Correspondence Analysis and Data Warehousing with Microarray Data

Sat 30-11-02 am

Kurt Fellenberg

Contents

Abbreviations

' - minutes

AI - **a**rtificial **i**ntelligence

ANN - **a**rtificial **n**eural **n**etwork

BLOB - **b**inary **l**arge **o**bject

CA - **c**orrespondence **a**nalysis

CART - **c**lassification **a**nd **r**egression **t**rees

CAST - **c**lustering **a**ffinity **s**earch **t**echnique

cDNA - **c**omplementary DNA

CGI - **c**ommon **g**ateway **i**nterface

CLICK - **cl**uster **i**dentification via **c**onnectivity **k**ernels

cond. - (experimental) condition

DBMS - **d**ata**b**ase **m**anagement **s**ystem

DNA - **d**eoxyribo**n**ucleic **a**cid

EBI - **E**uropean **B**ioinformatics **I**nstitute

ERM - **e**ntity-**r**elationship **m**odel

EST - **e**xpressed **s**equence **t**ag

exp. - experiment

GO - **g**ene **o**ntology

HMS - **h**ybridization-**m**edian-determined **s**caling

HTML - **h**yper**t**ext **m**arkup **l**anguage

ID - **id**entifier

ISIS - **i**dentifying **s**plits w**i**th clear **s**eparation

LIMS - **l**aboratory **i**nformation **m**anagement **s**ystem

MCE - **m**ulti**c**onditional **e**xperiment

M-CHIPS - **m**ulti-**c**onditional **h**ybridization **i**ntensity **p**rocessing **s**ystem

MDS - **m**ulti**d**emensional **s**caling

meas. - measurement (here referring to a dataset that comprises one value per spot for each spot on the array)

MIAME - **m**inimal **i**nformation **a**bout **m**icroarray **e**xperiments

min - minutes (also referred to by a ')

mRNA - **m**essenger RNA

NBS DES - **N**ational **B**ureau of **S**tandards (USA) **D**ata **E**ncryption **S**tandard

OD - **o**ptical **d**ensity

OMG - **o**bject **m**anagement **g**roup

ORF - **o**pen **r**eading **f**rame

OS - **o**perating **s**ystem

PCA - **p**rincipal **c**omponent **a**nalysis

PKC - **p**rotein **k**inase **C**

REVEAL - **rev**erse **e**ngineering **al**gorithm

RNA - **r**ibo**n**ucleic **a**cid

SAGE - **s**erial **a**nalysis of **g**ene **e**xpression

SNP - **s**ingle **n**ucleotide **p**olymorphisms

SQL - **s**tructured **q**uery **l**anguage

tab - tabulator

UML - **u**nified **m**odeling **l**anguage

WT - **w**ild **t**ype

WWW - **w**orld **w**ide **w**eb

XML - e**x**tensible **m**arkup **l**anguage

Analysis of microarray data:

- Planar embedding of hybridization intensities
- Storage of experiment descriptions

Kurt Fellenberg
Theoretical Bioinformatics Dept.
German Cancer Research Center, Heidelberg

Requirements:

- Integrated DB storage of microarray data stemming from
  → radioactive (monochannel) and
  fluorescent (multichannel) labelling
  → different organisms / fields of research
  (yeast, *Arabidopsis*, human tumor bisopsies, ...)
  incl. gene- and experiment annotations

- Preprocessing

- Data analysis and visualization

... in a nutshell

Microarray technology provides access to expression levels of thousands of genes at once, producing large amounts of data. However, the data show a considerable level of noise, low-level signal intensities are unreliable and datasets commonly comprise outliers. Moreover, a gene set observed to have a certain expression profile of interest will contain a considerable number of false-positives because of the large number of genes under study compared to the small number of conditions. Therefore, in addition to the ability to make amenable both genes and conditions, analysis has to meet certain requirements. It has to be capable of integrating multiple repeat hybridizations for each experimental condition. In addition, the method has to suppress noise and should not be distracted by outliers.

The talk presents a storage system as well as methods to study interdependencies among large-scale microarray data. I applied correspondence analysis as an explorative statistical tool to study interdependencies both between and among sets of variables, i.e. genes and hybridizations that result from expression profiling. Data are carefully preprocessed and correspondence analysis is performed in a way that integrates replicated hybridizations, accounts for noise, and circumvents outliers, thus adapting the method to the particular pitfalls of microarray data. Correspondence analysis is a projection method. Much like principal component analysis it displays a low dimensional projection of the data, e.g. into a plane. However, it does this for two variables simultaneously revealing associations between them. To introduce the method, I show its application to the well-known *Saccharomyces cerevisiae* cell-cycle synchronization data of Spellman *et al.* (Mol. Biol. Cell 9 (1998), 3273-3297). Furthermore, correspondence analysis has been applied to a non-time-series data set of our own, thus supporting its general applicability to microarray data of different complexity, underlying structure and experimental strategy (both two-channel fluorescence-tag and radioactive labeling).

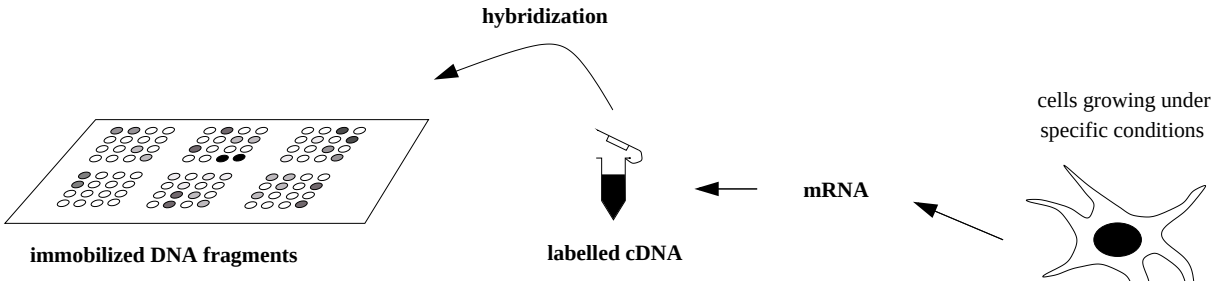
Any method which is, like correspondence analysis, suitable for the analysis of hybridization signals, is best used having access to a database holding the large datasets in a defined common format, ready for preprocessing and analysis. However, it is not sufficient to provide this platform only for hybridization intensities. It is equally necessary to supplement the intensity data by information about genes that are represented by the array spots, and about the experimental conditions for biological interpretation. For interpretation of large data sets,

these annotation data should be in a format amenable to computer aided analysis because they are too numerous for visual inspection. Including annotated experimental parameters into statistical analysis offers the opportunity to identify the global players behind transcription patterns.

Free-text annotations of recent microarray databases are not suited for direct statistical access. Parameter sets used for experiment annotation still change continously, and standards only comprise minimal conventions that do not enable extensive description. Complex and highly diverse experimental settings cause a high complexity and diversity in experiment descriptions, requiring also a higher flexibility in data storage than that achieved by standard database solutions. This is true in particular when data are stored in a statistically accessible format restricted to defined values. A structure which is independent of the particular parameter set enables updates of annotation hierarchies during normal database operation without altering the structure.

A system has been developed and implemented to meet the above requirements and to integrate correspondence analysis into a larger framework of data platform and supplemental methods. It has been named M-CHIPS (Multi-Conditional Hybridization Intensity Processing System). It allows for statistical data analysis of all of its components including the experimental annotations. It addresses the rapid growth of the amount of hybridization data, more detailed experimental descriptions, and new kinds of experiments in the future. Although different organism-specific databases may contain different parameter sets for experiment annotation, they share the same structure and therefore can be accessed by the very same statistical algorithms.

Correspondence Analysis



convert into numerical values

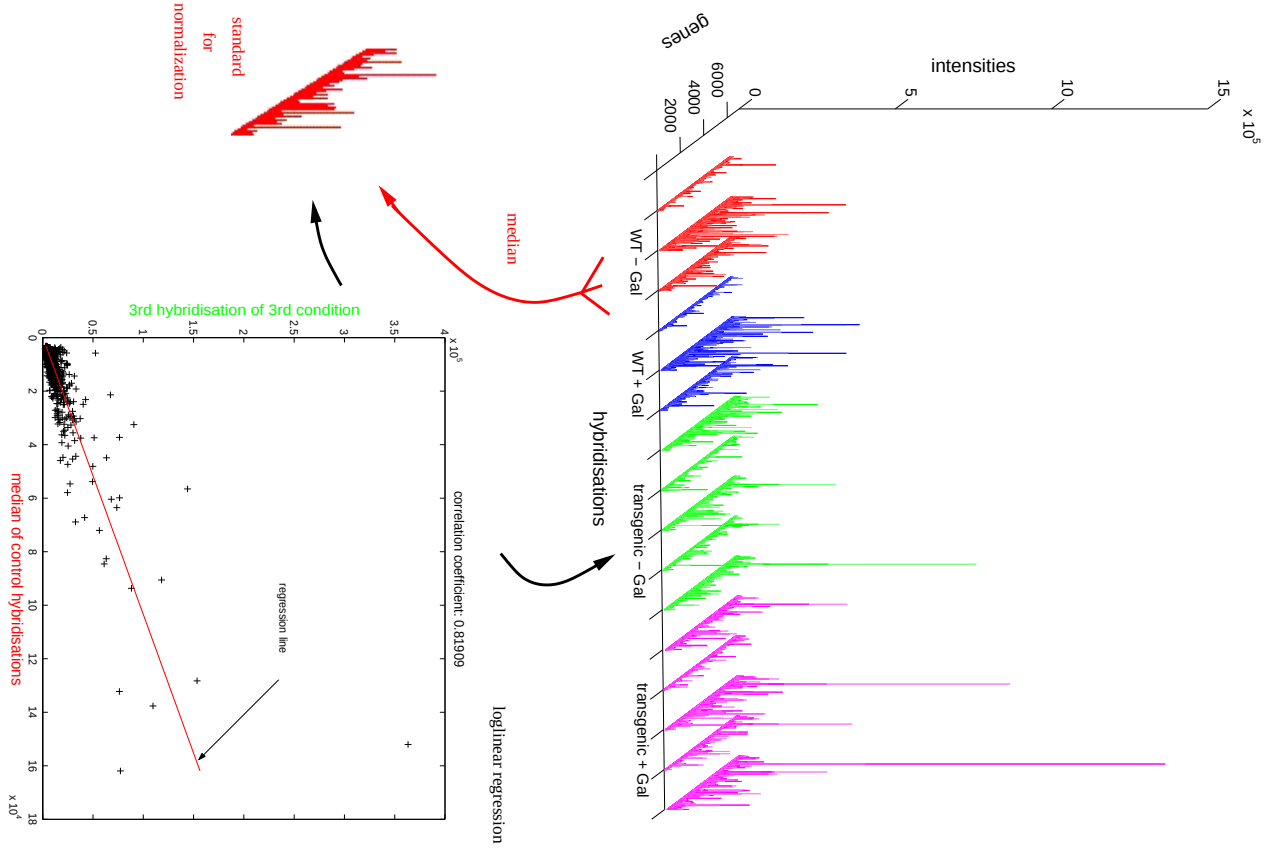
Intensity Table

	control condition			condition 1			condition 2		
	hybr. 1	hybr. 2	hybr. 3	hybr. 4	hybr. 5	hybr. 6	hybr. 7	hybr. 8	hybr. 9
gene 1	14,243	12,154							
gene 2	5,323	27,152							
gene 3	10,300	1,407	···						
gene 4	1,007	3,101							
gene 5	100,232	120,993							
gene 6									
gene 7	⋮								
gene 8									
gene 9									
⋮									

mRNA is prepared from cells growing under specific experimental conditions. It is labeled, i.e converted to more stable cDNA by reverse transcription using radioactively or fluorescence-tagged nucleotides and hybridized to an array. The scheme depicts only radioactive-label, i.e. a single-channel setup for simplicity. The detected signals are then converted into numbers by imaging software. I will refer to a set of conditions as a multi-conditional experiment when all hybridizations are done with reference to one and the same control condition. Multiple measurements for each condition, involving repeated sampling, labeling and hybridization,

offer the opportunity of extracting more robust signals. For the simple case of one channel per hybridization and with repeatedly performed hybridizations for each experimental condition, I will call the individual data set a measurement and represent it by a separate column in the table. One condition of a multi-conditional experiment can thus comprise several columns.

However, the intensity measurements in this table must not be taken at face value. Different levels of background may result in additive offsets, or different amounts of mRNA or different label incorporation rates may lead to multiplicative distortions among the measurements. Therefore the columns of the table have to undergo a normalization procedure, correcting for affine-linear transformation among the columns.



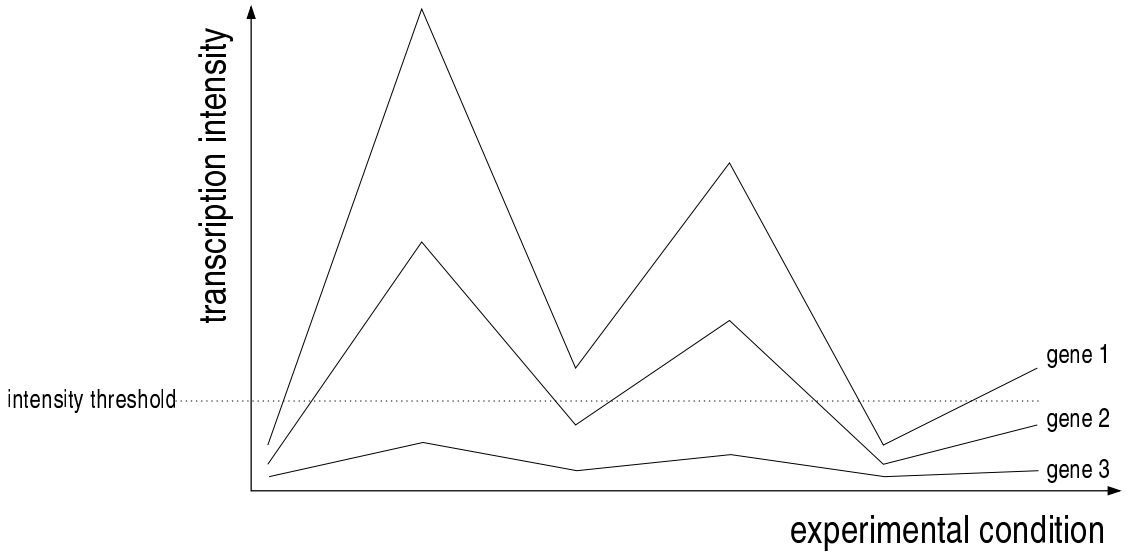
One measurement is fitted versus a control measurement. The performance may be judged from the scatterplot of the raw data (measurement versus control measurement). In this plot, a regression line represents the multiplicative distortion (slope) and additive offset (intersect) determined by the fitting algorithm. The performance of the fit is visible in how well the regression line matches the central dense part of the cloud. Furthermore it can be observed which properties of the raw data led to an eventually suboptimal result. The scale of the plot can be switched between linear and double-logarithmic. In log scale, the regression line appears as a curve whose curvature depends on the additive offset between the two measurements.

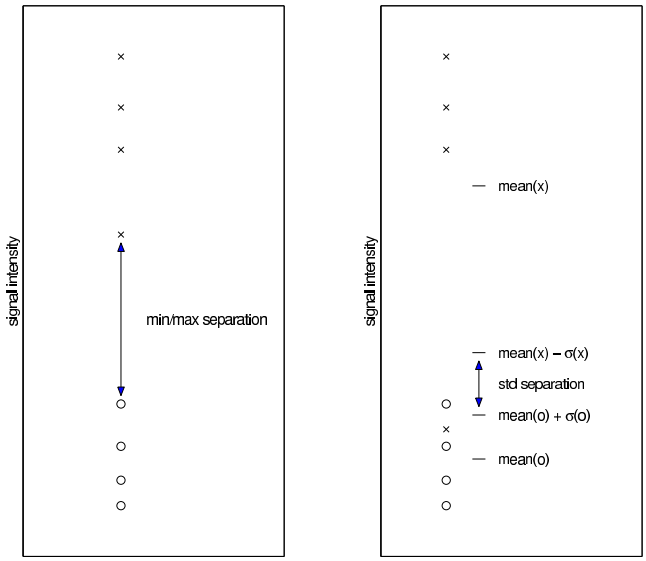
In order to normalize a whole multiconditional experiment, the above step is iterated. All measurements are iteratively normalized with respect to one and the same control condition, such that they can be compared afterwards. M-CHIPS discriminates between mono- and multichannel experiments, applying different control measurements and iteration steps. For monochannel (e.g. radioactive) data, each measurement is normalized versus the genewise median of the hybridizations for the control condition, resulting in absolute intensities. For multichannel hybridizations, the channel belonging to the control condition serves to normalize the other channel(s) of the same hybridization. Here, the normalized intensity values are not analyzed as such, but result in intensity ratios, calculated immediately after normalization. Normalization requires, that each hybridization comprises one channel obtained from the same control condition.

Filtering: Extraction of genes showing

- signal intensities clearly above the detection limit,
- significant relative change, and
- good reproducibility of this change

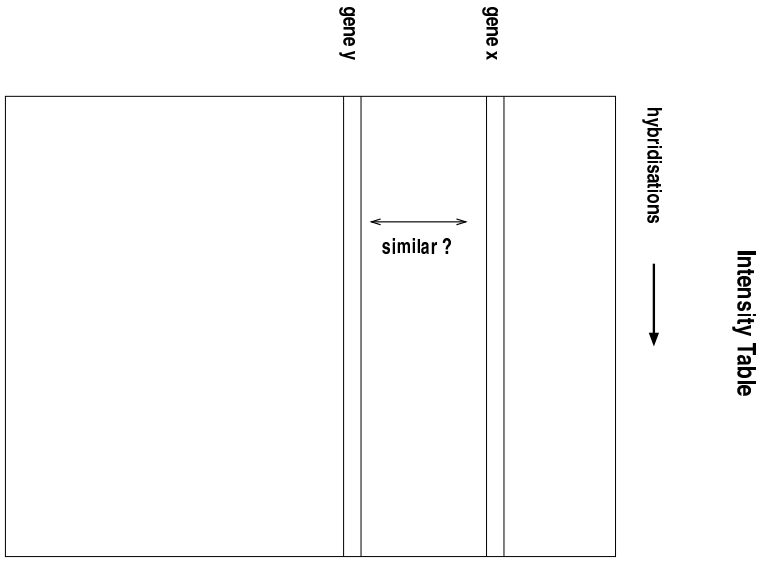
Subsequently it is advisable to disregard all genes which do not appear to be expressed under any of the conditions, or the transcription values which do not reproducibly change between the different conditions under study.



