Storage and analysis of microarray data

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Kurt Fellenberg
aus Gummersbach

Heidelberg, 2002

Berichterstatter:

Prof. Dr. Dietmar Schomburg

Prof. Dr. Heinz Saedler

Tag der letzten mündlichen Prüfung: 8. Mai 2002

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

Abstract

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 present work 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 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.

Introduction

Microarray technology

Cells accomplish metabolic processes, they comply with their growth program, adapt to changing environments, or communicate with other cells by accurately controlled expression of appropriate proteins. The protein portfolio is tailored to the particular requirements of the cell. To this end, gene-coding sequence in the DNA is transcribed and leaves the cell nucleus as messenger RNA (mRNA). In the surrounding cytoplasm, each mRNA molecule conducts the synthesis of the particular protein encoded. Presence and amount of a particular mRNA regulates the presence and the amount of the encoded protein.

For a particular cell status, the level of mRNA can be measured in parallel for thousands of genes by DNA chip technology. RNA is prepared from the cells and is reversely transcribed into more stable DNA, simultaneously incorporating radioactive or fluorescent labels. The labeled DNA is then applied to a DNA chip, also referred to as DNA array or microarray.

The microarray consists of support material, onto which DNA fragments of different sequence, representing genes, have been spotted. These spots serve to measure the level of the applied DNA obtained from the RNA sample. For DNA of the same kind, complementary single strands will bind ('hybridize'), resulting in double stranded DNA. Therefore, the level of labeled DNA bound to a particular spot will correspond to the level of the particular kind of mRNA in the cells. The amount of label is measured for each spot.

Thus, microarray technology provides insight into the transcriptional status of the cell ('transcriptome'), measuring RNA levels for thousands of genes at once [1, 2, 3, 4]. As such it has become an important tool in functional genomics [4, 5, 3, 6].

In addition to the advantage of parallel investigation of many genes, the benefit of the technique lies in its broad range of applications. Applications range from the study of organisms with a particular gene inactivated ('knockout mutants') to the investigation of the adaptation of cells to different environmental conditions (e.g. time series). Microarrays can be used in pharmaceutical studies to gain information for discovery and design of effective substances for medical therapy [7, 8]. Moreover, microarrays can be employed for diagnostic purposes,

e.g. for the detection of dispositions for hereditary diseases [9]. In combination with mass spectrometry, they enable the identification of SNPs (single nucleotide polymorphisms) [10]. For cancer research, genome-wide transcription is measured to classify tumor samples, i.e. to predict class-membership for new samples [11, 12].

Ongoing sequencing projects promise to yield complete gene sets for most model organisms in the near future, which can then be mounted on DNA chips. This will increase the number and diversity of transcriptional assays performed.

Experimental setting

A microarray consists of DNA fragments immobilized on a solid support. Common support materials are nylon, polypropylene or glass. The surface is often chemically treated prior to DNA immobilization to improve DNA binding properties. Immobilization on the support is done by either 'spotting' cDNA fragments or synthesizing oligonucleotides 'on chip' [13, 14, 15, 16, 17].

Transcriptional profiling with microarrays involves several steps (Fig. 1). mRNA is prepared from cells growing under certain experimental conditions. For each condition the prepared mRNA is separately subjected to reverse transcription with radioactively or fluorescent-tag labeled nucleotides. Radioactive labeling is often carried out in combination with nylon or polypropylene supports, while flurescence-labeled targets are mostly hybridized to glass chips. Whereas radioactive labeling has not yet been reported to be used with more than one kind of label per hybridization, fluorescent dyes of different color offer the opportunity to yield more than one channel. With two-channel fluorescent-tag labeling each hybridization involves additional application of a differently labeled cDNA, stemming from a control condition. Subsequently, the labeled cDNA mixture is hybridized to the microarray. After detection of the signals, image analysis programs are used to determine spot intensities [18].

Resulting data

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. Data thus produced by image analysis may be regarded as a table, each row representing a gene, each column an experimental condition. However, 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 hybridization and represent it by a separate column in the table. One condition of a multi-conditional experiment can thus comprise several columns (Fig. 1).

