, Mar 2005.

[MSS+05] Munneke, Schlauch, Simonsen, Beavis, and Doerge. Adding Confidence to Gene Expression Clustering. *Genetics*, Jun 2005.

[MTMG03] S Monti, P Tamayo, JP Mesirov, and T R Golub. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52(1-2):91–118, 2003.

[NMB+03] CL Nutt, DR Mani, RA Betensky, P Tamayo, JG Cairncross, C Ladd, U Pohl, C Hartmann, ME McLaughlin, TT Batchelor, PM Black, A von Deimling, SL Pomeroy, TR Golub, and DN Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–1607, Apr 2003.

[PLN02] P Pavlidis, DP Lewis, and WS Noble. Exploring gene expression data with class scores. In *Proc Pacific Symposium on Biocomp*, pages 474–85, 2002.

[PTG+02] SL Pomeroy, P Tamayo, M Gaasenbeek, LM Sturla, M Angelo, ME McLaughlin, JY Kim, LC Goumnerova, PM Black, C Lau, JC Allen, D Zagzag, JM Olson, T Curran, C Wetmore, JA Biegel, T Poggio, S Mukherjee, R Rifkin, A Califano, G Stolovitzky, DN Louis, JP Mesirov, ES Lander, and TR Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–42, 2002.

[R D05] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[RBM+01] DS Rickman, MP Bobek, DE Misek, R Kuick, M Blaivas, DM Kurnit, J Taylor, and S M Hanash. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res*, 61(18):6885–6891, Sep 2001.

[RDML04] J Rahnenführer, FS Domingues, J Maydt, and T Lengauer. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[RL04] Volker Roth and Tilman Lange. Feature Selection in Clustering Problems. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[RMO+04] ME Ross, R Mahfouz, M Onciu, HC Liu, X Zhou, G Song, SA Shurtleff, S Pounds, C Cheng, J Ma, RC Ribeiro, JE Rubnitz, K Girtman, WK Williams, SC Raimondi, DC Liang, LY Shih, CH Pui, and J R Downing. Gene expression profiling of pediatric acute myelogenous leukemia. *Blood*, 104(12):3679–87, Dec 2004.

[SCG+01] F Schacherer, C Choi, U Götze, M Krull, S Pistor, and E Wingender. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, 17:1053–7, 2001.

[Sch97] GD Schuler. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Molecular Medicine*, 75(10):694–8, Oct 1997.

[SFR+02] D Singh, PG Febbo, K Ross, DG Jackson, J Manola, C Ladd, P Tamayo, AA Renshaw, AV D'Amico, JP Richie, ES Lander, M Loda, PW Kantoff, TR Golub, and WR Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–9, Mar 2002.

[Spe03] T Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, Florida, USA, 2003.

[STM+05] Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, Sep 2005.

[TSKS04] A Tanay, R Sharan, M Kupiec, and R Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A*, 101(9):2981–6, Mar 2004.

[Tuk77] J W Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading Masschusetts, USA, 1977.

[vHHPV01] A von Heydebreck, W Huber, A Poustka, and M Vingron. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17(Suppl 1):S107–14, 2001.

[VS04] Sudhir Varma and Richard Simon. Iterative class discovery and feature selection using Minimal Spanning Trees. *BMC Bioinformatics*, 5(1):126, Sep 2004.

[WBD+01] 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 U S A*, 98(20):11462–7, Sep 2001.

[WJS+04] H Willenbrock, AS Juncker, K Schmiegelow, S Knudsen, and L P Ryder. Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. *Leukemia*, 18(7):1270–1277, Jul 2004.

[YRS+02] EJ Yeoh, ME Ross, SA 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):133–143, Mar 2002.

[ZKZL00] A Zien, R Kuffner, R Zimmer, and T Lengauer. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol*, 8:407–417, 2000.