BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 10 2005, pages 2424–2429
doi:10.1093/bioinformatics/bti367

Gene expression

Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data

Christian H. Busold[1,*], Stefan Winter[2], Nicole Hauser[3], Andrea Bauer[1], Jürgen Dippon[2], Jörg D. Hoheisel[1] and Kurt Fellenberg[1]

[1]Division of Functional Genome Analysis, Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany, [2]Institut für Stochastik und Anwendungen, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany and [3]Genomics—Proteomics—Screening (GPS), Fraunhofer-Institut für Grenzflächen- und Bioverfahrenstechnik (IGB), Nobelstrasse 12, D-70569 Stuttgart, Germany

Received on August 31, 2004; Revised on February 1, 2005; Accepted on February 28, 2005
Advance Access publication March 3, 2005

ABSTRACT

Motivation: The functional interpretation of microarray datasets still represents a time-consuming and challenging task. Up to now functional categories that are relevant for one or more experimental context(s) have been commonly extracted from a set of regulated genes and presented in long lists.

Results: To facilitate interpretation, we integrated Gene Ontology (GO) annotations into Correspondence Analysis to display genes, experimental conditions and gene-annotations in a single plot. The position of the annotations in these plots can be directly used for the functional interpretation of clusters of genes or experimental conditions without the need for comparing long lists of annotations. Correspondence Analysis is not limited in the number of experimental conditions that can be compared simultaneously, allowing an easy identification of characterizing annotations even in complex experimental settings. Due to the rapidly increasing amount of annotation data available, we apply an annotation filter. Hereby the number of displayed annotations can be significantly reduced to a set of descriptive ones, further enhancing the interpretability of the plot. We validated the method on transcription data from *Saccharomyces cerevisiae* and human pancreatic adenocarcinomas.

Availability: The M-CHiPS software is accessible for collaborators at http://www.mchips.org

Contact: c.busold@dkfz.de

Supplementary information: http://www.dkfz.de/mchips/supplements/supplement_busold_bioinf_OS.pdf

1 INTRODUCTION

Microarrays have become a common tool to query expression levels of thousands of genes or even complete genomes, resulting in large amounts of data to be analyzed. However, the identification of significantly regulated genes is only the first step in analysis. Having lists of differential genes at hand, researchers rely on gene-annotation data to deduce biological meaning. This involves gathering annotations from databases and identifying the characteristics of the extracted set of genes. The Gene Ontology (GO) consortium initiated a standardization of annotation terms making them applicable for different organisms and facilitating data exchange among laboratories and databases (GO Consortium, 2000, 2001). These characteristics render GO annotations a powerful tool for the interpretation of microarray data.

Up to now, methods allowing the analysis of microarray data in the context of GO terms (Hosack *et al.*, 2003; Beißbarth and Speed, 2004; Al-Shahrour *et al.*, 2004; Robinson *et al.*, 2004) are not capable of displaying genes, experimental conditions and annotations in a single plot. These methods are based on a two-step process: initially, a list of regulated genes needs to be computed. Based on this, annotation terms which are significantly over- or under-represented within this list—compared to a reference list that usually consists of all genes on the array—are extracted. Current methods allow only for a comparison of two experimental settings at a time and/or are restricted to one of the main GO categories or to a specific GO level. Furthermore the resulting annotations are presented in long lists. Such a one-dimensional setting does not display the actual similarities among annotations in an optimal way, especially with regard to the rapidly growing annotation data available.

Here, we present a method not limited in the number of experimental conditions that can be compared. In our approach, GO annotations are analyzed at all possible levels of detail simultaneously. Correspondence Analysis (CA) (Hayashi, 1952; Lebart *et al.*, 1984; Greenacre, 1993a) is not limited in the number of experimental conditions that can be compared and has successfully been applied in the analysis of microarray data (Kishino and Waddell, 2000; Fellenberg *et al.*, 2001). The analysis can be enhanced by adding annotations as supplementary rows (Greenacre, 1993b). They appear at indicative positions (centers of gravity) in the final visualization, without contributing directly to it. That means the plotted positions of genes would be the same with or without the supplementary information. CA allows the visualization of genes, experimental conditions and gene-annotations in a single plot. This provides an overview of their associations and thereby an immediate identification of relevant GO annotations even in complex experimental settings.