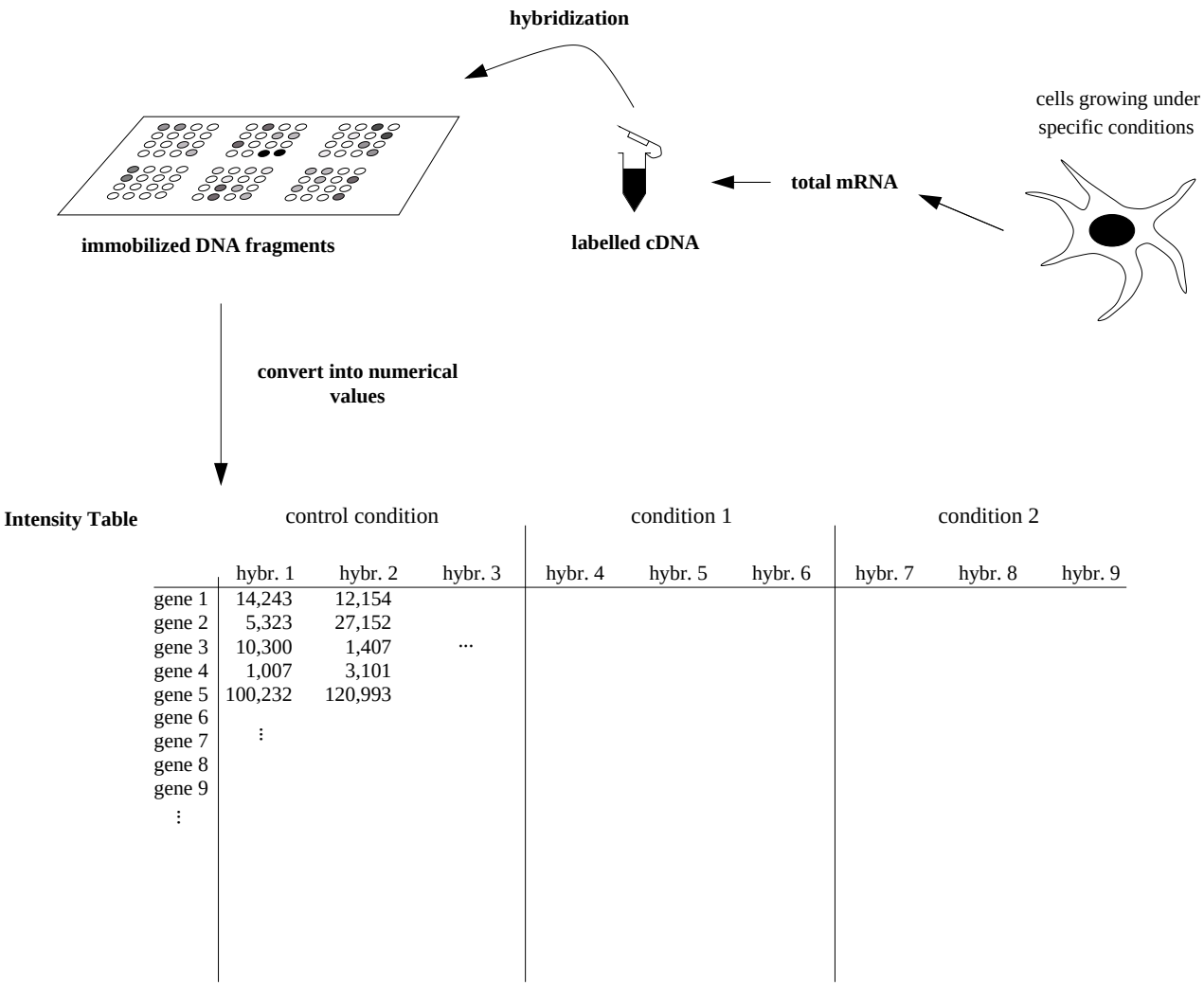

Figure 1: **Microarray hybridization**. mRNA is prepared from cells growing under specific experimental conditions. It is labeled, i.e converted to more stable cDNA by reverse transcription using radioactively or fluorescence-tagged nucleotides and hybridized to an array. The scheme depicts only radioactive-label, i.e. a single-channel setup for simplicity. The detected signals are then converted into numbers by imaging software. The output of several hybridizations can be regarded as a table with its rows representing the spotted elements. These may be genes or expressed sequence tags (ESTs). The columns of the table stand for the performed hybridizations. Courtesy of Benedikt Brors, modified.