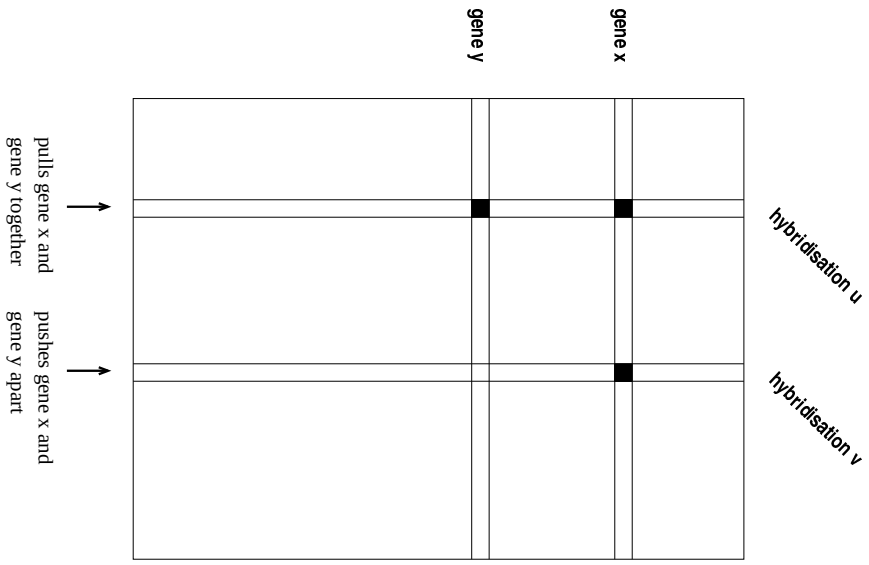


Distributions of repeated measurements are differential among the genes, depending on the intensity level. Usually, there are not more than three to five values per gene and condition

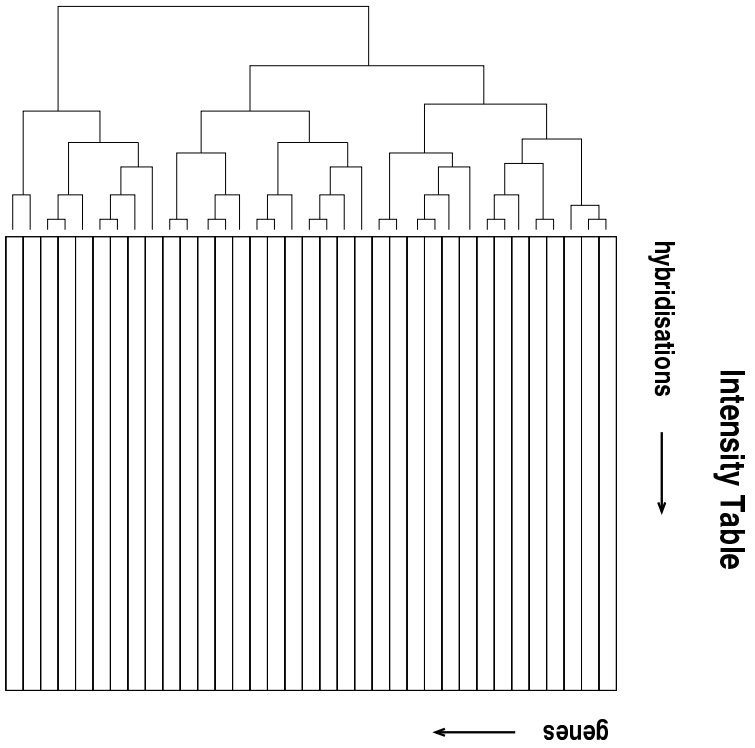
available for averaging. Here they are denoted as circles and crosses for control and non-control condition, respectively. I decided to rely on the minimal separation between two conditions (minmax-separation). Positive minmax-separation is restricted to well-sorted arrangements of the measurements of two conditions as shown in the left panel. Outliers as in the right panel lead to a negative minmax-separation. Tim Beißbarth developed the idea of diminishing the separation between the condition-means by one standard deviation (σ) of either condition set. The standard-deviation separation is less restrictive which is preferable when higher numbers of repeatedly performed measurements are available. In these cases it is desirable to tolerate single outliers in otherwise well-sorted sets of measurements.



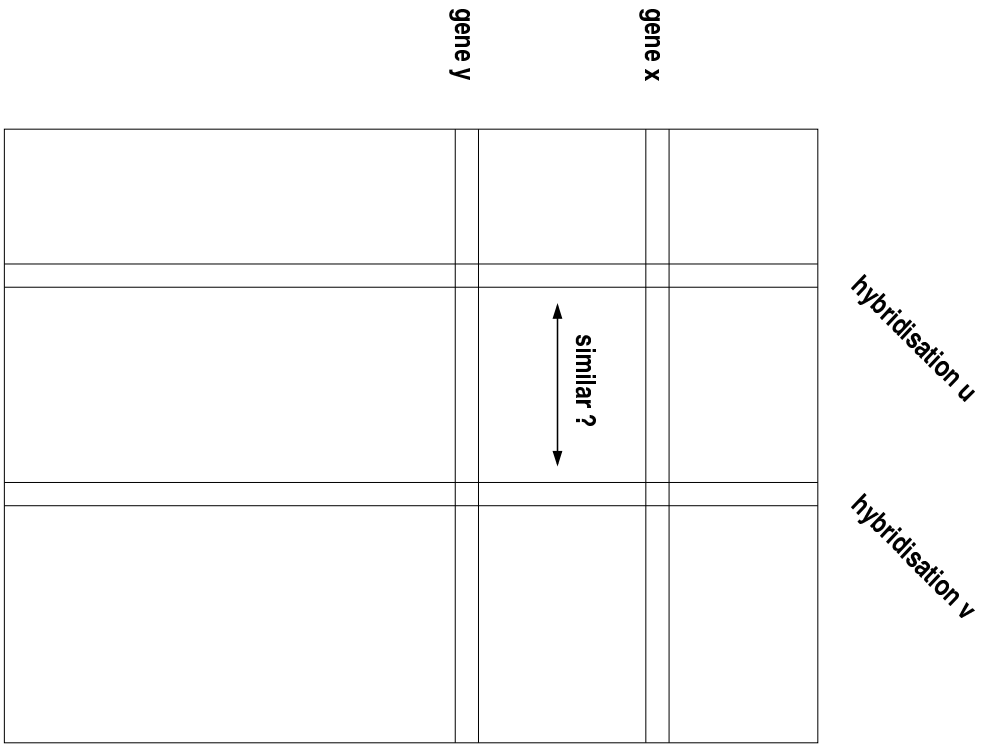
Given a thoroughly preprocessed data set one expects to be ready to tackle the biological questions of data interpretation. However, filtering the genes by applying the above constraints still results in large amounts of data. Typically, the so-to-speak "purified" table of signal intensities comprises several hundred to several thousand genes. Some of them show similar transcription profiles.



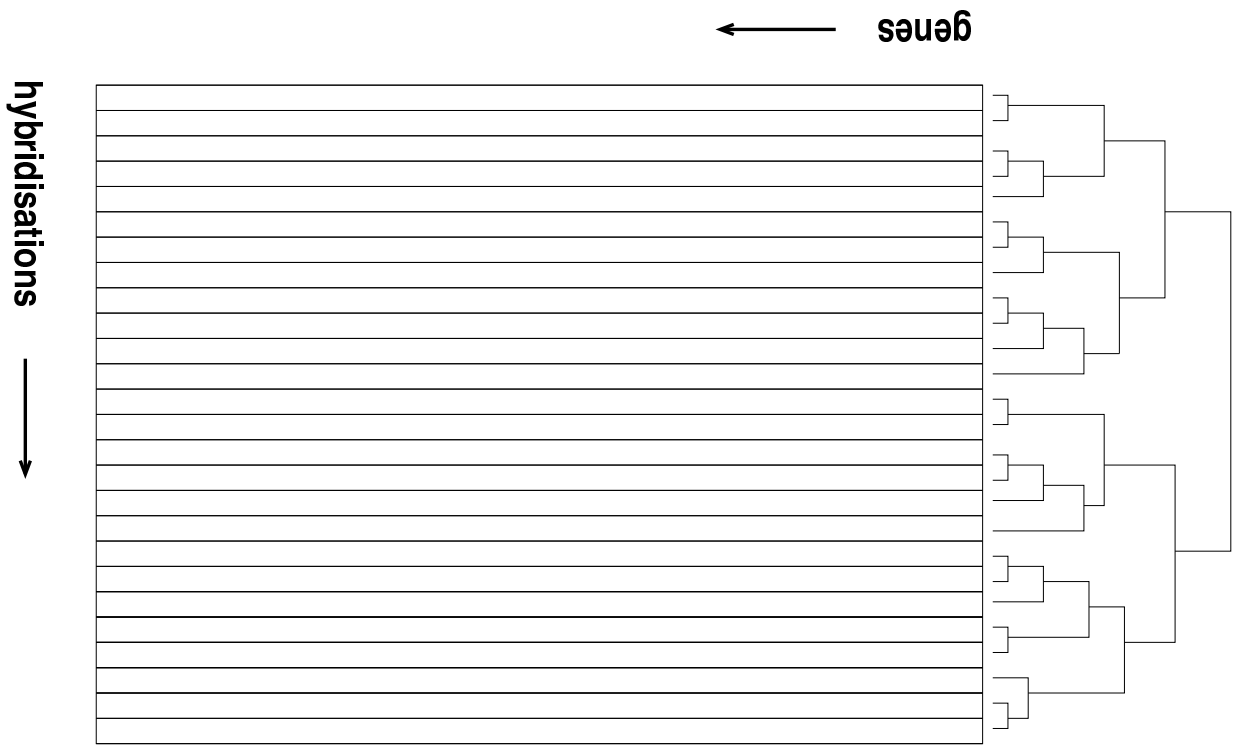
The degree of similarity may be measured by different metrics. Absolute or relative similarity of the values may be taken into account. There may be genes showing very similar values in some hybridizations while being differential in others.



A method frequently used for microarray data analysis is hierarchical clustering. It will result in a tree of genes. It can be cut at different levels resulting in different amounts of clusters.



The problem is symmetrical: The hybridizations show profiles across the genes ...



and may be hierarchically clustered as well.

Methods for the Analysis of Multivariate Array Data

Method	Output
○ Hierarchical clustering Eisen et al. (1998), PNAS 95: 14863	Tree
○ k-means clustering S. Tavazoie et al. (1999), Nat. Genet. 22: 281-285	Set of clusters
○ Learning Networks	
Neural Networks, e.g. Kohonen Maps Tamayo et al. (1999), PNAS 96: 2907	Set of clusters
Baysian Networks Friedman et al. (2000), Proc. RECOMB 2000: 127	Directed graph
○ Planar embedding	
Multidimensional Scaling Khan et al. (1998), Cancer Res. 58: 5009	Biplot
Principal Components Analysis Hilsenbeck et al. (1999), J. Natl. Cancer Inst. 91: 453	
Correspondence Analysis	
○ Other Methods Ben-Dor et al. (1999), J. Comput. Biol. 6: 281	Set of clusters

(incomplete list)

This is an incomplete list of some commonly applied methods. Naming all of the methods recently used for microarray data analysis would result in an outline of applied statistics.

more exploratory

Classification (supervised learning)

Take as input a grouping of objects, aim at delineating characteristic features common and discriminative to the objects in the group ("classifier"). For new objects, the classifier can be used to determine the appropriated group.

Clustering (unsupervised learning)

Which objects appear to be different, which similar? No group affiliations have to be known in advance. However, in practise parameters such as the topology of the map for SOMs or the expected no. of clusters for k-means clustering have to be selected.

Planar projection, also called planar embedding

Most exploratory. Showing how discrete or fuzzy cluster borders are, how well groups of objects are separated.

Most methods recently applied to microarray data fall into one of three groups, namely classification, clustering, or projection methods. Classification methods take as input a grouping of objects and aim at delineating characteristic features common and discriminative to the

Method	Output	Reference
Classification		
Weighted voting	Classifier	[17]
CART	Tree	[7, 11]
Support vector machines	Classifier	[8, 16]
Artificial neural networks (ANN)	Classifier	[21]
k-nearest neighbors	Classifier	[30]
ISIS	Bipartitions	[28]
Bayesian regression	Classifier	[29]
Clustering		
Hierarchical clustering	Tree	[12]
k-means clustering	Set of clusters	[27]
Clustering affinity search technique (CAST)	Set of clusters	[4]
Kohonen maps	Set of clusters	[26]
Cluster identification via connectivity kernels (CLICK)	Set of clusters	[24]
biclustering	Set of clusters	[10]
Gene shaving	Set of clusters	[18]
Planar embedding (projection)	2D- or 3D-projection plot	
Multidimensional scaling		[20]
Principal components analysis		[19]
Singular value decomposition		[2]
Other methods		
REVEAL	Directed graph	[23]
Bayesian networks	Directed graph	[15]

Table 1: **Methods frequently used for microarray data analysis.**

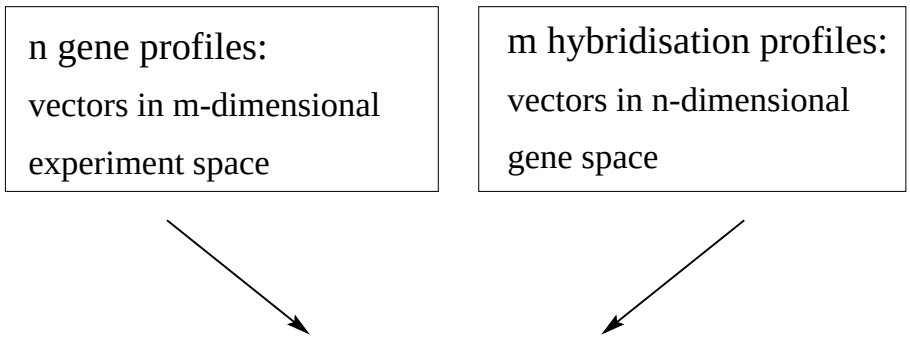
objects in the groups. The characteristic features are referred to as *classifier*. For new objects, the classifier can be used to determine the appropriate group. For cancer research, these objects may consist of different tumor cell lines or of tumor samples of different tumor-types, stage or grade, often supplemented by normal tissue of the particular organ [17]. Examples of classification methods range from linear discriminant analysis [14] to support vector machines [8] or classification and regression trees (CART, [7,11]). Clustering allows investigation of which genes or hybridizations appear to be different, and which transcription profiles appear to be similar. Examples of clustering techniques are k-means clustering [27], hierarchical clustering [12], and self-organizing maps [26]. Clustering tends to be more explorative than classification. No group affiliations have to be known in advance. However, parameters such as the topology of the map for self-organizing maps or the expected number of clusters for k-means clustering have to be selected. Varying parameters may result in altered output, and inappropriate parametrization in uninformative results.

Projection methods produce a low dimensional projection of an originally high dimensional data set. One can, for example, represent genes as numerical vectors with the number of elements of each vector being the number of hybridizations involved. Therefore those vectors could be plotted as points in hybridization dimensional space, if only the number of dimensions were small enough for visualization. Methods such as multidimensional scaling (MDS) [5] or principal component analysis (PCA) [19, 22] as well as the technique described here, project these points into a two or three dimensional subspace so that they can be plotted. Such an embedding attempts to represent objects such that distances among points in the projection resemble their original distances in the high dimensional space as closely as possible. An example of the above mentioned objects is hybridizations as vectors in gene space.

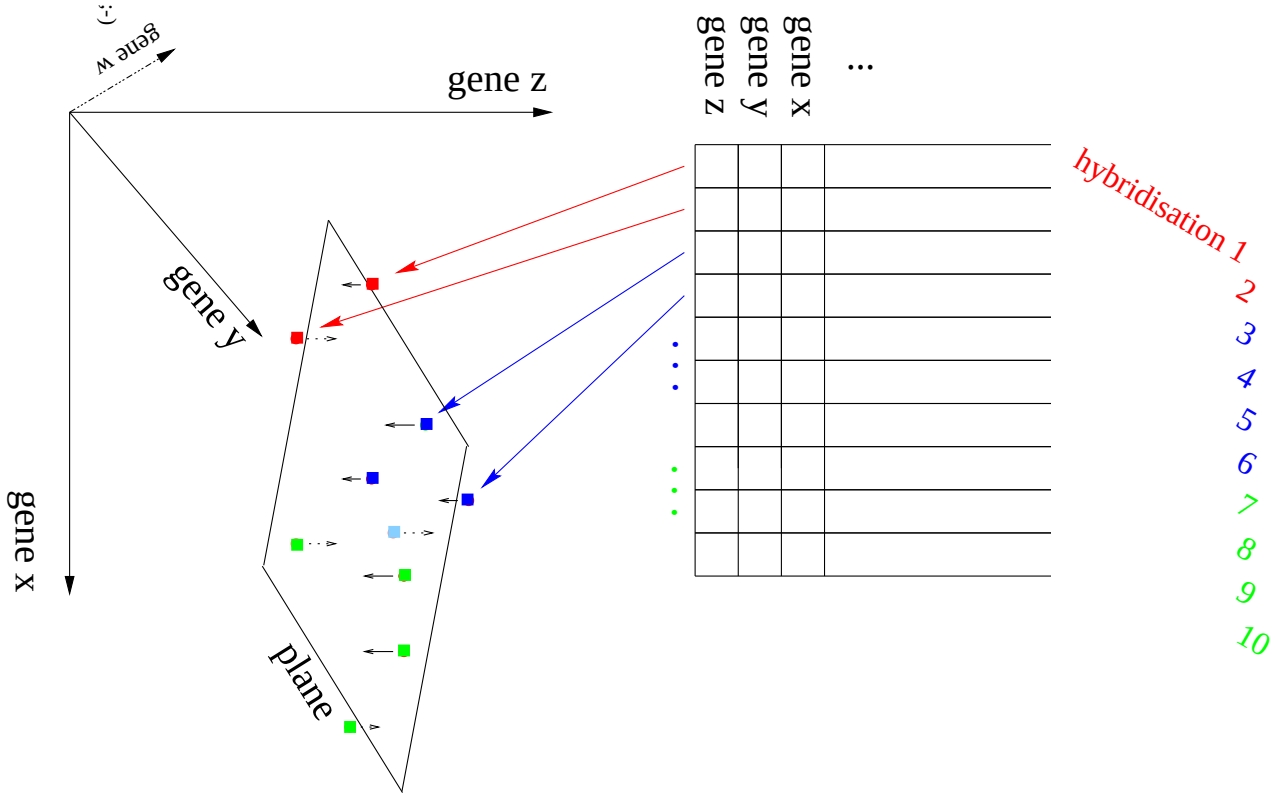
Correspondence Analysis

- Visualises hybridisations and genes at the same time
- Reveals interdepencencies (’correspondence’) between hybridisations and genes
- Exploratory:
  - no parametrisation needed
  - characterises predominant variations among the data points