We have integrated all the steps involved in such an analysis with M-CHiPS (multi-conditional hybridisation intensity processing system) (Fellenberg *et al.*, 2001, 2002). This allows researchers to conduct a complete microarray data analysis, i.e. normalization, filtering, clustering and functional categorization, in one system.

*To whom correspondence should be addressed.

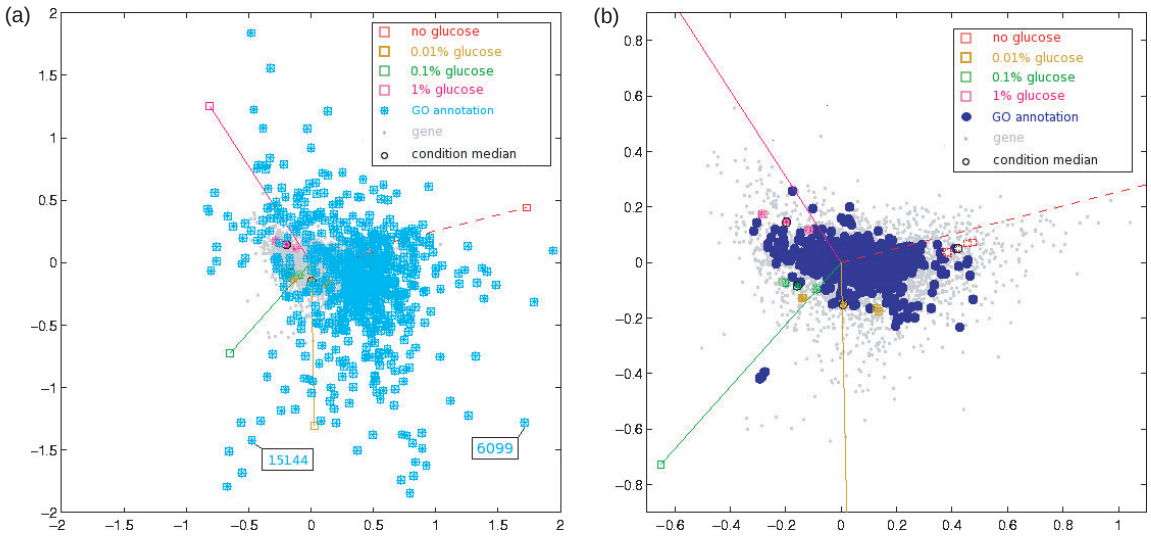

Fig. 1. Correspondence Analysis with GO annotations coded as boolean columns (**a**) and 'sum of genes' rows (**b**). Panel (a) shows GO annotations added to the data-matrix as supplementary columns. They are displayed as cyan squares. In contrast panel (b) shows annotations added as supplementary rows (after summing up the profiles of the annotated genes), here depicted as solid blue circles. In both panels genes are marked as gray dots, experiments as full squares, color coded according to the experimental conditions they belong to (see legend in upper right corners). Numbers adjacent to the annotations reflect the GO-Ids (truncated of 'GO:' and leading zeros). Representations for each condition are depicted in standard coordinates as colored empty squares terminating the lines in the plot (e.g. the green square terminating the green line denotes the standard coordinate of the condition with 0.1% glucose).

2 SYSTEMS AND METHODS

2.1 Obtaining gene annotations

For the experiment described, arrays consisted of 6103 yeast-genes (Hauser *et al.*, 1998). Mapping of these genes to GO terms was carried out using the systematic gene-name (e.g. *YBR166C*) and the common name (*TYR1*) in the association file provided by *Saccharomyces* Genome Database (SGD). A total of 3060 distinct GO annotations were found, which annotate 5506 (90.22%) genes on the array. In Figure 1 only genes that had a maximal normalized intensity value ≥ 10 in any experimental condition (intensity filter) were submitted to analysis. In Figs 2 and 3, genes were filtered out whose transcription intensities remained below 30 and/or showed a minmax-separation [a quality filter that assesses how well repeatedly measured genes are separated under two different conditions (Beißbarth *et al.*, 2000)] of <0.3. In these two figures, the y-coordinates of the data-points were multiplied by -1 (i.e. mirrored at x-axis) for better interpretability. In Figure 3 terms with a mean Spearman's correlation coefficient of <0.8 were discarded.