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 hybridizations. Therefore the columns of the table have to undergo a normalization procedure, correcting for affine-linear transformation among the columns. 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.

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. Furthermore, microarray data do not consist of the transcription intensities alone. Data sets should also include information about the immobilized DNA fragments as well as a detailed description of the experimental steps performed, imposing high demands both on data analysis and storage. The storage should be interconnected with analysis, i.e. it should hold the data in a format suitable for computer-based analysis, because visual inspection is impractical due to the large volume of these data.

Current methods of data storage and analysis

Data storage

To enable interpretation of large data sets, the data produced need to be stored in a suitable way to allow for global comparison [6]. 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 (Fig. 2).

Several database projects are currently addressing these questions. While ExpressDB (Harvard, [19]) 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[1] (NCGR), GEO[2] (NCBI), ArrayDB (NHGRI, [20]), ArrayExpress (EBI, [21]), and RAD[3](UPenn, [22]), the last one combining both objectives.

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

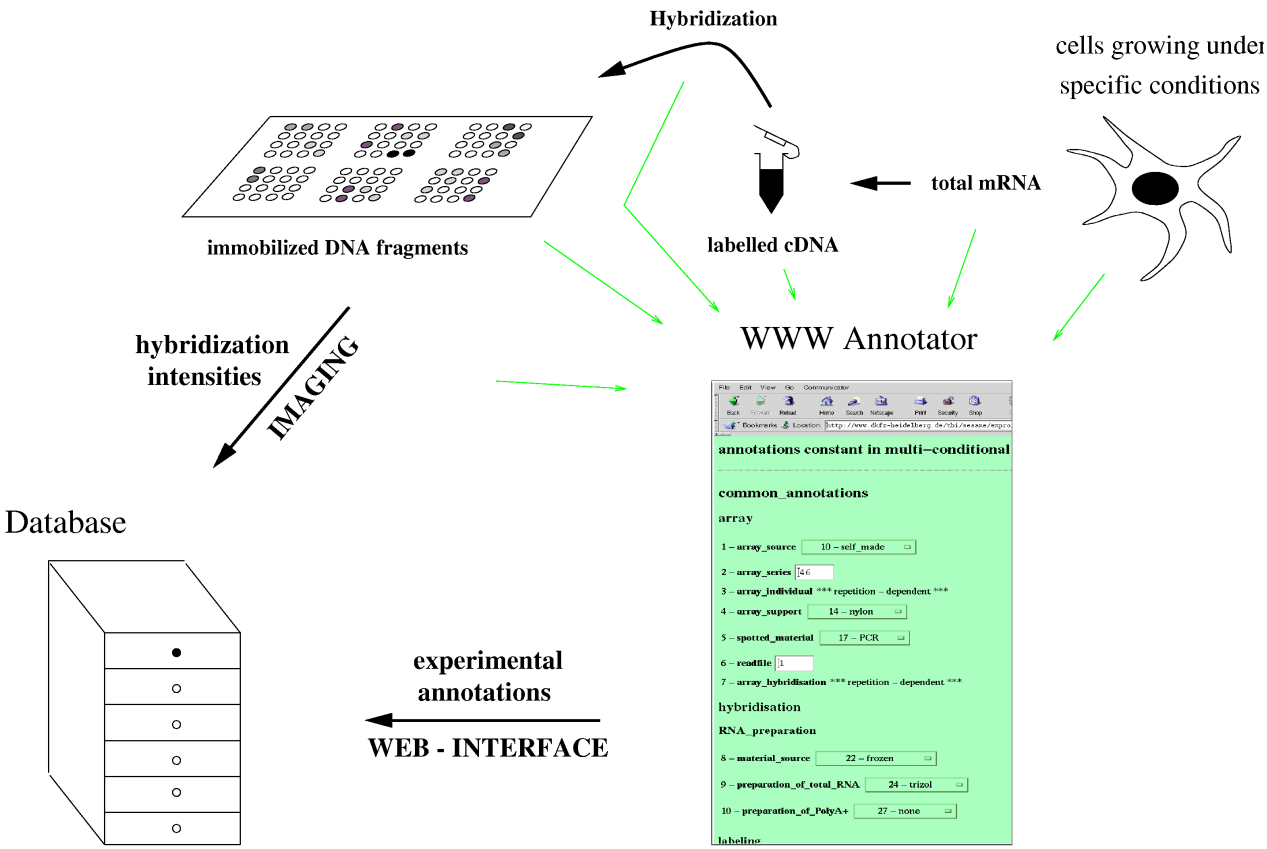

Figure 2: **Data upload.** Along with the transcription intensities, experiment annotations have to be stored. These should explicitely characterize the sample and its treatment, RNA preparation and labeling steps, hybridization and washing as well as the imaging process in sufficient detail. Courtesy of Benedikt Brors, modified.

Data analysis

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 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-type, stage or grade, often supplemented by normal tissue of the particular organ [11]. Examples of classification methods range from linear discriminant analysis [23] to support vector machines [24] or classification and regression trees (CART, [25,26]). 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 [28], and self-organizing maps [29]. 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) [30] or principal component analysis (PCA) [31, 32] as well as the technique mentioned in this study, 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 (Fig. 3). Vice versa, the rows of the data table, i.e. the n genes, can be represented in m-dimensional hybridization space. MDS or PCA can be used to visualize either the former or the latter.

Such a projection plot is an explorative way to visualize the underlying structure of a data set. It shows which clusters are well separated and whether cluster borders are smooth. It also allows visual judgement of the number of clusters.

Table 1 shows some of the methods recently used for microarray data analysis. Among them, hierarchical clustering is most frequently applied.