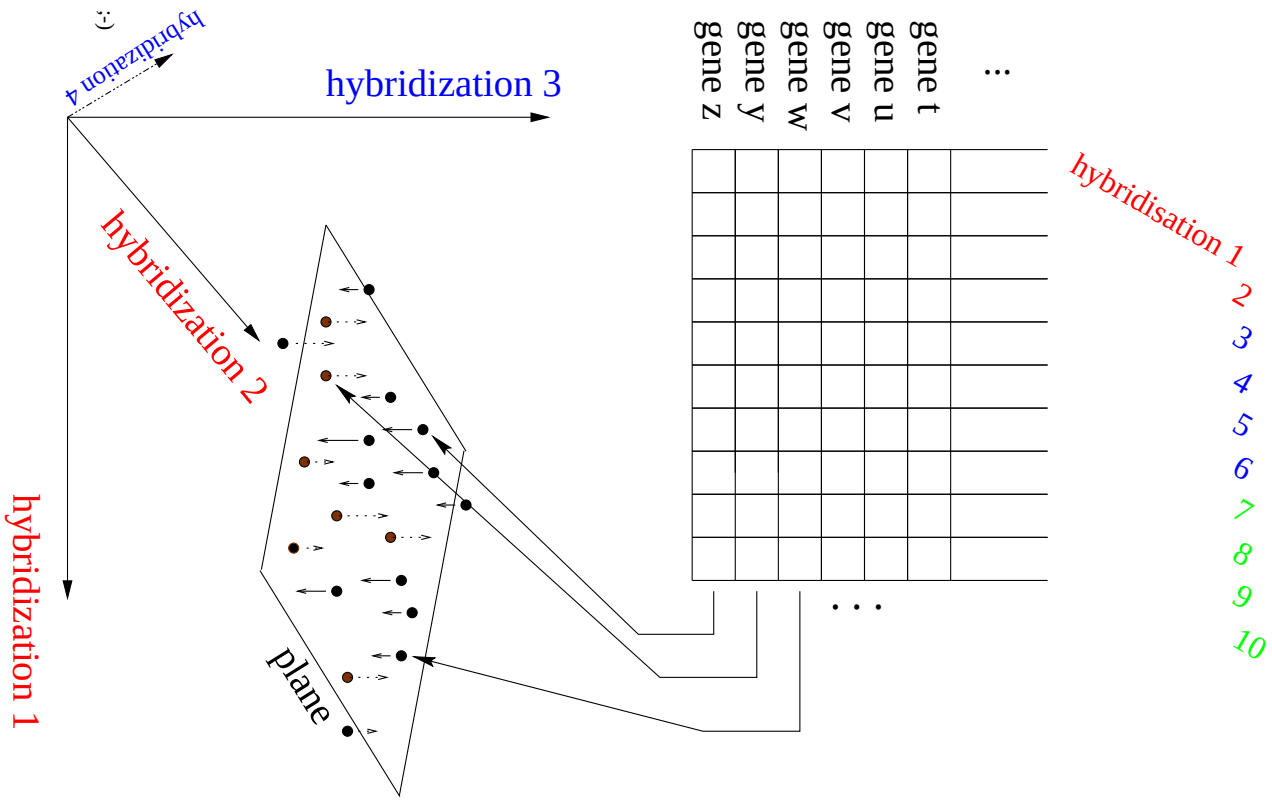
Brief methodology



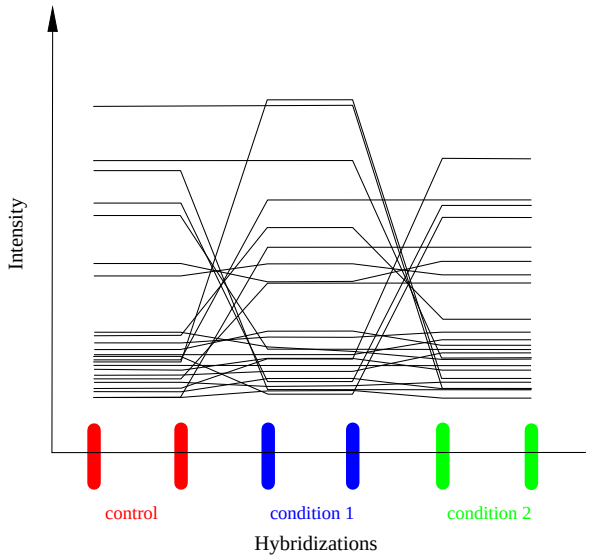
- projection into a common subspace of low dimensionality for visualisation
- conserving point to point distances (total variance) as well as possible (-> explained variance)



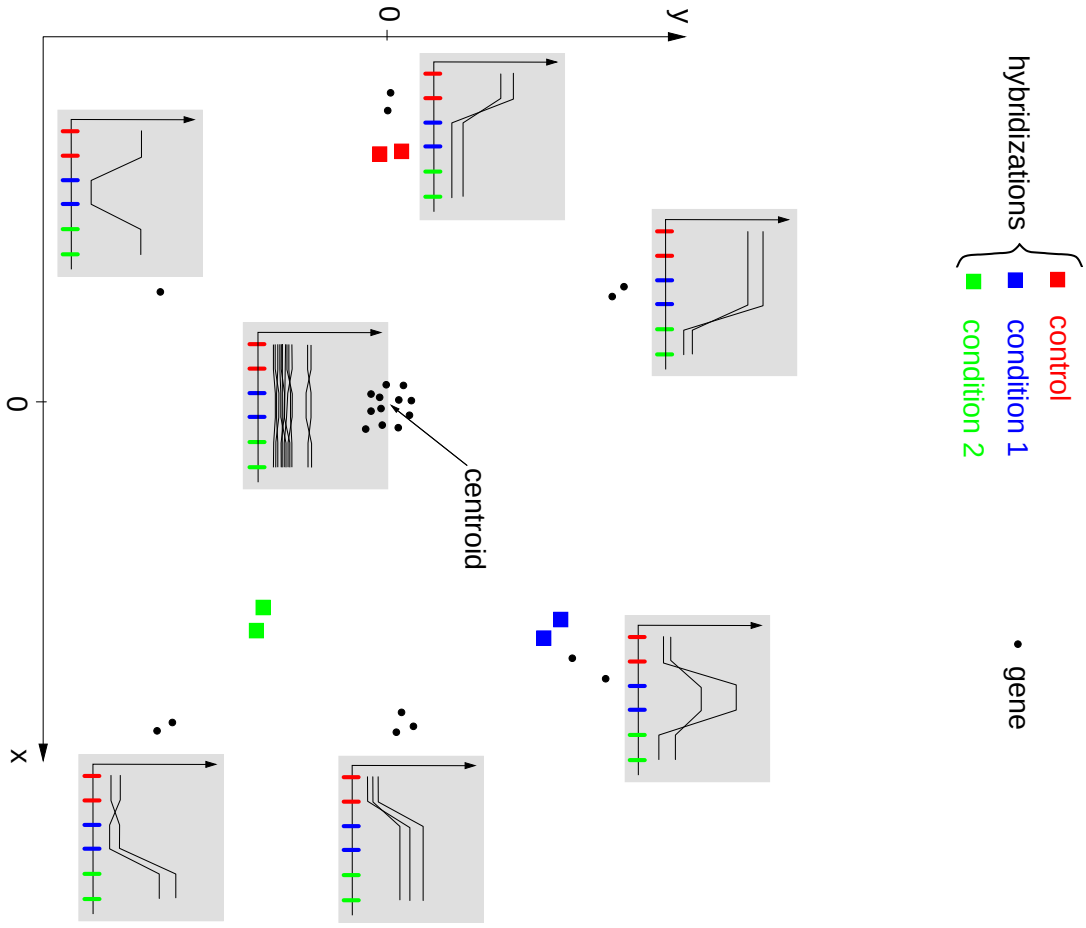
The m columns of a table of n genes $\times$ m hybridizations are represented in n-dimensional gene space (three dimensions are shown). n ranges from a few hundred to tenths of thousands. Most microarrays comprise several thousand elements. A plane is selected such that the distance of the hybridization vectors to the plane is minimal, thus conserving point-to-point distances among these vector points as well as possible.



Vice versa, one can regard the gene vectors in hybridization-dimensional space, projecting the genes as done above for the hybridizations.



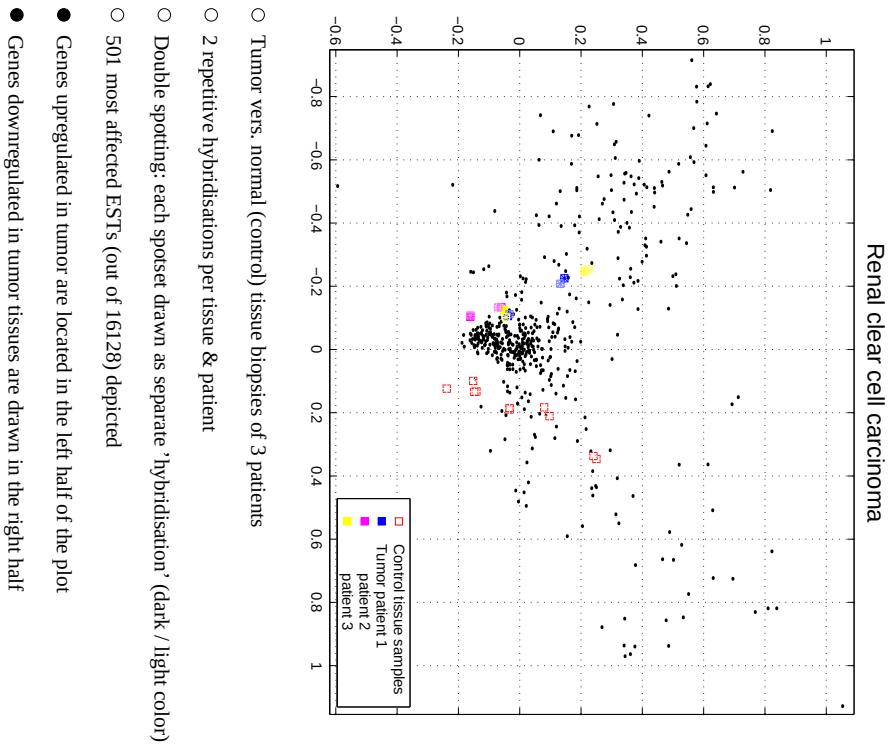
To look into the interpretablility of such a plot, let me introduce a constructed (virtual) data example. It resembles real data in that the majority of the genes is lowly or not transcribed to a measurable amount. It comprises only 24 genes and differs from the real world in perfect reproducibility among the two hybridizations of each experimental condition.



This is the expected output, demonstrating the properties of such projections more clearly than possible by showing a single plot of real data. Gene-clusters are shown together with the according gene profiles. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless. The following properties of such a plot are useful for its interpretation.

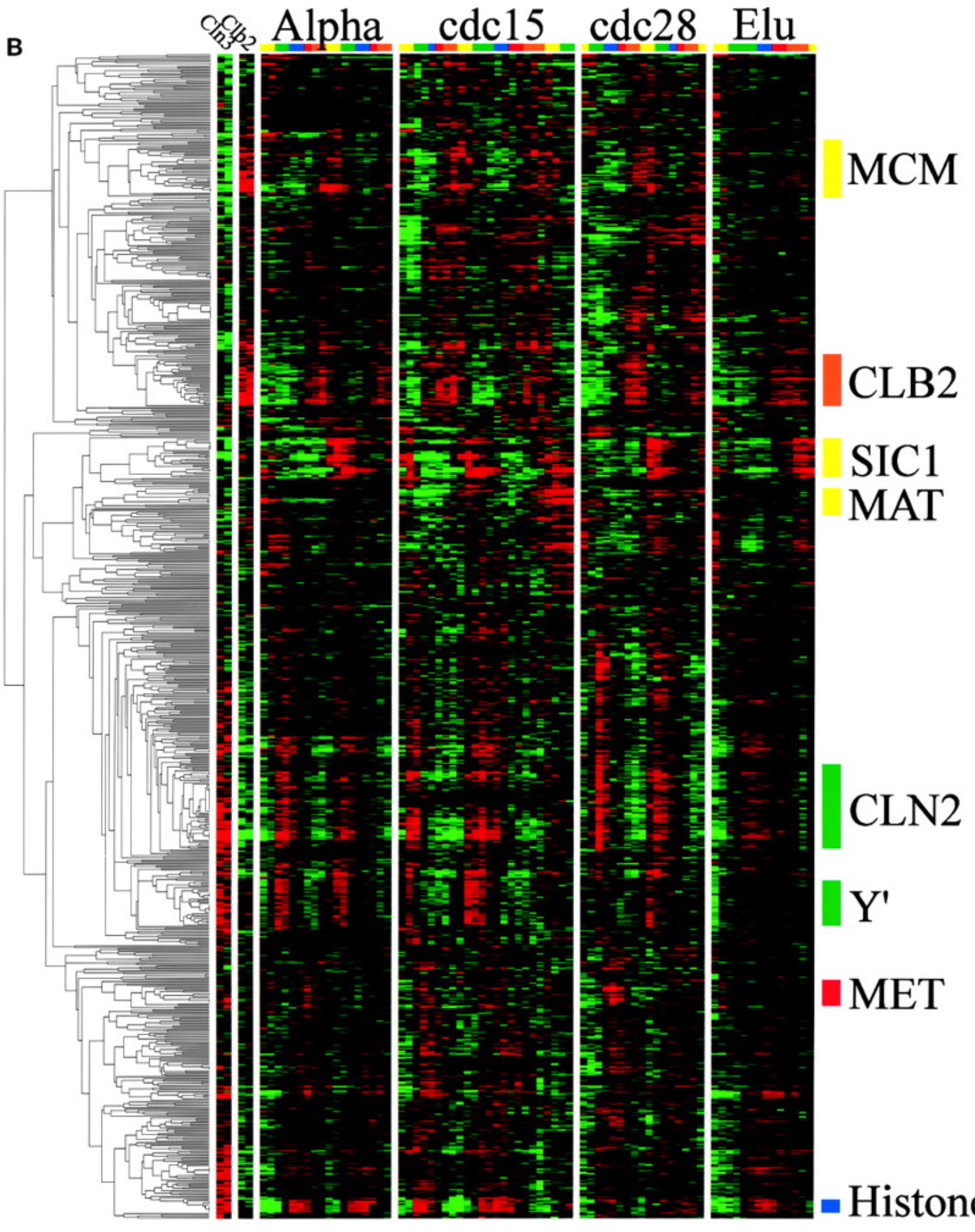
- Hybridizations showing high similarity in expression profile, for example because they belong to the same experimental condition, have a short distance in the 24-dimensional gene space, and therefore they will be neighbors in the projection as well.
- Genes with high intensities in a condition are located in the direction of this condition. The two genes located in the direction of the blue condition (upper right corner) are both upregulated particularly in the blue condition.
- Genes particularly downregulated under this condition are located at the opposite side of the centroid. One can regard this gene (lower left corner) as being downregulated in the blue condition. Another valid interpretation is, that it is located in the direction of the bisection line between the red and the green condition because it is equally abundant in these two conditions.

- All genes with unchanged expression, or those not expressed to a measurable amount in any of the conditions under study are located near the centroid. For experiments with comprehensive or complete gene sets, i.e. sets not particularly selected for high expression, the genes that are not detectable will be the majority. The CA plot will show a centric cloud of many genes lacking significantly changed expression throughout the experiment. The outer regions of the plot will contain the so-called 'differential' genes. Their distance to the centroid will reflect the significance of displaying differing expression from the 'average' ones in terms of χ^2 - statistics, which are placed at the center of the plot.



Now knowing at least roughly how to interpret such a plot, let's have a look on a simple real data example. Biopsies of both tumor tissue - the tumors being renal clear cell carcinoma - and normal tissue of the same patient (as a control) have been sampled and hybridised, imaged, normalized, quality filtered and projected.

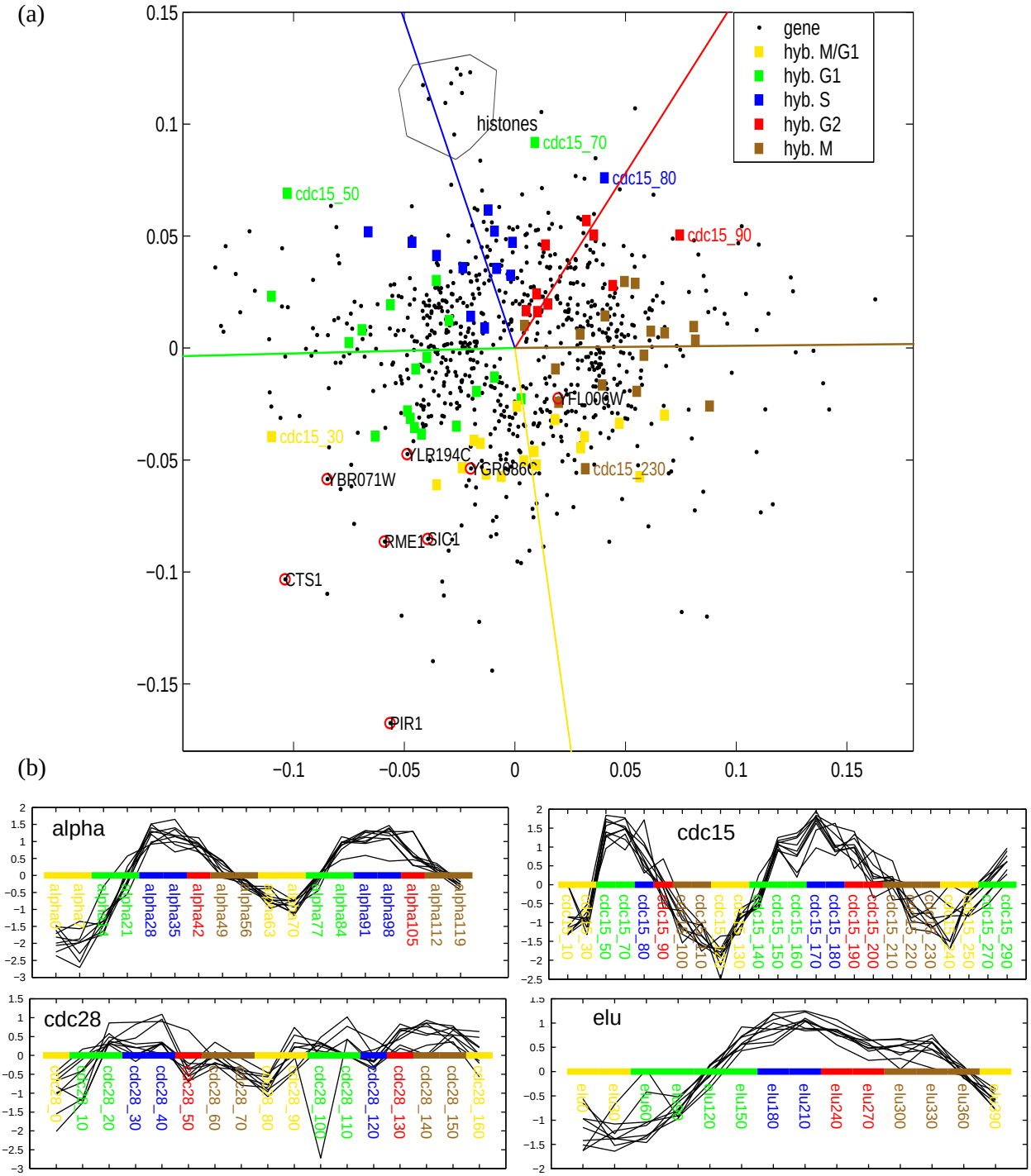
The tumor tissues on the left half of the plot - tagged by three colors for three different patients - are separated from the normal tissue samples on the right. And all the genes upregulated in the tumor are located on the left side, all those downregulated in the tumor we find in the right half of the plot.


B
Clb2
Cln3
Alpha
cdc15
cdc28
Elu
MCM
CLB2
SIC1
MAT
CLN2
Y'
MET
Histone

To introduce the method, its performance is demonstrated on a well-known data set. This set comprises the hybridizations referred to by Spellman *et al.* which are publicly available[1]. Spellman *et al* arrested the *S. cerevisiae* cell cycle by four different methods, namely α factor-, *CDC15*- and *CDC28*-based blocking, and elutrition. At certain timepoints after releasing the block, samples from each of the methods had been drawn, their cell cycle phase had been classified and the transcriptional status assayed by microarray hybridization.