In the human cancer case study (Fig. 4), genes were filtered out whose transcription intensities remained below 3 000 000 (saturation effects) and above 50 000 and/or showed a minmax-separation of <0.2. The mapping of genes to GO terms was based on the RZPD clone Id. Here, terms containing less than three features and a mean Spearman's correlation coefficient of <0.7 were discarded.

GO terms are hierarchically structured from a most general term (i.e. Gene Ontology) down to specific ones. In the association files a gene is always annotated by the node which carries the most detailed information known about the gene-product. The so-called 'true path' rule defines that all parent nodes of each annotation are also true, i.e. valid for a particular gene-product (GO Consortium, 2001). To utilize this information for CA, we associated the gene not only to the annotation specified in the file but also to each parental node up to the root to enable an analysis at all possible levels of detail at the same time.

Prior to analysis, the root term (GO) and the three main category terms (biological process, cellular component and molecular function) are deleted as well as the annotation 'unknown' from each category, since they do not carry any useful information for interpretation of gene-clusters in CA. Our implementation furthermore enables to focus analysis on any combination of the main GO categories.

2.2 Filtering annotations

As an initial step all annotations comprising only one gene are discarded. The normalized transcription intensity values of the genes being associated to a particular annotation were used to calculate the average of all pairwise correlation coefficients. If anti-correlation of genes in an annotation is considered to be descriptive as well, absolute values of the correlation coefficients can be used instead.

Filtering based on GO evidence codes is implemented, though not applied for the data shown.

2.3 ROC-curves to evaluate filter performance

Receiver operating characteristics (ROC) graphs are a tool to assess the performance of a classifier based on an evaluation set, for which the true categories of all elements are known. This method is described in detail in Swets and Pickett (1982). We applied it to assess the performance of potential filter parameters. To this end, we picked a set of 65 'standard' GO annotations undoubtedly categorizable 'by eye' based on overall similarity and other criteria of the expression profiles of the genes contained therein.

2.4 M-CHiPS system

The majority of the code is written in MATLAB, with extensions in C and PERL. The data is stored in a PostgreSQL database (version 6.5) running on a Sun E450 server. The system is accessible via an x-connection. Experimental annotations can be defined, altered and stored via a web interface. A stand-alone application is currently under development and will be available from the authors by the summer of 2005.

3 RESULTS

To demonstrate the applicability of our approach we analyze a dataset of the model organism *Saccharomyces cerevisiae* focusing on the

well-studied glucose pathway (Yin *et al.*, 2003). The relevant data is publicly available from http://mips.gsf.de/proj/eurofan/eurofan_2/b2/dkfz/results_mce37.txt. *S.cerevisiae* had been grown in media containing different amounts of glucose (0, 0.01, 0.1 and 1%). For each of these conditions, RNA had been isolated, processed and hybridized to microarrays (Hauser *et al.*, 1998). The data had been normalized and filtered (Beißbarth *et al.*, 2000; Fellenberg *et al.*, 2001) such that the data-matrix being submitted to CA holds normalized transcription intensities, rows depicting the genes, columns the experimental conditions. Conditions are represented by the gene-wise median of repeatedly performed hybridizations. The gene annotations were filtered such that only GO terms containing a minimum of five genes are displayed in the analysis.

3.1 Boolean implementation

Adding supplementary variables to the analysis (Micciolo *et al.*, 1985; Greenacre, 1993b) is a well-known method to enhance the interpretability of a CA plot. Moreover, Hoffman and Franke (1986), Charnomordic and Holmes (2001) and Dieterich *et al.* (2003) have shown the applicability of CA for analyzing Boolean matrices. The association of gene-products with annotations can be considered a Boolean variable—annotations for a specific gene-product are either available or not. Accordingly, each annotation is represented by a 0/1-vector, being 1 for each gene that is associated to the particular annotation. The annotation vectors are added to the data-matrix as supplementary columns, which do not contribute to the computation of the principal axes [i.e. are plotted *without mass* (Greenacre, 1993b; Fellenberg *et al.*, 2001)].

In the resulting plot (Fig. 1a), the displayed χ^2-distance is a measure of association among the individual rows (genes) and among the columns (experimental conditions, single hybridizations and GO annotations). Column profiles similar to the average profile (showing no strong association to any row) are plotted close to the centroid of the map (and *vice versa*). Profiles highly dissimilar to the average profile are plotted near the margins of the plot; in Figure 1a, for example, annotation GO:0006099 is plotted in a marked position near the margins of the plot (lower right corner) indicating that the annotation's profile differs significantly from the average profile. A high similarity of two profiles is reflected by a small distance of the data points in the plot. Points which have anti-correlated profiles are located in opposite directions from the centroid. More details on interpretation of CA plots are given in Greenacre (1984, 1993a), Kishino and Waddell (2000) and Fellenberg *et al.* (2001).

In the Boolean implementation, the annotations (depicted as cyan squares) are predominant over genes and conditions, i.e. the profiles of a large number of annotations deviate much stronger from the average profile than even the most differential profiles of the experimental conditions and genes. Thus, only annotations are found near the margins of the plot (Fig. 1a), while both genes and experimental conditions are essentially limited to a small area in the middle of the plot. This interferes with an efficient interpretation. Nevertheless, annotations such as 'tricarboxylic acid cycle' (GO:0006099) and 'carbohydrate transporter activity' (GO:0015144) can already be found in marked positions indicating the relevance of these annotations.

3.2 Intensity based implementation

The interpretability of the plot can be enhanced by diminishing the predominance of the annotations. To this end, we code the annotations as supplementary rows (instead of columns), whose elements are computed based on the expression intensities of the annotated genes. For each annotation, a representative expression profile is calculated by the experiment-wise sum of the annotated genes: let x_{ij} be the normalized expression intensities for gene $i = 1..n$, in condition $j = 1..m$; $A_k \subset \{1..n\}$ denote the set of genes annotated to GO term k. $\sum_{i_k \in A_k} x_{i_k j}$ is used as a representative gene profile for term k. These vectors are added as supplementary rows to the data-matrix (Fig. 1b). As with supplementary columns, supplementary rows do not contribute to the computation of the principal axes (Greenacre, 1993b; Fellenberg *et al.*, 2001). Summation of the expression intensities of the annotated genes places the corresponding annotation in the center of gravity of these genes, e.g. the position of annotation GO:0006099 in Figure 3 represents the center of gravity (centroid) of the annotated genes (tagged by red circles).

Table 1. Gene Ontology annotations forming the cluster in Figure. 1(b)[a]

GO identifier	(Main category) GO term
GO:0008643	(P) carbohydrate transport
GO:0015749	(P) monosaccharide transport
GO:0008645	(P) hexose transport
GO:0015144	(F) carbohydrate transporter activity
GO:0015145	(F) monosaccharide transporter activity
GO:0015149	(F) hexose transporter activity
GO:0015578	(F) mannose transporter activity
GO:0005353	(F) fructose transporter activity
GO:0005355	(F) glucose transporter activity

[a]Cluster members listed with their corresponding GO-Id, main category (P = biological process, F = molecular function) and GO term.

In Figure 1b, in contrast to Figure 1a, the majority of annotations (depicted as blue dots) is densely concentrated around the origin of the map, showing no significant association to any condition. This satisfies the expectation that only a small percentage of all annotations should contain regulated genes in this experimental context. Thus, only a diminutive set of annotations should be found in marked positions near the plot margins.

A distinct cluster of annotations can be observed, which was not detectable in the Boolean approach. The cluster consists of nine different annotations, whose GO identifiers and corresponding terms are listed in Table 1. All annotations describe, at different levels of detail, the activation of carbohydrate-transport into the cell. These are linked to a set of 13 genes, all of which belong to more than one annotation of the cluster (Table 2).

The position of the cluster indicates negative association with the control condition (no glucose in medium) and positive association with the remaining conditions with stronger association, i.e. up-regulation, in response to low glucose signals (0.01 and 0.1% glucose). This is consistent with prior findings identifying *HXT1* to *HXT7* as key enzymes for the uptake of glucose with *HXT2*, *HXT6* and *HXT7* being important for growth on 0.1% glucose (Yin *et al.*, 2003).

The identified annotation cluster is still distinguishable when only genes with the most reproducible expression profiles are analyzed (Fig. 2). The annotations contained in the cluster remain constant (as shown in Table 1). Only the number of distinct genes is reduced to six (marked bold in Table 2).