The transperancy has been reproduced from P. T. Spellman, G. Sherlock, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell, 9:3273-3297, 1998. It shows gene expression during the yeast cell cycle. Genes correspond to rows, and the time points of each experiment are the columns. The ratio of induction/repression is shown for each gene such that the magnitude is indicated by the intensity of the colors displayed. If the color is black, then the ratio of control to experimental cDNA is equal to 1, whereas the brightest colors (red and green) represent a ratio of 2.8:1. Ratios >2.8 are displayed as the brightest color as well. In all cases red indicates an increase in mRNA abundance, whereas green indicates a decrease in abundance compared to the control samples (stemming from asynchronous cultures of the same cells growing exponentially at the same temperature in the same medium). Gray areas (when visible) indicate absent data or data of low quality. Color bars on the right indicate the phase group to which a gene belongs (M/G1, yellow; G1, green; S, purple; G2, red; M, orange). These same colors indicate cell cycle phase along the top. Genes that share similar expression profiles are grouped. The dendrogram on the left shows the structure of the cluster.

[1] http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt



The planar embedding (of exactly the same data) produced by CA shows the hybridizations clearly separated according to their cell cycle phase. They are arranged in circular order of correct sequence. The lines denoting the direction of the hybridization medians emphasize this arrangement. The black dots correspond to genes. Genes that show strong expression in a certain phase are located in the direction determined by the hybridizations of this phase. The farther away from the center the genes are, the more pronounced is their association with that phase. Genes that are down-regulated in this phase appear on the opposite site of

the centroid. As an example of strong association with the S-phase, the gene profiles for the histone gene cluster, also marked by Spellman *et al.*, are encircled in black. Their profiles are shown below (b) which is further subdivided according to the method of cell cycle arrest that had been used. The red-encircled genes will be discussed below in the context of *CDC14* induction. Genes equally transcribed in most or all of the cell cycle states had been removed by Spellman *et al.*, causing a hole near the centroid of the CA plot where otherwise genes would lie that show little change.

Upon close inspection the biplot reveals interesting details about the data. It should be noticed that hybridization cdc15_30 (cdc15-based blocking, 30 min timepoint) classified as M/G1 (yellow) lies in the green (classified G1) sector rather than in the yellow one. Likewise, hybridization cdc15_70 is classified G1 but clusters together with the blue dots (S-phase), and one S-phase hybridization, cdc15_80, lies in the red sector of G2 hybridizations. All these outliers come from the series of hybridizations where the cell cycle arrest was achieved using *CDC15*-based blocking. This arrangement of cdc15 hybridizations suggests an improper phase classification for these samples.

This hypothesis can be validated based on the gene profiles. For the histones, the shift towards an earlier stage in cell cycle is visible in the upper right panel (b). Timepoints cdc15_30 through cdc15_90 show the upregulation of the histones already at the end of M/G1 (yellow) instead of G1 (green) as well as too early downregulation: the curves intersect the zero line (identity to the control channel) at cdc15_90, classified as G2 (red) instead of M (brown), as e.g. in the elutrition experiment. The nine histones are only a small subset of the 800 cell-cycle regulated genes. Profiles of other genes, though different from the ones plotted, also display shifting of the above timepoints to expression patterns associated to an earlier state in cell-cycle by the remaining timepoints (data not shown). CA computes the projection for timepoints cdc15_30 to cdc15_90 according to their expression patterns in the entirety of the geneset, independent of their phase classification. The CA plot displays them displaced in clockwise shift compared to equally colored squares, that is in positions inconsistent with their cell-cycle state classification. While clustering together the nine histone genes, the original figure by Spellman *et al.* does not properly show this shift.

- original data: $\mathbf{N}$ ($I \times J$ matrix) with elements n_{ij}

$$n_{+j} - \sum_{i=1}^{I} n_{ij};\ n_{i+} - \sum_{j=1}^{J} n_{ij};\ n_{++} - \sum_{i=1,j=1}^{I\ J} n_{ij}$$

- row weights: $r_i = n_{i+}/n_{++}$
- column weights: $c_j = n_{+j}/n_{++}$
- correspondence matrix $\mathbf{P}$: $p_{ij} = n_{ij}/n_{++}$

- χ^2 matrix $\mathbf{S}$, $s_{ij} = \frac{(p_{ij} - r_i c_j)}{\sqrt{r_i c_j}}$
- singular value decomposition: $\mathbf{S}$=$\mathbf{U}\Lambda\mathbf{V}^T$, diag. matrix Λ contains singular values λ_k of $\mathbf{S}$

- **principal coordinates:**

for gene i (principle axis $k = 1, ..., J$) $f_{ik} = \frac{\lambda_k u_{ik}}{\sqrt{r_i}}$

for hybridization j (in the same space) $g_{jk} = \frac{\lambda_k v_{jk}}{\sqrt{c_j}}$

- **standard coordinates:**

for gene i $f'_{ik} = u_{ik}/\sqrt{r_i}$

for hybridization j $g_{jk} = v_{jk}/\sqrt{c_j}$

- **HMS:**

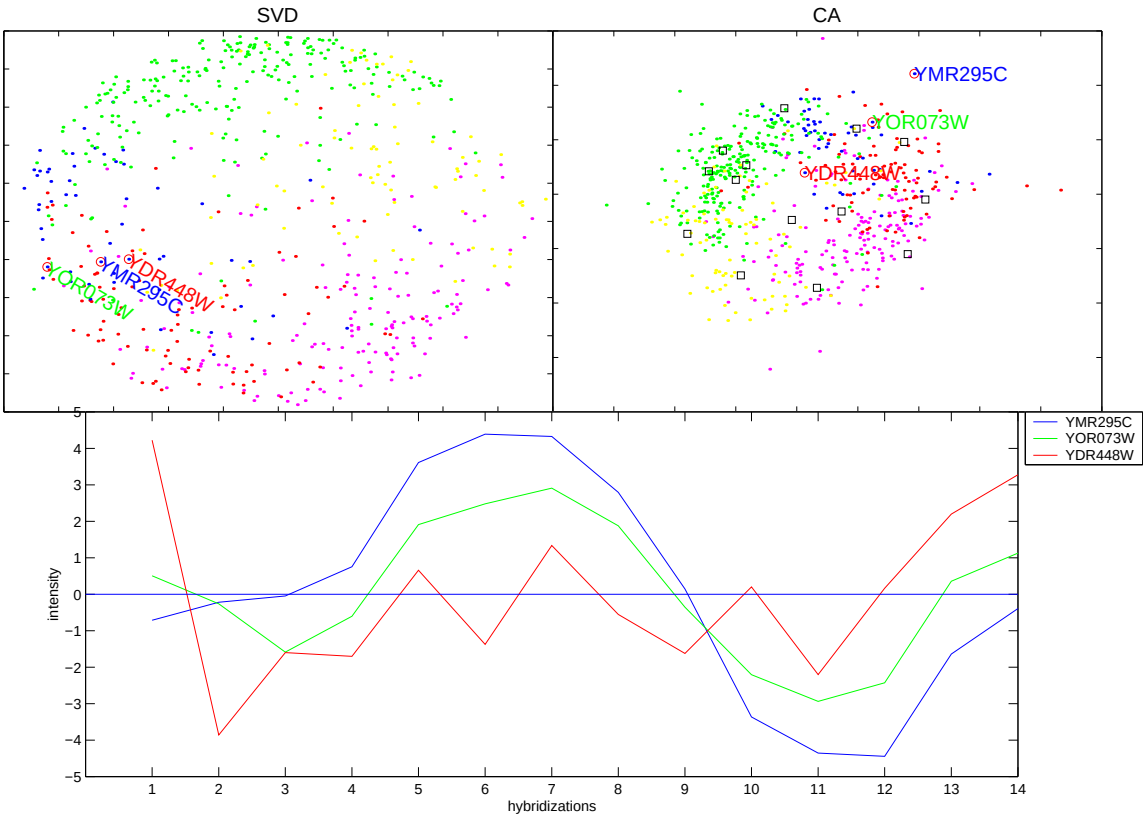
Let $\mathbf{N}$ contain the hybridization medians of original data matrix $\mathrm{N}^*$of elements n^*_{ij}

$\mathbf{N}$ is submitted to CA.

Let $\mathbf{P}^*$ have elements p^*_{ij}, $p^*_{ij} = n_{ij}/n^*_{++}$, then the principle coordinates for the supplementary hybridizations from correspondence matrix $\mathbf{P}^*$ are

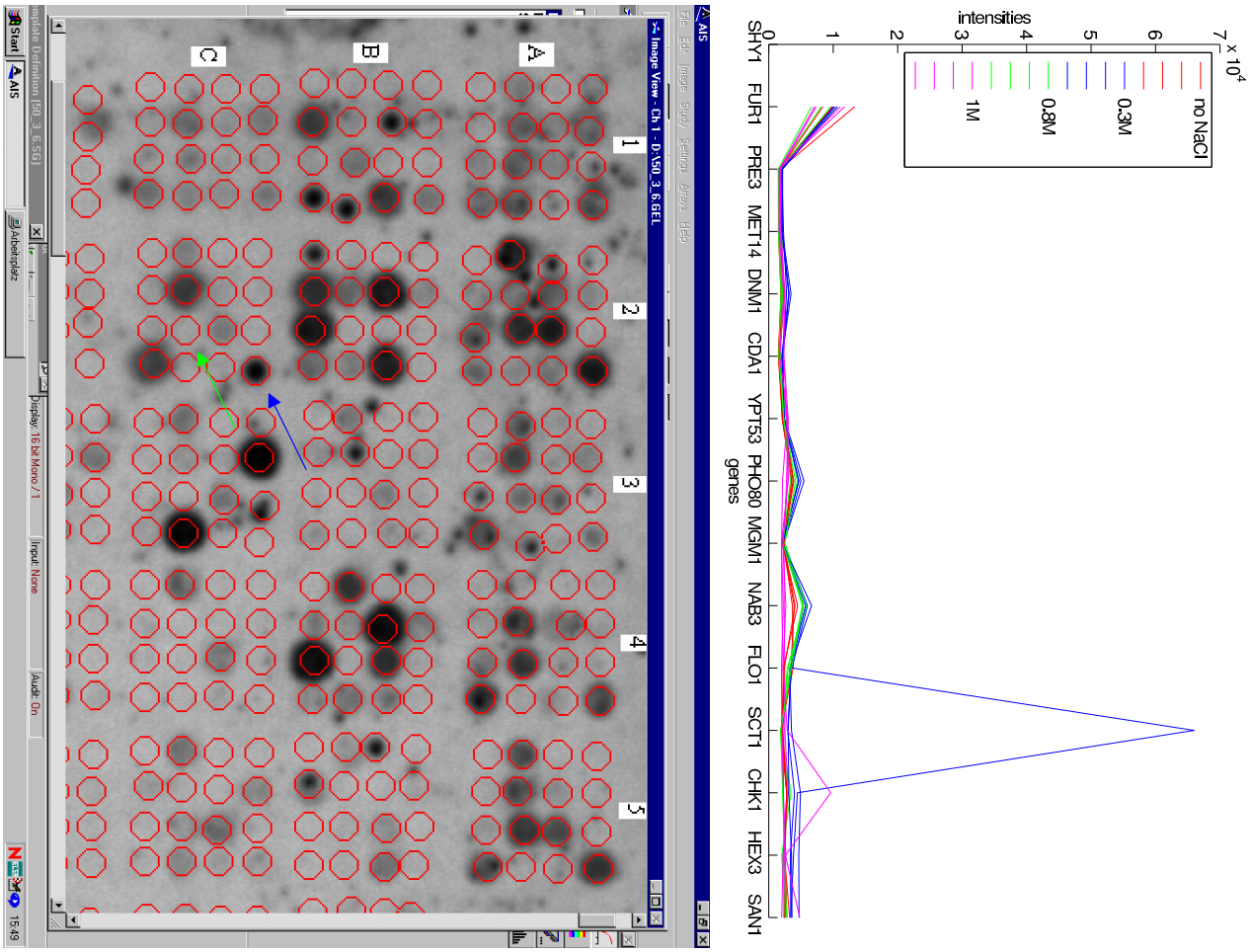
$$g^\star_{jk} = \frac{1}{\sum_i p^\star_{ij} f_{ik}} \frac{\sum_i p^\star_{ij} f_{ik}}{\lambda_k}.$$

Above two transperancies show all information needed to implement a simple CA algorithm. It can be easily done by using nested *for* loops. A much shorter implementation without loops can be achieved in any programming language supporting matrix multiplication and providing a routine for singular value decomposition, e.g. in MATLAB (see `http://www.ub.uni-koeln.de/ediss/archiv/2002/11w1296.ps`, Appendix B).



Alter *et al.* (O. Alter, P. O. Brown, and D. Bostein. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. U.S.A., 97:10101-10106, 2000.) successfully applied singular value decomposition to the analysis of the same data set. In CA plots, the distance of a given gene from the centroid represents the strength of its association with a hybridization lying in the same direction and vice versa. A direct comparison with phase and radius in the visualization of Alter *et al.*[2] shows that this is not necessarily the case in the singular value decomposition alone.

[2]as given e.g. at http://genome-www.stanford.edu/SVD/PNAS/Datasets/Sort_Elutriation.txt



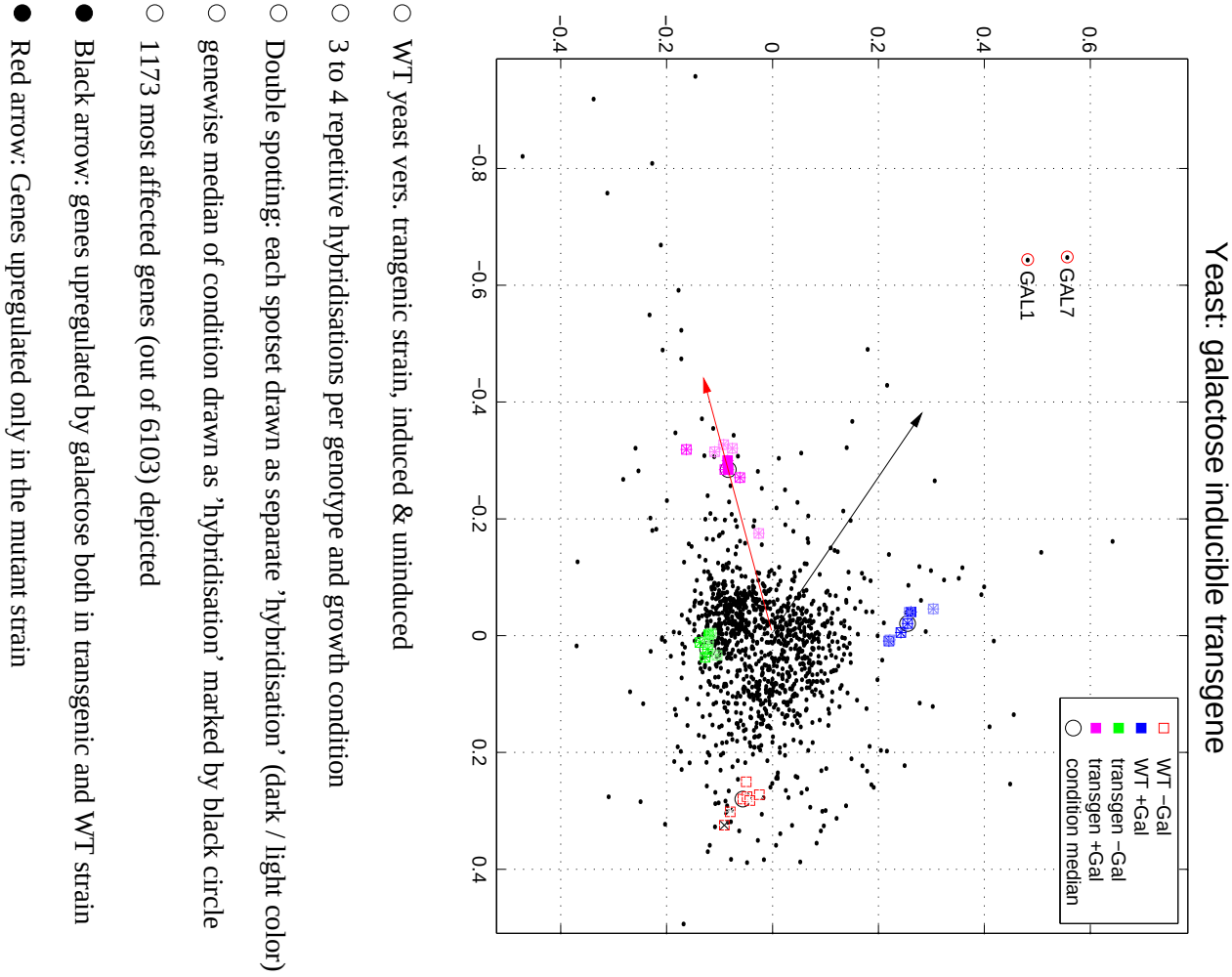
As shown above, it might be useful to visualize deviations of outlying measurements from the expected state. However, data sets frequently comprise severe outliers such as this one.

Wild type yeast was exposed to different concentrations of sodium chloride in the medium (see legend). Normalized transcription intensities of 14 genes are shown in a parallel coordinates plot, lines representing measurements and being color coded according to their particular experimental condition. The plot presents a typical subset of genes, representative with regard to the high number of genes not expressed to a measurable amount. Whereas the different conditions are reproducibly measured for most genes, SCT1 shows one far outlying signal for 0.3M NaCl (blue), which in this case is due to agglutinated label. In the corresponding image (below), the bright dots of unspecifically bound label are common to radioactively labeled targets, whereas the most severe outliers among multichannel data are frequently caused by highly flourescent dust (not shown). The ordinate shows arbitrary (machine dependent) intensity units.

For this reason, a thorough preprocessing is essential. Different normalization algorithms are applied to single and multichannel data for the different meaning of the particular raw intensities. Intensity-, ratio-, and reproducibility filters are applied to extract genes of marked expression for both types of data.

Genes with generally low reproducibility for most of the conditions under study are filtered out by the reproducibility filter. However, with increasing numbers of conditions, discarding all genes with low reproducibility in one of the conditions will leave no gene undiscarded. The same is true for the intensity filter. It is therefore reasonable to use these filters to discard only genes with low abundance or low reproducibility (often coinciding) in all the conditions under study. Thus, outliers as shown above have to be handled by other measures. Otherwise, they would seriously interfere with CA analysis, which in contrast to other methods is not similarity-driven but aims at displaying variance. Any difference to the default state (expected value) such as an outlier, will be regarded as important for the projection. The larger the difference, the more distinctly the corresponding point will be plotted.

We prevent this by choosing the principal axes according to the condition medians only (HMS).