Table 2. Transporter cluster in Figure 1b[a]

Genes	GO identifiers (truncated of 'GO:' and trailing zeros)
YPL026C (SKS1)	8643, 8645, 15749
YDL194W (SNF3)	5355, 15144, 15145, 15149
YPL244C (HUT1)	8643, 15144
YGL225W (GOG5)	8643, 15144
YHR094C (HXT1)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YMR011W (HXT2)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YDR345C (HXT3)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YHR092C (HXT4)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YHR096C (HXT5)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YDR343C (HXT6)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YDR342C (HXT7)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YJL214W (HXT8)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YJR158W (HXT16)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749

[a]The first column is comprised of the systematic gene name and the common name is given in brackets. The second column lists the annotations of the genes. Genes marked in bold fullfill stringent gene filter criteria (Fig. 2).

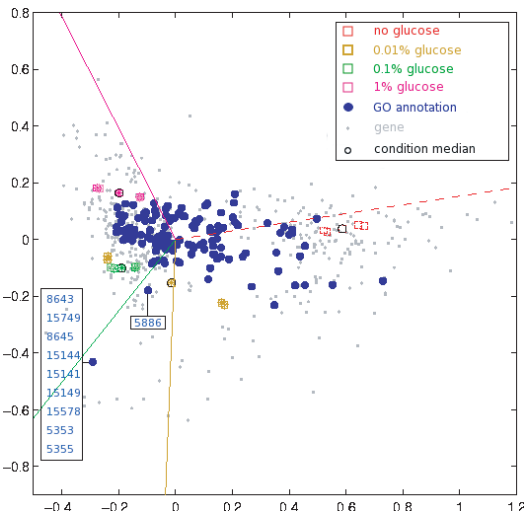

Fig. 2. Correspondence Analysis Map with stringent gene filter (444 genes remaining). GO annotations are added to the data-matrix as supplementary rows (solid blue circles); GO-Ids have been truncated as in the previous figures (see also legend in upper right corner). For better readability, annotations forming the transporter cluster are listed in the adjacent box.

Due to more stringent gene filters, the overall number of displayed annotations is reduced such that a further annotation ('plasma membrane', GO:0005886) becomes apparent, being associated with the same condition (0.1% glucose), i.e. being located in the same direction as the 'transporter-cluster'. The localization of transporter proteins in the plasma membrane is in agreement with the work of Boles and Hollenberg (1997).

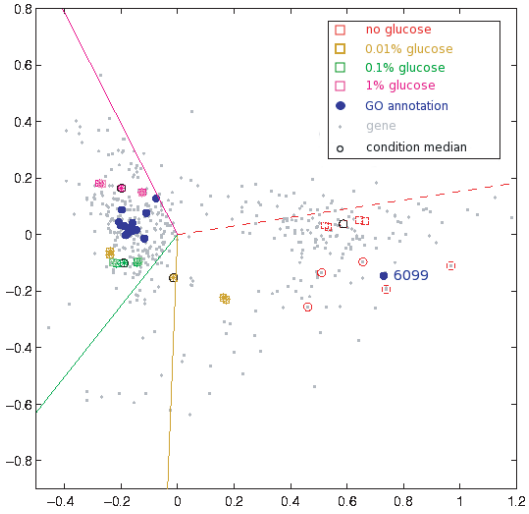

Fig. 3. Correspondence Analysis Map containing filtered GO annotations. To filter out GO terms annotating inhomogenously transcribed gene sets, the mean of Spearman's correlation coefficient was applied. The plot follows the layout of the previous figures; see legend in upper right corner. Annotations forming the cluster left of the centroid are listed in Table 3. Genes annotated as 'tricarboxylic acid cycle' (GO:0006099) are encircled in red. Note that this annotation represents the center of gravity of the corresponding genes.

3.3 Identifying the most descriptive annotations

The results shown so far were based solely on filtering genes—displaying all associated annotations having a minimum of five genes. With increasing amounts of gene-information available, the display of all available annotations would severely reduce the clarity of the plot. Thus, for initial analysis, filtering out annotations that are not likely to provide useful information (i.e. annotations not functionally characterizing any experimental condition or cluster) leads to a better interpretation of the analysis plot and unravels information that would be lost otherwise.

We consider an annotation as helpfully descriptive if the associated genes show a common transcription profile (Fig. 4 of immunoglobulin related genes in Kishino and Waddell, 2000). To measure the descriptiveness of an annotation, we compared the non-parametric correlation coefficients of Spearman and Kendall. Their performance has been evaluated on a set of self-defined 'standard annotations' and analyzed by receiver operating characteristics (ROC) graphs (Swets and Pickett, 1982). The coefficients perform very similarly, even though at high cut-off values, the best classification was achieved with Spearman's rank correlation coefficient (data not shown).

Figure 3 shows the resulting CA-map when applying Spearman's rank correlation coefficient as an annotation filter. The total number of displayed annotations was reduced to 15 (Table 3), enhancing the clarity of the plot. Here, annotation 'tricarboxylic acid cycle' (GO:0006099, lower right quarter of the plot) is associated with the control condition, suggesting that the annotated genes are repressed

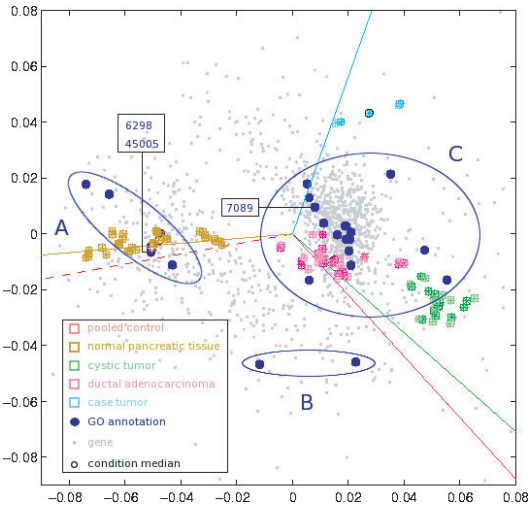

Fig. 4. Correspondence Analysis Map of human pancreatic cancer study. Comparison of ductal adenocarcinomas (pink), cystic tumors (green) and normal tissue (yellow). A possible new tumor entity is colored cyan. The plot follows the layout of the previous figures: see legend in lower left corner. GO annotations are grouped in three different clusters named A, B and C .

Table 3. Gene Ontology annotations displayed in Figure 3.[a]

GO identifier	(Main category) GO term
GO:0000027	(P) ribosomal large subunit assembly and maintenance
GO:0000028	(P) ribosomal small subunit assembly and maintenance
GO:0042257	(P) ribosomal subunit assembly
GO:0005840	(C) ribosome
GO:0005830	(C) cytosolic ribosome (sensu Eukarya)
GO:0005842	(C) cytosolic large ribosomal subunit (sensu Eukarya)
GO:0005843	(C) cytosolic small ribosomal subunit (sensu Eukarya)
GO:0030529	(C) ribonucleoprotein complex
GO:0004553	(F) hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0015926	(F) glucosidase activity
GO:0006096	(P) glycolysis
GO:0006099	**(P) tricarboxylic acid cycle**
GO:0006445	(P) regulation of translation
GO:0006450	(P) regulation of translational fidelity
GO:0008652	(P) amino acid biosynthesis

[a]Main GO category for each annotation is given in brackets (P = biological process, F = molecular function, C = cellular component). The annotation marked in bold is positively associated to the control condition. The remaining annotations are negatively associated to the control condition.

in the presence of glucose. This interpretation is supported by what is known about the regulation of tricarboxylic acid (TCA)-cycle-genes (Boles and Hollenberg, 1997).

In the opposite direction to the centroid, a cluster of annotations can be found, which apart from the obvious annotations ‘glycolysis’ and ‘glucosidase activity’, is mainly comprised of annotations referring to the ribosome (Table 3). Their positions in the CA-map indicate up-regulation of the corresponding genes at 0.1 and 1% glucose, consistent with Yin *et al.* (2003). In the presence of sufficient amounts of glucose, yeast cells invest energy in the production of ribosomes to enable rapid growth and reproduction. This is also reflected by the up-regulation of genes responsible for ‘amino acid biosynthesis’ (GO:0008652), which is essential for prolonged growth.

3.4 Human cancer case study

To demonstrate the performance of our approach in a more complex experimental context, we analyzed the dataset described in Esposito *et al.* (2004). RNA from human ductal adenocarcinomas, cystic tumors and normal pancreas tissue was extracted, labeled and hybridized to a cDNA microarray. The resulting data was processed analogous to the previous dataset. Annotations were added as supplementary rows to the data-matrix (intensity based implementation) and filtered by Spearman’s correlation coefficient such that the number of displayed annotations was reduced to 35 (Fig. 4).