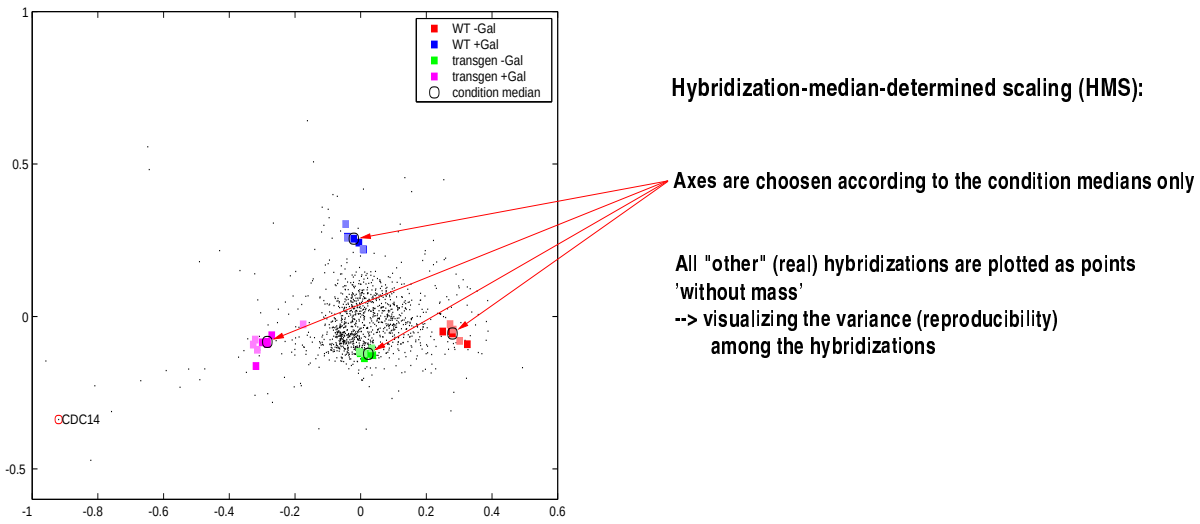


Let me first introduce a new data example. Here, a transgene was transfected into yeast cells under the control of a galactose inducible promoter:

Red empty boxes are WT yeast without galactose in the medium, blue WT with galactose, green the transgenic strain without galactose and pink transgenic with galactose and here we would expect the transgene to be induced and genes that follow the transgene. The aim was

to separate those genes from all the genes which are upregulated by galactose in yeast anyway, that is: also in the WT strain.

The bisection line between WT and transgenic strain with galactose (black arrow) points to the genes induced by galactose both in the transgenic and in the WT strain. There we find genes like Gal 7 and Gal 1. The red arrow points to the genes upregulated specifically only in the transgenic strain - those are the genes the experimenter intended to look at with this experiment.



Typically, replicate hybridizations are performed for each condition under study leading to several values for one gene/condition pair. The number of such repeated hybridizations is often small. I therefore represent these values by their gene-wise median rather than their gene-wise average because the median is less sensitive to outliers. The need remains, though, to visualize also the original data and not only the median since they contain valuable information about experimental variance and quality of individual hybridizations. In fact, CA offers the possibility to reflect both aspects. To this end, CA is first effected by using the gene-wise medians, determining the coordinate system to embed the original hybridization intensities. These data points are then referred to as supplementary points or points without mass. Thus the share of noise belonging to an experimental condition is shown by the spread of its hybridizations around the median. As the dimensions of the data are reduced by using medians of hybridizations per experimental condition, I refer to this strategy as hybridization-median determined scaling (HMS).

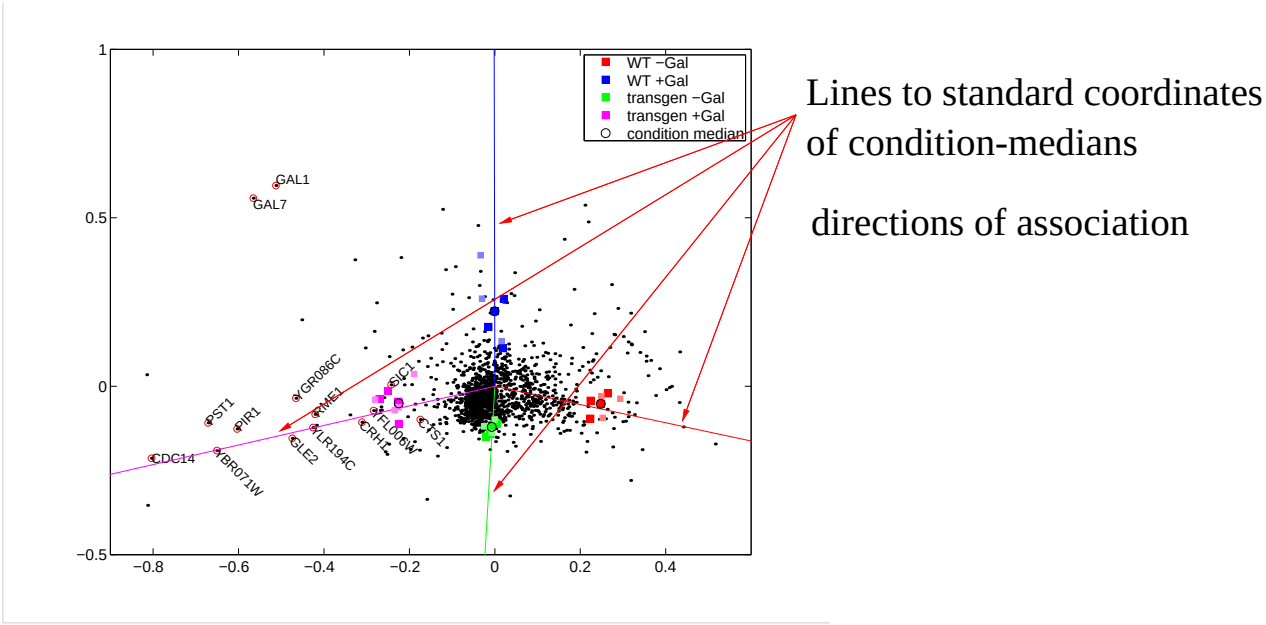
The embedding for hybridizations without mass is computed as follows. Let the matrix $\mathbf{N}$ contain only the hybridization medians and let $\mathbf{N}^\star$ of elements $n^\star_{ij'}$ be the original data matrix containing all the hybridizations. $\mathbf{N}$ is submitted to CA. Let $\mathbf{P}^\star$ have elements $p^\star_{ij'} = n^\star_{ij'}/n^\star_{++}$. The principal coordinates for the supplementary hybridizations from correspondence matrix

$\mathbf{P}^\star$ are then calculated as

$$g^\star_{j'k} = \frac{1}{\sum_i p^\star_{ij'}} \sum_i \frac{p^\star_{ij'} f_{ik}}{\lambda_k}.$$

In our own data sets, a single hybridization consists of two corresponding spot sets because each cDNA had been spotted twice on the array. I refer to these spot sets as *primary* and *secondary spots*. They tend to show a higher correlation than hybridizations belonging to the same experimental condition. Plotting them separately (duplicating the number of supplementary points) provides an atomic unit of distance in the biplot, where no units are assigned to the axes.

Projection methods generally aim at explaining the major trends in the data while at the same time ignoring minor fluctuations. HMS has been demonstrated to further enhance this effect (Fellenberg et al., attached).



Due to all these precautions and given a sufficient number of repeated hybridizations, the variance explained by a CA plot will largely reflect biological changes, displaying the significance of differences both among the genes and among the hybridizations in terms of the χ^2-statistic. The power of the CA technique however is that it is able to show associations between genes and hybridizations. To fully exploit this property, it is necessary to examine the exact directions of gene-association with the experimental conditions. These are given by the standard coordinates of the according condition medians rather than by their principle coordinates. An experiments represented in standard coordinates can be viewed as a virtual gene having its entire mass (intensity) in this particular experiment. Thus, it is the gene of highest possible association with this experiment, able to "represent" the experiment in "gene-space". Plotting the standard coordinates directly would cause all principle coordinates to shrink into a small area in the middle of the plot. The introduction of lines representing the standard coordinates

is of great help in the interpretation of the plots, relating genes and conditions to each other and circumventing direct plotting.

Adapting CA to microarray data

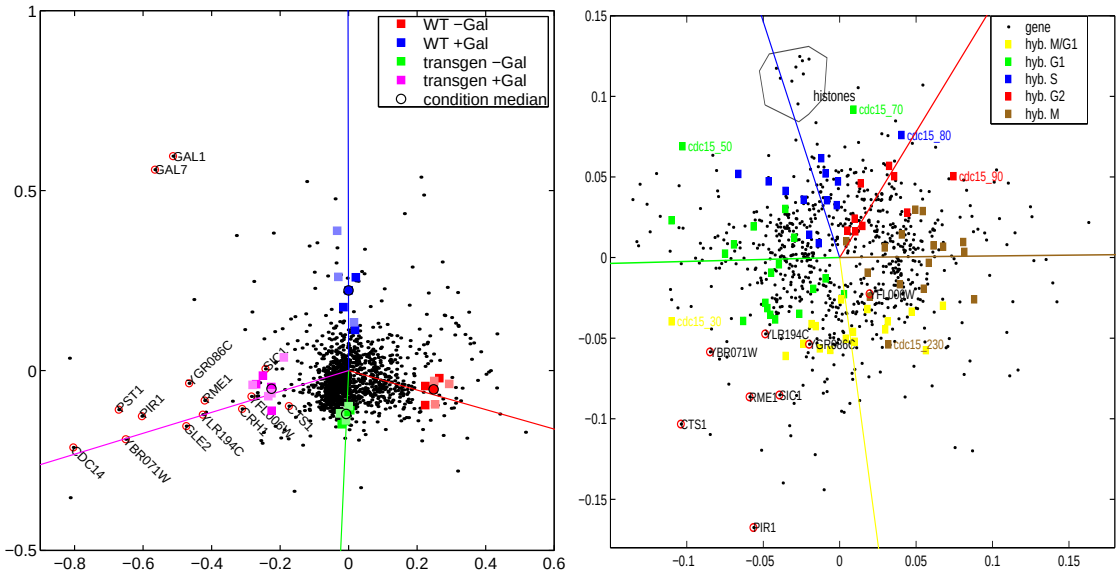
particular requirements	solutions
hybridizations show differential background, label-incorporation, ...	log-linear regression
ratios of low-level intensities are not reliable	→ stabilize by additive shift → intensity filtering
noise, systematic errors	→ reproducibility filtering
low numbers of repeated measurements	→ adapted reproducibility measure → visualization of all measurements
frequently occuring outliers	→ choosing axes according to medians of each condition (HMS)

It is equally important to tackle the problem of unduly high ratios in the low intensity region. As already mentioned, only those genes are filtered out that are low in every condition under study. To lower the impact of low intensities on the intensity ratios, the normalization method described in

- T. Beissbarth, et al. Processing and quolity control of DNA array hybridization data. Bioinformatics, 16:1014-1022,2000.

has been modified, additively shifting the normalized matrix back to its original expression level. To exemplify the benefit of simply adding a certain number to all of the values, consider that a shift from 0.02 to 0.04 resembles upregulation by factor 2, whereas a change from 1000.02 to 1000.04 does not.

Above transperancy summarizes all the discussed measures adapting CA to the particular requirements of microarray data.



- SIC1, known to be accumulated in a Cdc14p dependent fashion [1]
- CTS1, belongs to the cluster of SIC1 co-regulated genes [2]
- RME1, CRH1, PST1 known to be cell-cycle regulated with peaks in mitosis/G1 transition, G1 or late G1, respectively but have not yet been described in association with Cdc14p activity.
- YBR071W, PIR1, YGR086C, YLR194C, YFL006 not yet annotated to be cell-cycle regulated but in agreement with the data of ref. 2 (mitosis/G1 transition)
- GLE2 (nuclear pore protein): ? unknown function in Cdc14p activation context

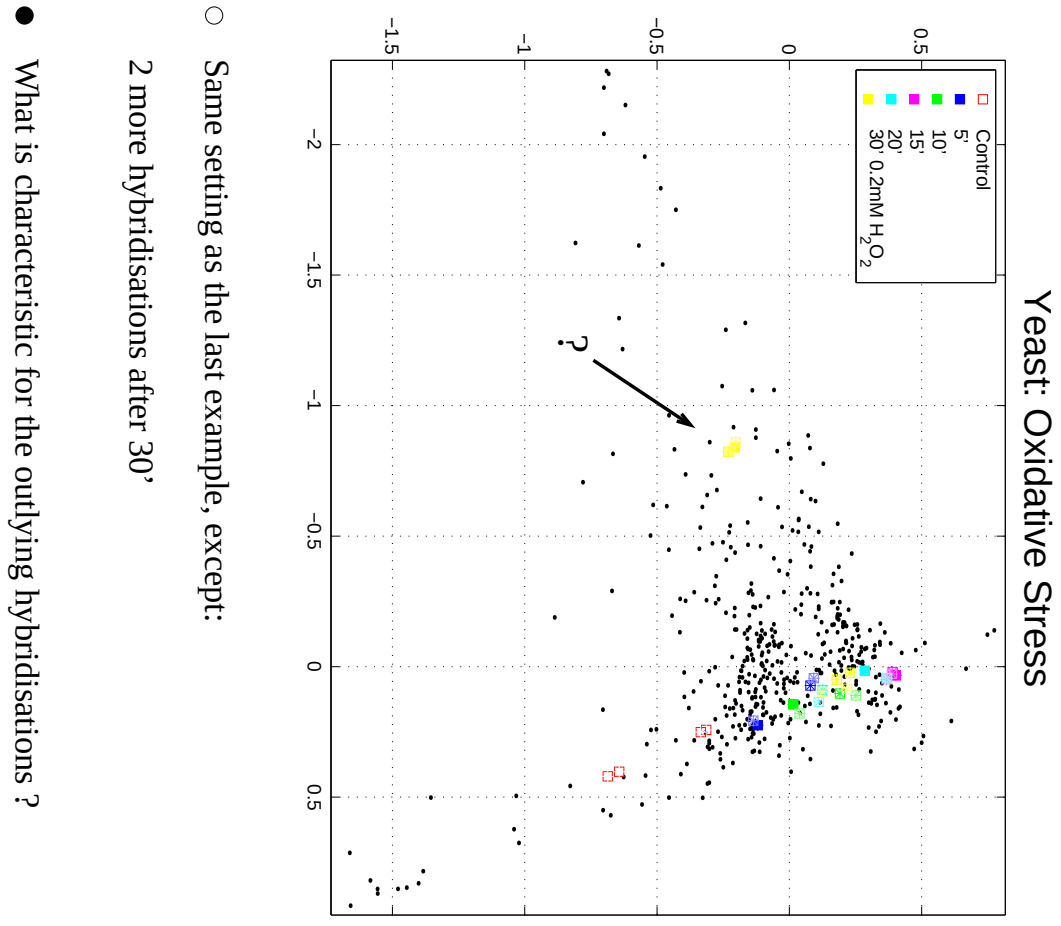
[1] D. O. Morgan. Regulation of the APC and the exit from mitosis. Nat. Cell Biol., 1(2):E47-53, 1999.
[2] P. T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast [...]. Mol. Biol. Cell, 9:3273-3297, 1998.

Some Biology: Genes in the direction of galactose induced transgenic yeast are those specifically upregulated upon *CDC14* induction as opposed to genes activated by galactose also in the WT strain, like *GAL1* and *GAL7*. This subtraction has been achieved purely computationally and is based on the provision of galactose activated genes in wild type as a separate condition. The set of genes associated specifically to the Cdc14p overproducing condition comprises *CDC14* itself as well as *SIC1*, known to be accumulated in a Cdc14p dependent fashion [1] and *CTS1* which belongs to the cluster of *SIC1* co-regulated genes [2]. *RME1*, *CRH1* and *PST1* are known to be cell cycle regulated with peaks in mitosis/G1 transition, G1 or late G1, respectively but have not yet been described in association with Cdc14p activity. *YBR071W*, *PIR1*, *YGR086C*, *YLR194C*, and *YFL006W* have not been annotated to be cell cycle regulated, but these results show that they are. This is in agreement with the data of Spellman *et al.* (right panel, genes marked by red circles), which also show these genes to be transcribed during mitosis/G1 transition. The role of the nuclear pore protein *GLE2* in a Cdc14p activation context remains unclear.

The biological context of CDC14 is sketched in the following transperancy.

The last data example relates to the second part of the talk. A time course has been recorded for wild type *S. cerevisiae* cells under oxidative stress. The thin black arrow draws the chronological progression of the experiment. The cells responding to 0.2mM hydrogenperoxide in their medium show quite a leap in expression behaviour between 15 and 20 minutes that includes the downregulation of genes which had been switched on in the initial phase of the response. Four of those are flagged. Their gene profiles are plotted below. They are switched on initially

and are being downregulated somewhere between 15 and 20 minutes.



In the above example there is obviously something wrong. It is exactly the same experimental setting as before but now the yellow 30' condition is divided into two clusters located far away from each other and distorting the nice picture of the previous plot.

And we want to know why. What is wrong with the outliers? In other words: Can we find features in the experimental description which are characteristic for the outlying cluster? Are there annotation values overrepresented in the cluster? Or are there values missing or underrepresented in the cluster?

Data Warehousing

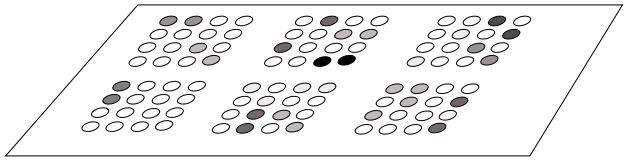
To enable interpretation of large data sets, the data produced need to be stored in a suitable way to allow for global comparison [3]. For rapid and simple access, data should be stored in common format, e.g. in a database, rather than in unequally structured flat files. Database repositories provide the convenience of consistent view, defined interfaces and increased access performance. Build-in methods for multiuser operation as well as a centralized administration enable high standards for data security in addition.

The advantages of standardized storage apply not only to the signal intensities for each item in an array but also to all available descriptions of the sample from which the RNA has been derived, and all details of its treatment.

Several database projects are currently addressing these questions. While ExpressDB (Harvard, [1]) aims at storing data from nearly all available platforms, i.e. cDNA and oligonucleotide chips as well as SAGE (serial analysis of gene expression), a different focus has been to develop systems for consistent description of the samples used and the genes mounted on the array, e.g. in GeneX[3] (NCGR), GEO[4] (NCBI), ArrayDB (NHGRI, [13]), ArrayExpress (EBI, [6]), and RAD[5](UPenn, [25]), the last one combining both objectives.

[3]http://www.ncgr.org/research/genex/
[4]http://www.ncbi.nlm.nih.gov/geo/
[5]http://www.cbil.upenn.edu/RAD2

M-CHIPS Database Charge



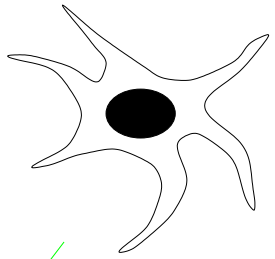
immobilized DNA fragments

Hybridization

labelled cDNA

mRNA

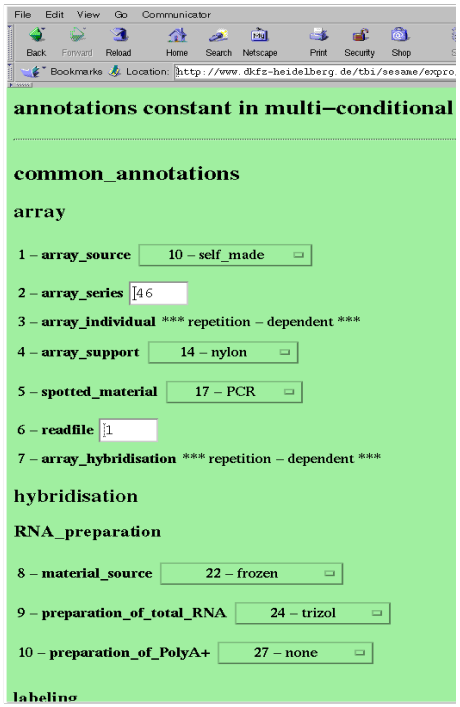
cells growing under specific conditions



hybridization intensities

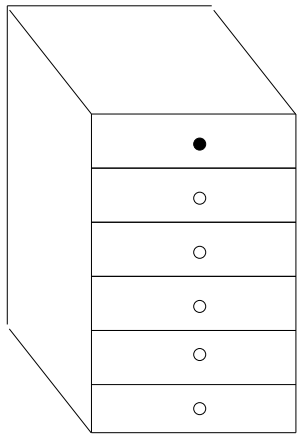
IMAGING

WWW Annotator



33

PostgreSQL Database



experimental annotations

WEB - INTERFACE

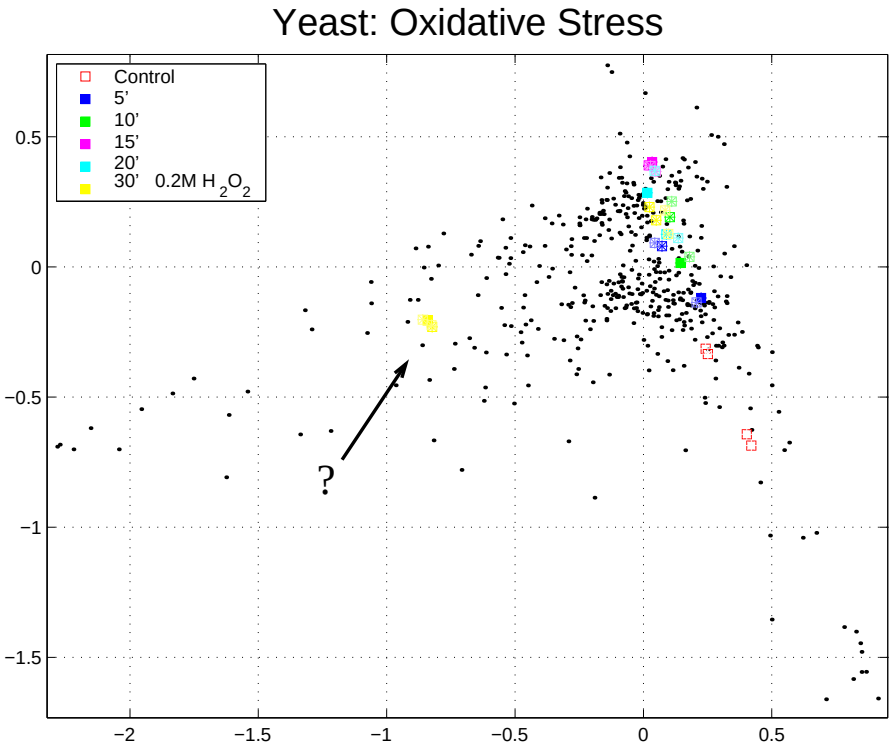
Data upload. Along with the transcription intensities, experiment annotations have to be stored. These should explicitely characterize the sample and its treatment, RNA preparation and labeling steps, hybridization and washing as well as the imaging process in sufficient detail.

Arabidopsis experiment annotation:

see http://www.dkfz-heidelberg.de/tbi/services/mchips/arabidopsis.html

(11 other organisms at
http://www.dkfz-heidelberg.de/tbi/services/mchips/#annos)

Experiment annotations may comprise, among other things, the description of environmental conditions, genotypes, clinical data, type of tissue, estimated degree of contamination by other cell types, or the sampling method. Annotations related to the hybridization protocol, properties of the individual array or imaging process are also included. Because both sample biology and experimental settings (protocols) are complex, the list of parameters to account for is too large to be investigated by eye even for small sets of hybridizations. Because visual inspection is impossible, automatic (computer based) analysis is needed.



What is characteristic for the outlying hybridisations ?

In practise: Selecting these outliers, scanning for at least 2-fold over-or underrepresented annotation values results in values belonging to only 8 out of 111 annotations, listed in the next transperancy.

Automatic analysis of experimental annotations:

Yields 6 out of 72 annotations characteristic for the outlying hybridisation cluster

More than or exactly 2x over/underrepresented:

annotation 2: array_series
value 59 is 7x overrepresented (2/2 in cluster : 2/14 in total
value 61 is absent (0/2 in cluster : 12/14 in total

annotation 3: array_individual
value 1 is absent (0/2 in cluster : 2/14 in total)
value 2 is absent (0/2 in cluster : 2/14 in total)
value 3 is absent (0/2 in cluster : 2/14 in total)
value 4 is absent (0/2 in cluster : 2/14 in total)
value 5 is absent (0/2 in cluster : 4/14 in total)
value 6 is 7x overrepresented (2/2 in cluster : 2/14 in total)

annotation 7: array_hybridisation
value 5 is absent (0/2 in cluster : 1/14 in total)
value 6 is absent (0/2 in cluster : 1/14 in total)

annotation 39: experimentator
value 104: bastuk is absent (0/2 in cluster : 2/14 in total)

annotation 1053: temporary_additive
value 1123: none is absent (0/2 in cluster : 2/14 in total)

annotation 1055: incubation_period
value 5 is absent (0/2 in cluster : 4/14 in total)
value 10 is absent (0/2 in cluster : 2/14 in total)
value 15 is absent (0/2 in cluster : 2/14 in total)
value 20 is absent (0/2 in cluster : 2/14 in total)
value 30 is 3.5x overrepresented (2/2 in cluster : 4/14 in total)

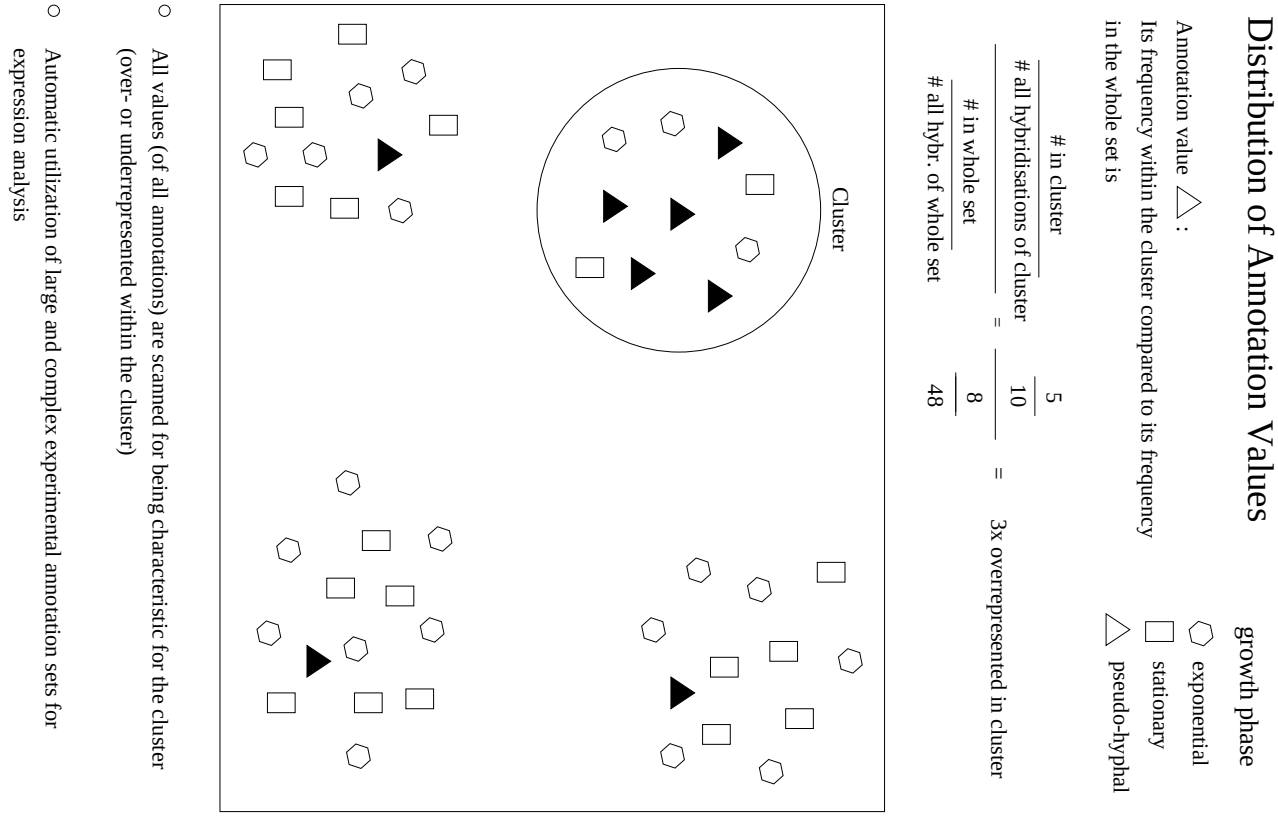
These annotations are possible candidates to explain the cluster formation. Some can be excluded when considering their meaning in the experimental context. The annotation 'incubation period' records the time points, and 'temporary additive' describes whether or not hydrogen peroxide was present in the growth medium, both only reflecting that the selected measurements belong to the 30 min timepoint.

'Label incorporation rate' and 'total activity' of incorporated label can be also disregarded for characterization of the cluster, because values annotated for the measurements in the cluster show up in mid-range for both annotations in the list.

The absence in the cluster of a particular 'experimentator', who performed two out of the twelve measurements outside the cluster is unlikely to explain the difference between cluster and other measurements. The same applies to not rehybridizing the array for the 5th or 6th time (annotation 'array_hybridization').

The first two annotations listed mean that the entire cluster was hybridized on 'array individual' 6 which is the only one stemming from 'array series' (i.e. production batch) 59, whereas all other arrays were from series 61. From other experiments, sufficient comparability among

arrays of the same production series has been observed, whereas arrays of different batches could not be directly compared. The differential array batch used for hybridization in the selected measurements causes their profiles to be different. The CA plot shows them clearly separated not only from the remaining measurements of the 30 minutes timepoint but also from all other measurements. This artifact distorts the projection of an otherwise sound and revealing dataset - omitting the two outlying measurements for analysis results in the sound and revealing CA plot on the last but one transperancy of the CA part.



How to obtain such a list? Instead of using the χ^2-test statistic to determine, which annotation values are characteristic, let's consider a simple way to access these associations. Consider the yeast specific enumeration-type annotation 'growth phase' that can take 3 different values, namely 'exponential', 'stationary' or 'pseudo-hyphal'. The corresponding hybridization data points are drawn as rectangles, hexagons and triangles, respectively. Focusing on the triangles, one can count their frequency in the encircled hybridization cluster, which is $\frac{1}{2}$ (5 out of 10) as well as in the entire set ($\frac{8}{48} = \frac{1}{6}$). Dividing the first by the second frequency suggests a 3-fold over-representation of the value 'pseudo-hyphal' in the selected cluster. In the same manner, all values of all annotations can be scanned for being characteristic, i.e. over- or underrepresented in a hybridization cluster, thus enabling automated analysis of large and complex data sets. The resulting (characteristic) experimental parameters are candidates for explaining the cluster formation, i.e. they are candidates for being the active players which

drive the cells to the observed transcriptional state.

Simple as it may be, this method already provides good analytical access to long lists of annotations and huge sets of hybridizations, which could hardly be evaluated by visual inspection. While this is a simple and easy to explain way to do so, statistical tests would certainly better suit this task. However, any statistical analysis will require that the variables (annotations) are of categorical range and that instances of occurance can be counted for any annotated value:

Free text annotation:

- misspellings
- meaning of words depend on context
- different researchers use different words to represent the same item

interfere with counting such values !

Misspellings, different textual representations of semantically identical items, and, *vice versa*, ambiguous words whose meaning depends on the context, interfere with counting such values. With these limitations to access for computer based, i.e. statistical, analysis, global studies of large data sets will not be possible.

Free text annotation

- assigning defined values by text mining
  - big effort
  - recovering a share of the original information
  - percentage may be low

or

- data cemetery

--> loss of information

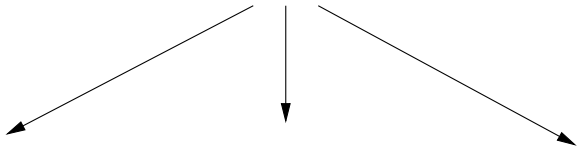
Defined values

- instances of occurance are countable

--> 100% information accessible

Statistical Access to Experiment Annotations

PROBLEM: **Values are not directly countable in free text annotation**



solution 1
pattern matching

solution 2
few free text fields

solution 3
no free text at all

Instead of tolerating free text annotation in addition to a greater or smaller share of "controlled vocabulary", we do without any freetext.

No Freetext

Freetext grants the flexibility to annotate any value.

Now one has to define annotations and values, e.g.

- **a list of valid enumeration type values for each categorical variable**
- **a unit for each continous variable**

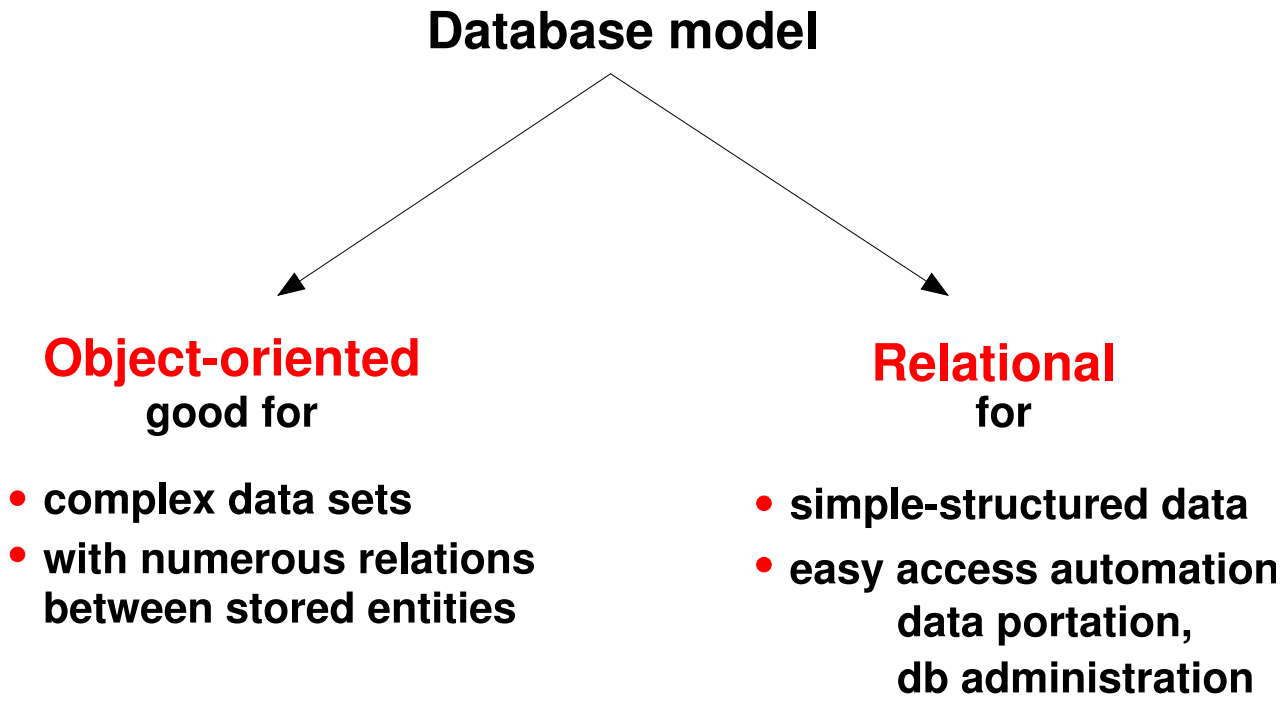
PROBLEM:

List of both variables and their defined values grows with every new experimental design

--> DB structure must be more flexible

to cope with evolving definitions

While in free text descriptions the number of occurences of a value is not directly countable, dispensing with free text also causes problems. An arbitrary-length free text field allows to annotate each possible value and may also take any number of such atomic pieces of information. In contrast, the type of annotation described above is restricted to predefined values. New annotations and/or new values for existing annotations have to be added constantly as new experiments are designed. This requires the ability to define new annotations rapidly without altering the database scheme, i.e. during normal database operation. The absence of highly flexible free text annotations has to be compensated for by increased flexibility in database storage.

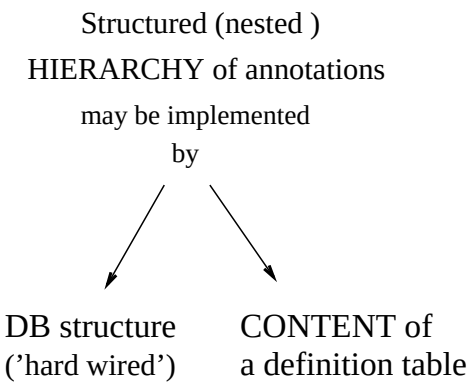


In principle, a microarray database could be either object-oriented or relational. The object-oriented model is chosen for complex data sets where numerous relations exist between the stored entities. In contrast, relational databases are convenient for simple-structured data and easy to handle with respect to access automation, data portation, and database administration. A microarray database will consist mostly (more than 99% of storage space in our databases) of intensity data which can be perfectly stored in tables and show few relations to other items. I therefore decided to focus on the relational rather than the object oriented model due to the simplicity and good portability among different database management systems (DBMS).

A relational database consists of

- **relations**, also called tables. Such a table relates between
- **attributes** also referred to as data fields or columns of such a table and may contain an arbitrary number of
- **tuples**, also termed records or datasets, which are represented as the rows of a table.

In addition to 'table', 'column', and 'row', I will frequently use the formal relational terms *relation*, *attribute*, and *tuple*, respectively.



Given the tool of a relational DBMS, storage of e.g. experiment annotations can be implemented in different ways.

by STRUCTURE

annotations stored as attributes
(= columns, fields)

experiment	array_source	array_support	spotted_material
1	self_made	nylon	PCR
2	genome_systems	polypropylene	PCR
3	self_made	glas	colonies
...			

by CONTENT

... stored in table content

DEFINITIONS

annotation	defined value
array_source	self_made
array_source	genome_systems
array_support	glass
array_support	nylon
array_support	polypropylene
spotted_material	colonies
spotted_material	PCR

ANNOTATIONS

experiment	annotation	value
1	array_source	self_made
1	array_support	nylon
1	spotted_material	PCR
2	array_source	genome_systems
...		

The parameter names such as "array source", let us refer to them as "annotations", may become the attributes (column names) of a single table. Another possibility is to make them the content (the tuples) of a first table, whose only purpose is to define the annotations along with the values they may take. Here, a second table is needed to store the actual values taken in particular experiments.

	by STRUCTURE	by CONTENT
new annotation	new attribute (column) have to be found and handled by algorithms	new entry in definition table ==> no problem
new kind of annotations	new table have to be found and handled by algorithms	new entry in 'headings' table ==> no problem
hierarchy becomes more complex (increased nesting depth)	! new hierarchy has to be recognized or described externally	new column in 'headings' table ==> no problem

The increase in redundancy - annotations only take a tiny share of the storage space anyway - is more than compensated for by the increase in flexibility. New items can be inserted without changing the database structure nor any algorithm operating on it.