The separation of normal tissue from cancer samples is clearly visible along the x-axis, whereas the new tumor entity (Esposito *et al.*, 2004) separates from ductal and cystic tumors along the y-axis. Three clusters of annotations can be distinguished: annotations associated to normal tissue (A), associated to ductal and cystic (B) and generally tumor associated (C). The complete list of displayed annotations is listed in Table S1 (supplementary material).

Amongst others, annotations ‘mismatch repair’ (GO:0006298) and ‘maintenance of fidelity during DNA-dependent DNA replication’ (GO:0045005) are contained in the normal-associated cluster A. The genes annotated to them are *MSH2*, *MSH3* and *MLH1*. These are DNA mismatch repair enzymes. It has been reported that a loss of their function could be associated with invasive bladder cancer (Thykjaer *et al.*, 2001) and these genes could be potential prognostic factors in colorectal cancers (Jansson *et al.*, 2003; Plaschke *et al.*, 2004).

Annotations in cluster B should be evaluated cautiously. Even though there are several different features (i.e. clones) associated to each of the annotations, effectively there is only one gene per annotation, since the annotated clones contain identical genes. This suggests that the addition of a software option to remove or concatenate clones from the same gene might be useful. In the tumor-associated cluster C, the annotation ‘traversing start control point of mitotic cell cycle’ (GO:0007089) can be found. The comprised genes (represented by three different clones), *CDC2* and *CDC25C*. It is well known that cyclin-dependent kinases play a crucial role in controlling the cell cycle (Murray, 2004) with *CDK10* having a potential role in regulating the G2/M phase (Kasten and Giordano, 2001).

4 DISCUSSION

To facilitate the functional interpretation of microarray data we present two approaches for the integration of gene annotation data in CA. The first approach builds on coding the GO annotations as Boolean variables (Hoffman and Franke, 1986; Charnomordic and Holmes, 2001; Dieterich *et al.*, 2003). However an intuitive interpretation of the plot is hampered by the predominance of the annotations (Fig. 1a). This is due to the Boolean nature of the annotation vector, resulting in large relative changes for which CA is sensitive. To increase the clarity of the plot, we combined the expression profiles of the annotated genes for each GO annotation by summation. These combined profiles function as representatives for the annotations and are added to the data-matrix as supplementary rows.

Apart from smoothly interfacing transcription data with Boolean variables, the intensity-based coding improves on first extracting relevant genes from transcription profiling before imposing a Boolean

variable (Dieterich *et al.*, 2003). The intensity-based coding could prove to be generally applicable to interface transcription data to further variables of Boolean nature, such as affiliation to regulatory sequences or methylation.

Alternatively to representing annotations by the sum of the annotated genes, the median expression profile could be used. In the majority of cases the resulting plots are virtually the same (data not shown). Nevertheless, for more general annotations (which comprise a higher number of genes), commonly only a small portion of the annotated genes exhibits differential expression. Representing these by their median will neutralize the influence of the few differential genes, such that the annotation will be plotted near the centroid, whereas summation accounts for the possibility of only a few differential genes in an annotation.

Correspondence Analysis not only allows plotting annotations as supplementary rows (Greenacre, 1993b), but also computing the principal components based on the annotations (i.e. give them mass), which might be helpful in some cases. However, in the *Saccharomyces* data discussed, plotting annotations with mass only marginally changes their position in the plot (data not shown). Moreover, in many cases annotations for the most differential genes are either not available or not relevant in the particular experimental context. Principal components soley computed on annotations would not display these genes at marked positions, such that this information would be lost.

For even better interpretability of the map, an annotation filter was applied which is based on the homogeneity of the expression profiles of the annotated genes. It leaves only the most descriptive annotations to be displayed in analysis: all annotations selected by the filter (Table 3) describe functional processes that are known to be regulated in the given experimental context. Without filtering, all of those (except for GO:0006099) were plotted in positions (Fig. 1b) masked by a large number of comparably uninformative GO annotations, whose genes are inhomogenously transcribed yet sum-up to profiles of considerable variability.

We validated our approach on the well-understood glycolysis pathway in the model organism *S.cerevisiae* and in the more complex setting of a human cancer case study. The identified annotations describe processes that are important in the development and progression of cancer and thus provide a valid overview of mechanisms relevant in cancer.