Thus, a "by content" - implementation deals with the complexity e.g. of experiment annotations already at the database level. It provides a standardized platform for algorithms (which may be complex enough without that task).

experimental annotations: DEFINITIONS

annotationheadings

heading1no	heading1	heading2no	heading2	heading3no	heading3
1	common_annotations	1	array	1	-
1	common_annotations	2	hybridisation	2	RNA_preparation
1	common_annotations	2	hybridisation	3	labeling
1	common_annotations	2	hybridisation	4	hybridisation_conditions
1	common_annotations	2	hybridisation	5	stringency_wash
1	common_annotations	2	hybridisation	6	detection
1	common_annotations	3	sample	7	-
2	organism_specific_annotations	4	genotype	8	-
2	organism_specific_annotations	4	genotype	9	-
2	organism_specific_annotations	4	genotype	10	auxotrophic_marker
2	organism_specific_annotations	5	growth_conditions	11	-
2	organism_specific_annotations	6	medium	12	-
2	organism_specific_annotations	6	medium	13	C_source
2	organism_specific_annotations	6	medium	14	additive

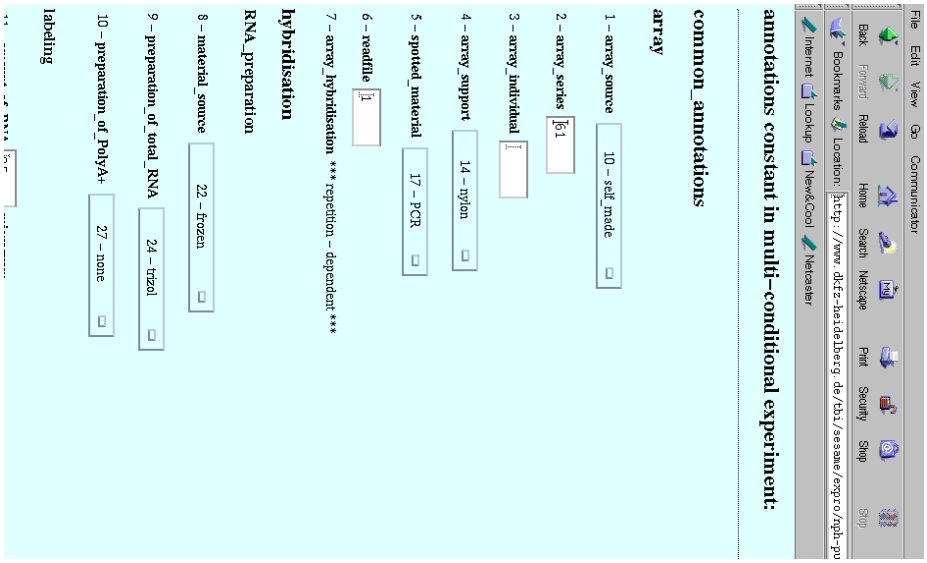
(14 rows)

annotations

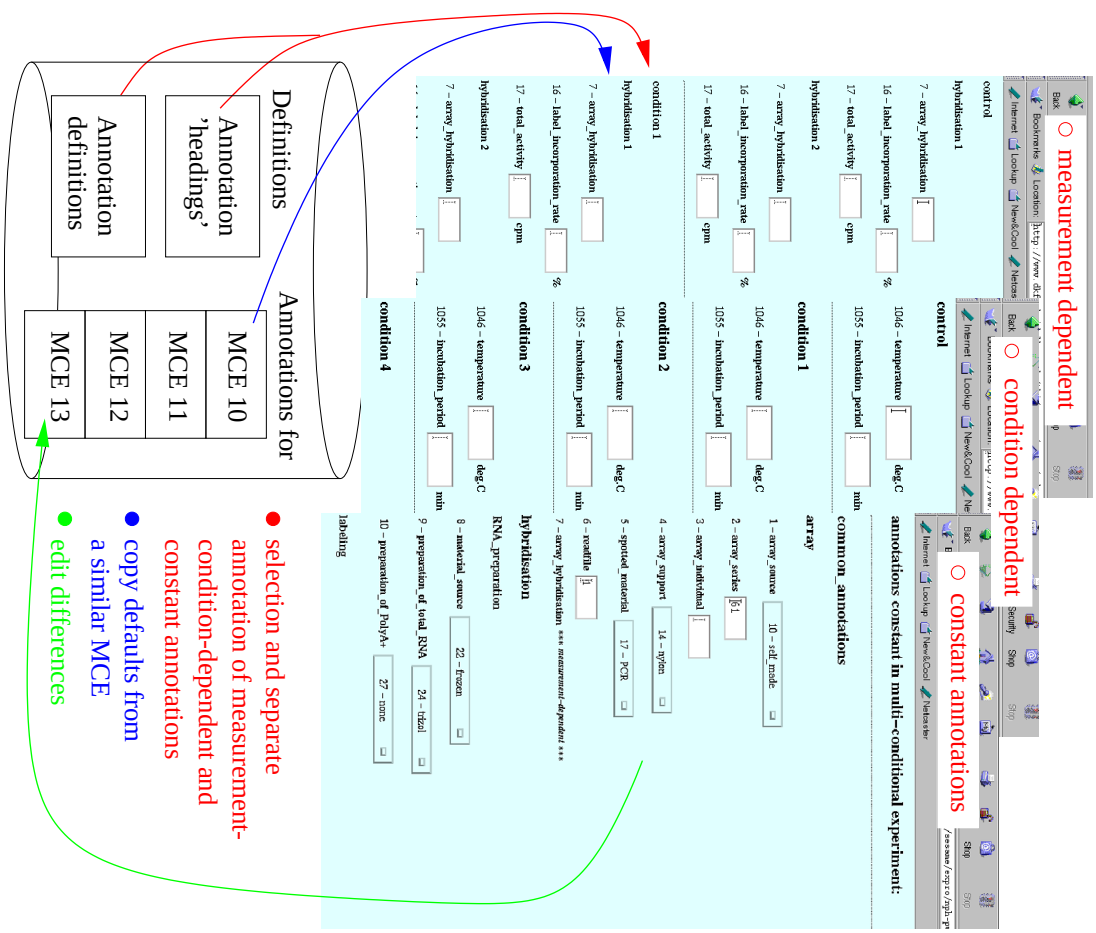
lastheadingno	ano	nextano	annotation	vno	nextvno	value
1	1	2	array_source	10	11	self_made
1	1	2	array_source	11	12	genome_systems
1	1	2	array_source	12	13	clontech
1	1	2	array_source	13	14	research_genetics
1	2	3	array_series	0	0	[]
1	3	4	array_individual	0	0	[]
1	4	5	array_support	14	15	nylon
1	4	5	array_support	15	16	polypropylene
1	4	5	array_support	16	17	glass
1	5	6	spotted_material	17	18	PCR
1	5	6	spotted_material	18	19	colonies
...						
13	50	51	galactose	0	0	[%]
13	51	52	ethanol	0	0	[%]
13	52	53	glycerol	0	0	[%]
14	53	54	temporary_additive	121	122	H2O2
14	53	54	temporary_additive	122	123	NaCl
14	53	54	temporary_additive	123	-1	none
14	54	55	concentration	0	0	[mM]
14	55	-1	incubation_period	0	0	[min]
12	48	49	base	118	120	SDC

(135 rows)

In practise, the definition table may be supplemented by another table storing a hierarchy. This one has 3 columns taking sections, subsections and subsubsections. Independent of the nesting depth, the numbering of the so-to-speak "smallest" (last) headings relates to the attribute lastheadingno in the annotations table, thus connecting the two tables.The attributes 'ano' and 'vno' are used as IDs to reference annotations or their values, respectively. The attributes 'nextano' and 'nextvno' point to the next entry, thus implementing a linked-list structure. Values that contain square brackets are not necessarily categorical but are meant to take a number, e.g. a production batch ID. If a unit can be defined for the value, it will be listed within the brackets.



The content of these definition tables serves as meta data to compile html forms used during the process of annotating an experiment.



A convenient way minimizing efforts in annotation. Every piece of information has to be entered only once. The annotation process may start with copying default values from the most

similar multiconditional experiment (MCE). Secondly, from the complete list of defined annotations the measurement-dependent ones (those taking different values for each hybridization or channel such as 'label_incorporation_rate') are selected and then annotated for each single measurement. Afterwards, from the remaining annotations, those being condition-dependent (taking different values for each experimental condition under study) for the particular experiment are chosen and annotated for each experimental condition. For the constant annotations, it suffices to edit few, if the questionaire is prefilled with default values copied from a similar experiment.

experimental annotations: annotated values

y1_constant_categorical_1

experiment	ano	annotation	vno	cvalue
1	10	preparation_of_PolyA+	27	none
1	30	background_correction	69	none
1	12	enzym	31	superscript
1	31	spot_detection	74	all_spots
1	13	priming	35	anchored_oligo_dT
1	32	spot_size	76	fixed
1	14	nucleotide	38	dCTP
1	33	intensity_measurement	79	sum_of_pixels_within_boundary
1	15	label	40	33P

(33 rows)

y1_constant_number_1

experiment	ano	annotation	vno	value
1	20	hybridisation_temp	0	65
1	21	buffer_volume	0	5
1	22	hybridisation_length	0	16
1	24	times	0	2
1	25	wash_length	0	30
1	26	temperature	0	65
1	27	exposure_time	0	72
1	46	temperature	0	30
1	49	glucose	0	2
1	2	array_series	0	53

(17rows)

y1_conditiondependent_1

experiment	condition	ano	annotation	vno	cvalue	nvalue
1	5	55	incubation_period	0	+++	50
1	0	3	array_individual	0	+++	3
1	1	3	array_individual	0	+++	4
1	2	3	array_individual	0	+++	5
1	3	3	array_individual	0	+++	6
1	4	3	array_individual	0	+++	7
1	5	3	array_individual	0	+++	8
1	0	55	incubation_period	0	+++	0

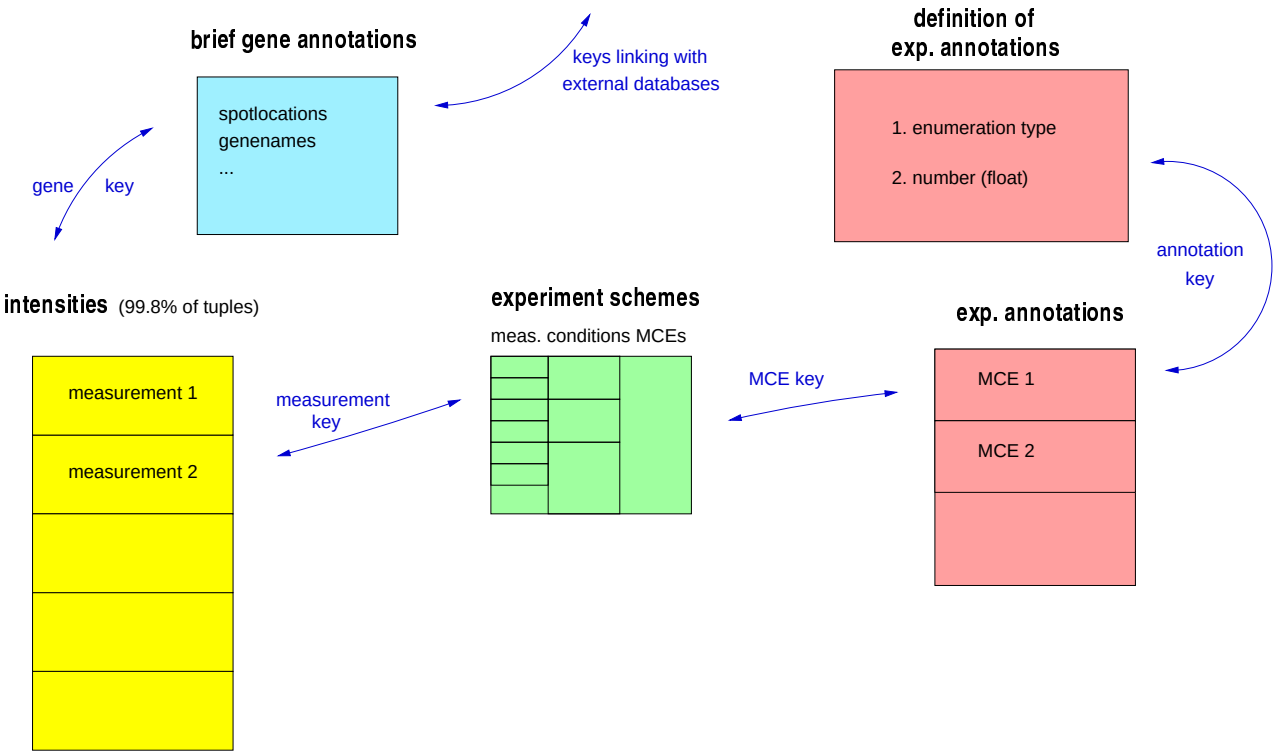
(12 rows)

y1_measurementdependent_1

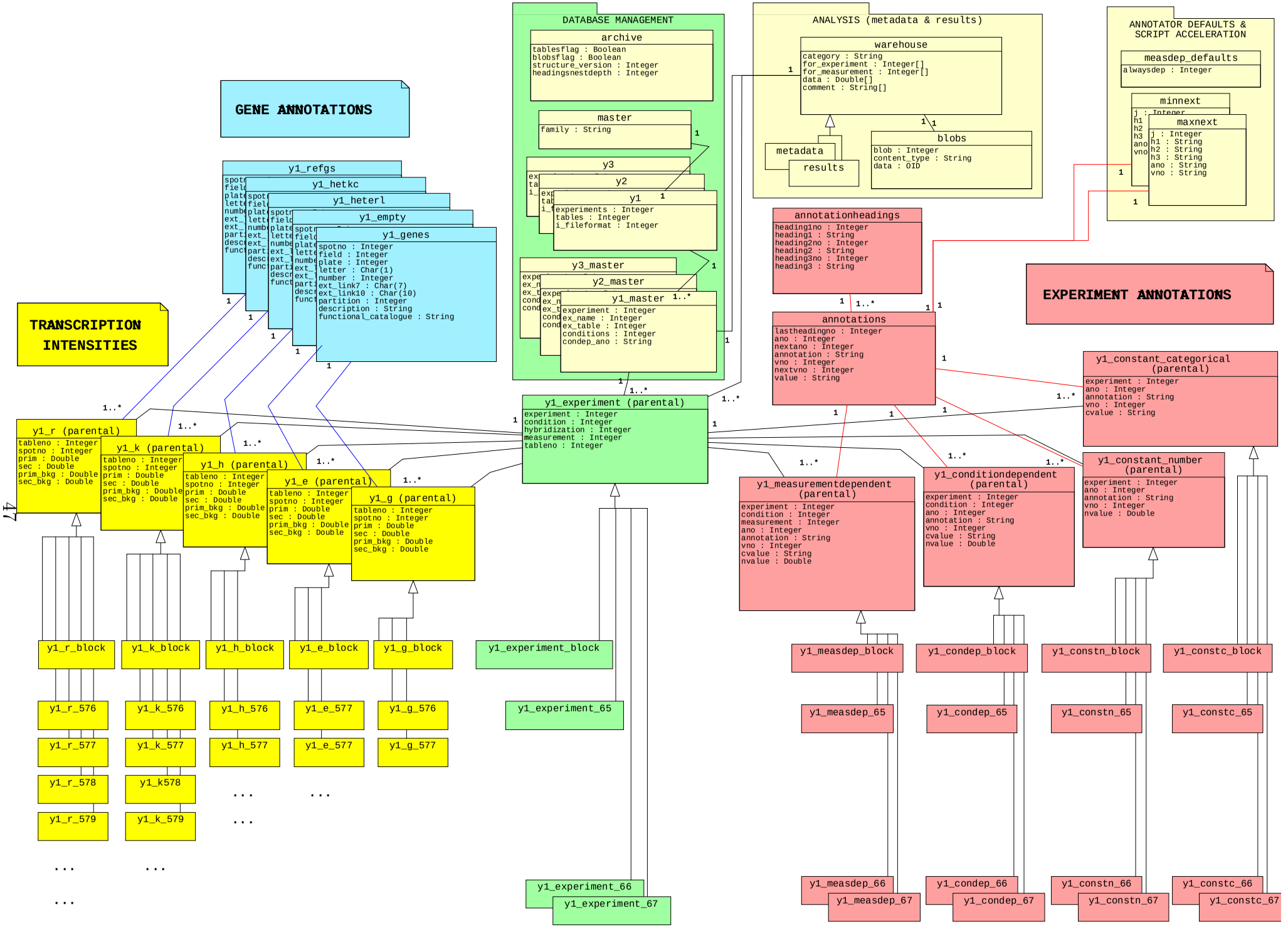
experiment	condition	measurement	ano	annotation	vno	cvalue	nvalue
1	5	2	16	label_incorporation_rate	0	+++	70
1	4	3	16	label_incorporation_rate	0	+++	70
1	5	3	17	total_activity	0	+++	10000000
1	4	0	17	total_activity	0	+++	10000000
1	0	1	7	array_hybridisation	0	+++	2
1	1	1	7	array_hybridisation	0	+++	3
1	2	1	7	array_hybridisation	0	+++	2

(54 rows)

While the contents of the definition tables are used as meta data by the web-based user interface to compile multiple-choice forms, the results of the annotation process are stored in annotation tables. These tables take the annotations along with their values taken for a particular MCE, condition or measurement.