In summary, we provide a method which facilitates the functional interpretation of microarray data by the integration of GO annotations in CA. Our method is not limited in the number of experimental conditions to be compared simultaneously and analyzes GO annotations at all possible levels of detail. Moreover, especially the ability of combining experimental conditions, differentially transcribed genes and shared GO annotations in a single plot is crucial for data interpretation, since the location of the various clusters in the plot provides additional information and allows easy identification of associations between and among them, even in complex experimental settings. Representing shared gene annotations as supplementary data rows is shown to add functionality to the already powerful CA approach. This facilitates the interpretation of microarray data.

REFERENCES

Al-Shahrour,F. *et al*. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Beißbarth,T. and Speed,T. (2004) Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Beißbarth,T. *et al*. (2000) Processing and quality control of DNA array hybridization data. *Bioinformatics*, **16**, 1014–1022.

Boles,E. and Hollenberg,C.P. (1997) The molecular genetics of hexose transport in yeasts. *FEMS Microbiol. Rev.*, **21**, 85–111.

Charnomordic,B. and Holmes,S. (2001) Correspondence Analysis with R. *Stat. Comp. Stat. Graph. Newsletter*, **12**, 19–25.

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

GO Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.

Dieterich,C. *et al*. (2003) Exploring potential target genes of signaling pathways by predicting conserved transcription factor binding sites. *Bioinformatics*, **19** (Suppl. 2), II50–II56.

Esposito,I. *et al*. (2004) Microcystic tubulopapillary carcinoma of the pancreas: a new tumor entity? *Virchows Arch.*, **444**, 447–453.

Fellenberg,K. *et al*. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.

Fellenberg,K. *et al*. (2002) Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. *Bioinformatics*, **18**, 423–433.

Greenacre,M.J. (1984) *Theory and Applications of Correspondence Analysis*, 1st edn. Academic Press, London, p. 223.

Greenacre,M.J. (1993a) *Correspondence Analysis in Practice*, 1st edn. Academic Press, London, pp. 36 and 181–183.

Greenacre,M.J. (1993b) *Correspondence Analysis in Practice*, 1st edn. Academic Press, London, pp. 95–102.

Hauser,N.C. *et al*. (1998) Transcriptional profiling of all open reading frames of *Saccharomyces cerevisiae*. *Yeast*, **14**, 1209–1221.

Hayashi,C. (1952) On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Institut. Statist. Math.*, **5**, 121–143.

Hoffman,D. and Franke,G. (1986) Correspondence analysis graphical representation of categorical data in marketing research. *J. Market. Res.*, **XXIII**, 213–227.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with ease. *Genome Biol.*, **4**, R70.

Jansson,A. *et al*. (2003) Combined deficiency of hMLH1, hMSH2, hMSH3 and hMSH6 is an independent prognostic factor in colorectal cancer. *Int. J. Oncol.*, **22**, 41–49.

Kasten,M. and Giordano,A. (2001) Cdk10, a Cdc2-related kinase, associates with the Ets2 transcription factor and modulates its transactivation activity. *Oncogene*, **20**, 1832–1838.

Kishino,H. and Waddell,P. (2000) Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 83–95.

Lebart,L., Morineau,A. and Warwick,K. (1984) *Multivariate Descriptive Statistical Analysis*. Wiley, New York.

Micciolo,R. *et al*. (1985) Correspondence analysis in a study of the clinical evolution of uncomplicated chronic relapsing alcoholic pancreatitis. *Stat. Med.*, **4**, 303–309.

Murray,A. (2004) Recycling the cell cycle: cyclins revisited. *Cell*, **116**, 221–234.

Plaschke,J. *et al*. (2004) Loss of MSH3 protein expression is frequent in MLH1-deficient colorectal cancer and is associated with disease progression. *Cancer Res.*, **64**, 864–870.

Robinson,P.N. *et al*. (2004) Ontologizing gene-expression microarray data: characterizing clusters with gene ontology. *Bioinformatics*, **20**, 979–981.

Swets,J.A. and Pickett,R.M. (1982) *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

Thykjaer,T. *et al*. (2001) Functional analysis of the mismatch repair system in bladder cancer. *Br. J. Cancer*, **85**, 568–575.

Yin,Z. *et al*. (2003) Glucose triggers different global responses in yeast, depending on the strength of the signal, and transiently stabilizes ribosomal protein mRNAs. *Mol. Microbiol.*, **48**, 713–724.