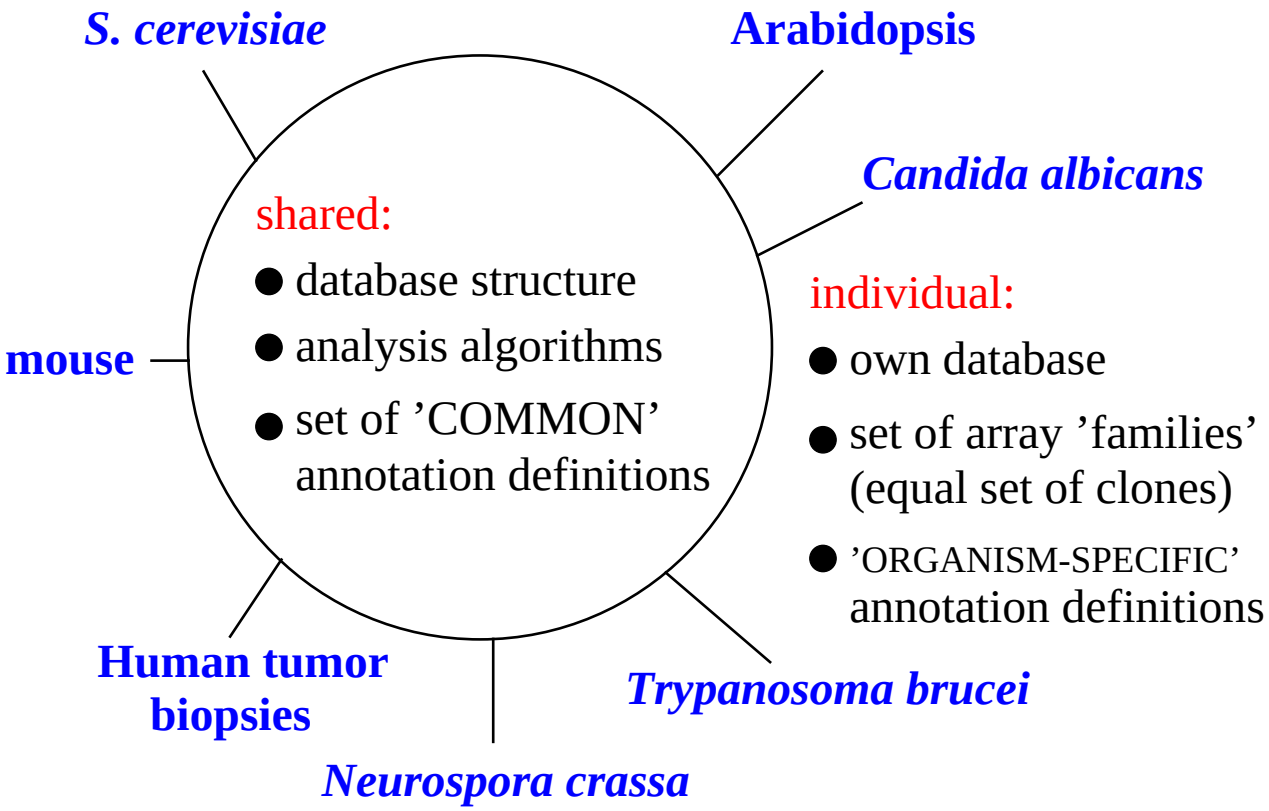
Above transperancy shows them (lower right boxes) in database context. Gene annotations, signal intensities (please note the percentage!), and experiment annotations are displayed in blue, yellow and red, respectively, also in the next transperancy. The gene annotations are linked both with the transcription intensities and with public external gene databases (e.g. GO) in order to enable explicit characterization of genes showing a particualar transcription behaviour. The intensities are stored as measurements. A measurement (i.e. a hybridization for radioactive or a single channel for multi-channel data) comprises a single value for each spot on the microarray. Experiment schemes record for each measurement which hybridization and experimental condition it belongs to, and which multiconditional experiment (MCE) this condition is contained in. The experiment schemes are the 'storekeepers' of the database, relating intensity data with experiment annotations, which allow for explicit characterization of measurements showing a particular expression pattern.


GENE ANNOTATIONS
TRANSCRIPTION INTENSITIES
EXPERIMENT ANNOTATIONS
DATABASE MANAGEMENT
archive
tablesflag : Boolean
blobsflag : Boolean
structure_version : Integer
headingsnestdepth : Integer
master
family : String
y3
y2
y1
experiments : Integer
tables : Integer
i_fileformat : Integer
y3_master
y2_master
y1_master
experiment : Integer
ex_name : Integer
ex_table : Integer
conditions : Integer
condep_ano : String
ANALYSIS (metadata & results)
warehouse
category : String
for_experiment : Integer[]
for_measurement : Integer[]
data : Double[]
comment : String[]
metadata
results
blobs
blob : Integer
content_type : String
data : OID
ANNOTATOR DEFAULTS & SCRIPT ACCELERATION
measdep_defaults
alwaysdep : Integer
minnext
maxnext
j : Integer
h1 : String
h2 : String
h3 : String
ano : String
vno : String
y1_refgs
y1_hetkc
y1_heterl
y1_empty
y1_genes
spotno : Integer
field : Integer
plate : Integer
letter : Char(1)
number : Integer
ext_link7 : Char(7)
ext_link10 : Char(10)
partition : Integer
description : String
functional_catalogue : String
annotationheadings
heading1no : Integer
heading1 : String
heading2no : Integer
heading2 : String
heading3no : Integer
heading3 : String
annotations
lastheadingno : Integer
ano : Integer
nextano : Integer
annotation : String
vno : Integer
nextvno : Integer
value : String
y1_experiment (parental)
experiment : Integer
condition : Integer
hybridization : Integer
measurement : Integer
tableno : Integer
y1_constant_categorical (parental)
experiment : Integer
ano : Integer
annotation : String
vno : Integer
cvalue : String
y1_constant_number (parental)
experiment : Integer
ano : Integer
annotation : String
vno : Integer
nvalue : Double
y1_r (parental)
y1_k (parental)
y1_h (parental)
y1_e (parental)
y1_g (parental)
tableno : Integer
spotno : Integer
prim : Double
sec : Double
prim_bkg : Double
sec_bkg : Double
y1_measurementdependent (parental)
experiment : Integer
condition : Integer
measurement : Integer
ano : Integer
annotation : String
vno : Integer
cvalue : String
nvalue : Double
y1_conditiondependent (parental)
experiment : Integer
condition : Integer
ano : Integer
annotation : String
vno : Integer
cvalue : String
nvalue : Double
y1_r_block
y1_k_block
y1_h_block
y1_e_block
y1_g_block
y1_experiment_block
y1_measdep_block
y1_condep_block
y1_constn_block
y1_constc_block
y1_r_576
y1_r_577
y1_r_578
y1_r_579
y1_k_576
y1_k_577
y1_k578
y1_k_579
y1_h_576
y1_h_577
y1_e_577
y1_g_576
y1_g_577
y1_experiment_65
y1_measdep_65
y1_condep_65
y1_constn_65
y1_constc_65
y1_experiment_66
y1_experiment_67
y1_measdep_66
y1_measdep_67
y1_condep_66
y1_condep_67
y1_constn_66
y1_constn_67
y1_constc_66
y1_constc_67

This scheme takes over arrangement and color code of the overview, dissolving it into database relations and their attributes. According to the Unified Modeling Language (UML) specifications[6] of the Object Management Group (OMG), a database relation - in the world of objects represented by a so called 'class' - is depicted as a box containing its name and, separated by a horizontal line, its attributes. Building on the Entity-Relationship-Model (ERM) of P. Chen [9], relationships between these relations (or classes) can be of three different kinds:

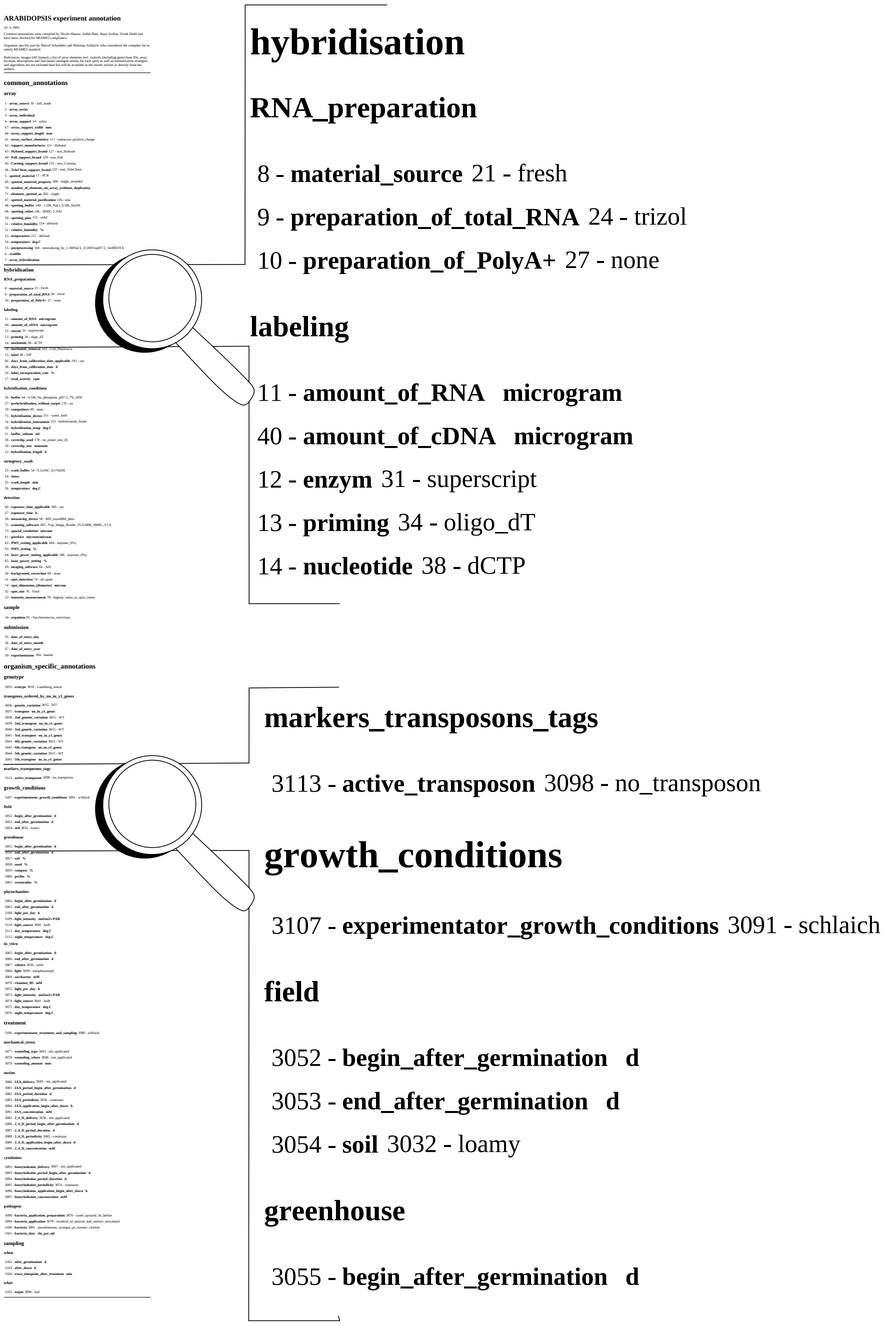
- 1-to-1 relationships are depicted as '1—1'. Each tuple (i.e. entry) of relation A corresponds to exactly one tuple stored in relation B.
- Many-to-1 relationships, drawn '1..*—1', indicate that each entry in B may correspond to more than one entry in A.
- Many-to-many relationships are resolved by a connecting intermediate relation (e.g. the green table in the center of the diagram).

Table inheritance - on a more abstract level represented by a generalized relationship of a subclass sharing the structure or behaviour of a superclass - is indicated by arrows. In M-CHIPS, all child tables have exactly the same structure as their parents (rather than showing additional attributes). The attributes of these child tables have been omitted in the diagram for visual clarity. For the same reason, tables of identical structure overlap.



[6] http://www.omg.org/technology/documents/formal/uml.htm

Each field of research is represented by an individual database containing a set of array 'families' (each standing for a particular kind of array with a certain spotting scheme). The field of research is represented by a particular set of 'organism-specific' annotation definitions (comprising e.g. medium components for yeast or tumor stage for human tumor samples). All these databases share the same structure and can therefore be handled and analyzed by the same algorithms. There is also a set of 'common' annotation definitions, i.e. those used by all users (e.g. label incorporation rate).


hybridisation
RNA_preparation
8 - material_source 21 - fresh
9 - preparation_of_total_RNA 24 - trizol
10 - preparation_of_PolyA+ 27 - none
labeling
11 - amount_of_RNA microgram
40 - amount_of_cDNA microgram
12 - enzym 31 - superscript
13 - priming 34 - oligo_dT
14 - nucleotide 38 - dCTP
markers_transposons_tags
3113 - active_transposon 3098 - no_transposon
growth_conditions
3107 - experimentator_growth_conditions 3091 - schlaich
field
3052 - begin_after_germination d
3053 - end_after_germination d
3054 - soil 3032 - loamy
greenhouse
3055 - begin_after_germination d
ARABIDOPSIS experiment annotation

These common annotations are related to the microarray technique, describing the array, RNA preparation, labelling, hybridization and washing conditions and signal detection. The second half of the list consists of organism-specific annotations.

An important issue for implementing and running a database is data integrity also called data consistency. Suppose a valid alteration of the data, defined by a block of sequentially performed operations (such a block is called a 'transaction') breaks down after doing half of the work. A table could have been deleted but remains registered in the table administrating system catalogue of the database system. Another example may be the task to add 500 Euro to everyone's salary in a table containing employees and now it is unknown which row was updated and which not. In both cases data integrity (database consistency) is violated. To prevent damage, the DBMS should be transaction-based. In above cases, the whole transaction will undergo a "rollback" upon occurance of the error, i.e. the database will be put back to the state before the start of the transaction. Furthermore it is important to prevent unauthorized access and to have at hand both global and partial backups to restore the complete system or accidentally deleted experiments, respectively.

- Performance
  - Minimizing query space:
    Separate storage for tuples normaly searched for at different times, e.g. genes, empty spots, external controls
  - Write / delete contra read access
    Hybridizations are quickly written / deleted as single tables, but queries are slowed down as number of tables increases:
    - fast writing / deleting hybridizations as separate tables
    - fast reading from a large block
- Towards ANALYSIS
  - Large and complex experimental annotation space doesn't have to be examined by eye:
    Characteristic annotation scores for a hybridization cluster revealed by mouseclick

While the tables containing the gene annotations have only as many tuples (table rows) as there are genes, transcription intensities add up to this number of entries for each single measurement. Gene and experiment annotations on average only take 0.35% of the storage space. Since this amount is far too small to be relevant for query performance, flexibility remains the only time-saving aspect related to experiment annotations. Performance considerations are related only to the hybridization intensities. Among all intensities, analysis focuses on spots that represent genes as opposed to empty spots and various kinds of controls. For this reason we use different tables to store these kinds of intensities, thus minimizing query space.

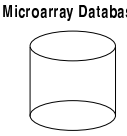
Having stored intensities and background for genes, empty spots and different categories of controls, fast querying of tuples for all these categories is mediated by so-called indices, which immediately guide the search to the specified tuples. If all measurements were stored in one large table per category, adding a new measurement would be slow because of the time necessary for recomputing the indices. Therefore, new measurements are inserted as separate tables, computing indices only for the new tuples.

However, database search is slowed down by increasing the number of separate tables because there is no global index immediately guiding the search to the table containing the tuples. Although high performance for write/delete operations is achieved, read access is slow for a large number of separate tables. In order to optimize both writing and reading operations, we write or delete measurements as separate tables, but read from large 'block' tables that are

filled by over-night jobs collecting measurements that are no longer to be altered or deleted. Thus, computation of large indices is performed at times of low traffic as an investment in query performance. Table inheritance is used as an elegant aid in keeping track of both single and block tables. Since each access to the intensity tables is directed via one of the parental tables, query syntax does not change when a set of tables is merged into one block. This block will be a child of a specific parental table as are the tables to be merged (UML scheme, small yellow tables). Thus the event takes place at the underlying database level, being completely insulated from the level of accessing algorithms for reduced complexity.

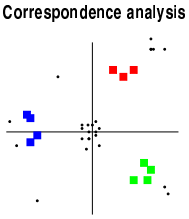
The only access property changed by this process is query speed. On a SUN E450 server under Solaris 2.7, a PostgreSQL 6.5.3 server process retrieves two consecutively uploaded hybridizations (comprising 6103 yeast genes in double spotting) out of 686 stored in separate tables on average in 85 seconds. The same query performs in 2.3 seconds, if the 686 hybridizations are assembled into one large table. Even retrieving two out of 2251 hybridizations takes only 2.8 seconds when all hybridizations are *en bloc*.

Summary



achievements:
- > 3000 hybridizations of 11 different fields of research
- in a common data format

⇒ statistical access also to experiment anntotations



advantages:
- integrated visualization of genes and hybridizations
- visualizes intricate details such as subtle deviations of objects from the expected state
- is explorative

achievements:
- adaption to the requirements of microarray data (normalization, filtering, HMS)
- integrated analysis of experiment annotations

This last transperancy summarizes the achievements we made with correspondence analysis ontop a customized data warehouse solution.

More information / references

- Further information incl. a detailed description our storage scheme,
- free-text free experiment annotation definitions for 11 different organisms, and
- public data (recently 292 public hybridizations)

can be obtained at `http://www.dkfz-heidelberg.de/tbi/services/mchips`.

Bibliography

[1] J. Aach, W. Rindone, and G. M. Church. Systematic management and analysis of yeast gene expression data. *Genome Res.*, 10:431–445, 2000.

[2] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10101–10106, 2000.

[3] D. E. Basset Jr., M. B. Eisen, and M. S. Boguski. Gene expression informatics – it's all in your mine. *Nat. Genet.*, 21(Suppl.):51–55, 1999.

[4] A. Ben-Dor and Z. Yakhini. Clustering gene expression patterns. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the 3rd annual conference on computational molecular biology (RECOMB 99)*, pages 33–42. ACM Press, 1999.

[5] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

[6] A. Brazma, A. Robinson, G. Cameron, and M. Ashburner. One-stop shop for microarray data. *Nature*, 403:699–700, 2000.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth and Brooks/Cole, Monterey, CA, 1984.

[8] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, and T. S. Furey. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.*, 97:262–267, 2000.

[9] P. Chen. The entity-relationship approach: Toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976.

[10] Y. Cheng and G. M. Church. Biclustering of expression data. In R. Altman, T. L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I. N. Shindyalov, L. F. Ten Eyck, and H. Weissig, editors, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 93–103. AAAI Press, 2000.

[11] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, Dept. of Statistics, UC Berkeley, CA, 2000.

[12] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, 1998.

[13] O. Ermolaeva, M. Rastogi, K. D. Pruitt, G. D. Schuler, M. L. Bittner, Y. Chen, R. Simon, P. Meltzer, J. M. Trent, and M. S. Boguski. Data management and analysis for gene expression arrays. *Nat. Genet.*, 20:19–23, 1998.

[14] R. A. Fisher. The use of multiple meausrements in taxonomic problems. *Ann. Eugenics*, 7:300–8, 1936.

[15] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. In R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the 4th annual international conference on computational molecular biology (RECOMB00)*, pages 127–135. ACM Press, 2000.

[16] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, 2000.

[17] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Holler, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[18] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression profiles. *Genome Biol.*, 1:research003.1–research003.21, 2000.

[19] S. G. Hilsenbeck, W. E. Friedrichs, R. Schiff, P. O'Conell, R. K. Hansen, C. K. Osborne, and S. A. W. Fuqua. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer Inst.*, 91:453–459, 1999.

[20] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. T. Trent, and P. S. Meltzer. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, 58:5009–5013, 1998.

[21] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, 7:673–679, 2001.

[22] I. Lefkovits, L. Kuhn, O. Valiron, A. Merle, and J. Kettman. Toward an objective classification of cells in the immune system. *Proc. Natl. Acad. Sci. U.S.A.*, 85(10):3565–9, May 1988.

[23] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, pages 18–29, 1998.

[24] R. Sharan and R. Shamir. Click: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:307–16, 2000.

[25] C. Stoeckert, A. Pizarro, E. Manduchi, M. Gibson, B. Brunk, J. Crabtree, J. Schug, S. Shen-Orr, and G. C. Overton. A relational schema for both array-based and sage gene expression experiments. *Bioinformatics*, 17(4):300–8, Apr 2001.

[26] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, 96:2907–2912, 1999.

[27] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22:281–285, 1999.

[28] 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.

[29] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. r. Olson JA, J. R. Marks, and J. R. 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.

[30] C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17 Suppl 1:S316–22, 2001.