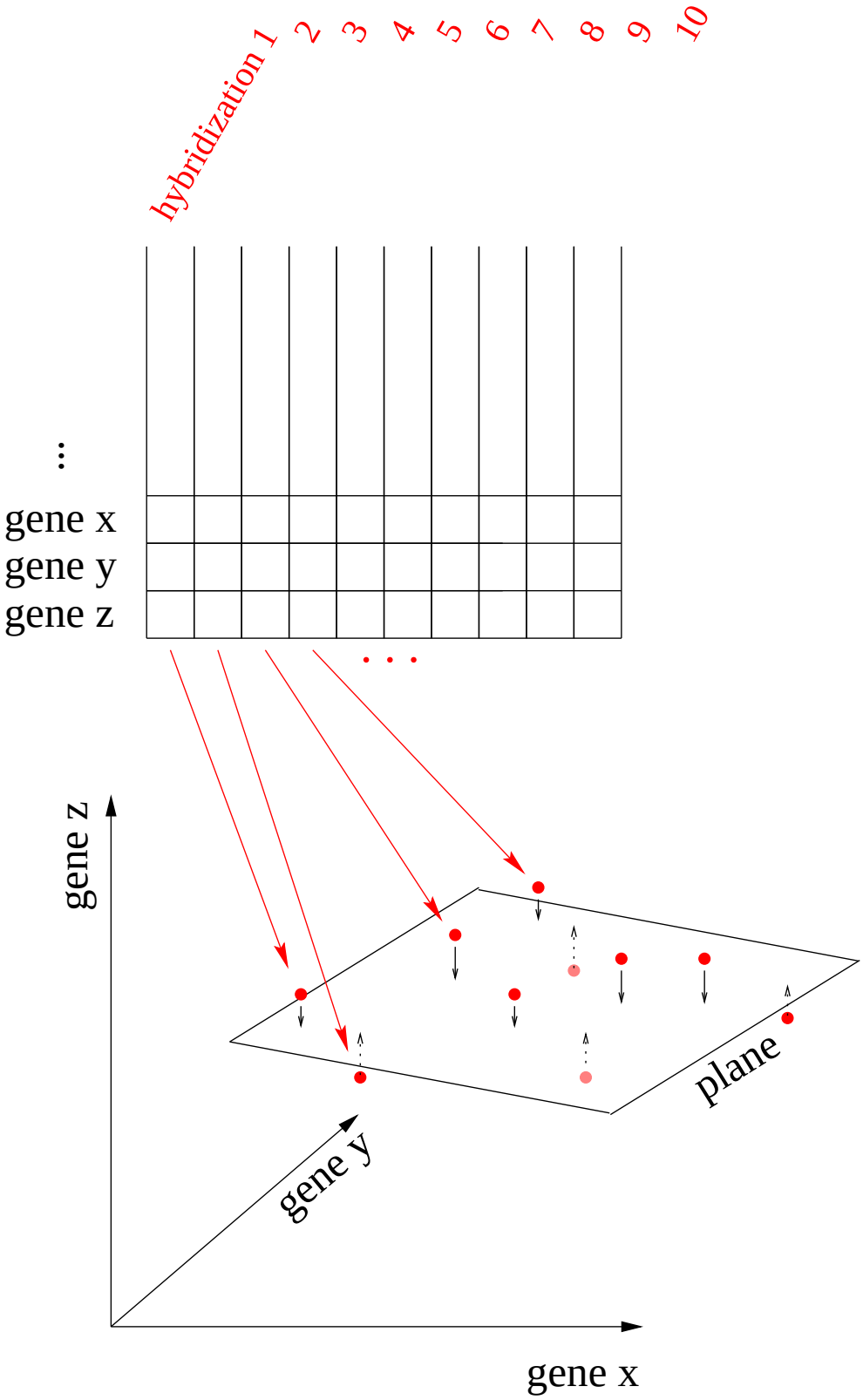

Figure 3: **Planar embedding.** The m columns of a table of n genes $\times$ m hybridizations are represented in n-dimensional gene space (three dimensions are shown). n ranges from a few hundred to tenths of thousands. Most microarrays comprise several thousand elements. A plane is selected such that 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.

Method	Output	Reference
Classification		
Weighted voting	Classifier	[11]
CART	Tree	[25, 26]
Support vector machines	Classifier	[24, 33]
Artificial neural networks (ANN)	Classifier	[34]
k-nearest neighbors	Classifier	[35]
ISIS	Bipartitions	[36]
Bayesian regression	Classifier	[37]
Clustering		
Hierarchical clustering	Tree	[28]
k-means clustering	Set of clusters	[27]
Clustering affinity search technique (CAST)	Set of clusters	[38]
Kohonen maps	Set of clusters	[29]
Cluster identification via connectivity kernels (CLICK)	Set of clusters	[39]
biclustering	Set of clusters	[40]
Gene shaving	Set of clusters	[41]
Planar embedding (projection)	2D- or 3D-projection plot	
Multidimensional scaling		[42]
Principal components analysis		[32]
Singular value decomposition		[43]
Other methods		
REVEAL	Directed graph	[44]
Bayesian networks	Directed graph	[45]

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

Interaction of storage and analysis

Statistical thinking, while being largely dispensible for small-scale experimental settings, is necessary for both the design and interpretation of microarray experiments [46]. Large amounts of data as well as high levels of detail in data annotation render visual inspection of these data impractical. They have to be investigated by automated computer-aided analysis. This emphasizes the need for computer-readable data as well as for appropriate methods for integrated analysis.

Current microarray database projects focus on the integration of multiple platforms and various fields of microarray experiments by means of flexible storage of transcription intensities and consistent, often hierarchical sample description. However, most of the valuable information contained in experiment annotation is currently not taken into account for analysis. This is due to the fact that the annotations are stored in a way not readily accessible for statistical methods. Frequencies of annotation values, e.g. within a set of experiments clustered by their expression patterns, ought to be countable.

Figure 4 shows a simple way to determine the association of transcription patterns with experimental parameters. The graph shows a virtual data set. It depicts a planar projection (e.g. a MDS or PCA plot) of 48 microarray hybridizations. 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 results in 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 computer-based 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. While this is a fairly simple method, it already provides good analytical access to long lists of annotations and huge sets of hybridizations, which could not be thoroughly evaluated by visual inspection. More sophisticated multivariate statistics can also be applied. However, any statistical analysis will require countablility of annotated 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.

Another problem interfering with in-depth analysis of experiment annotations is the need for reduction of the annotated parameters to achieve universal applicability. While it is suitable

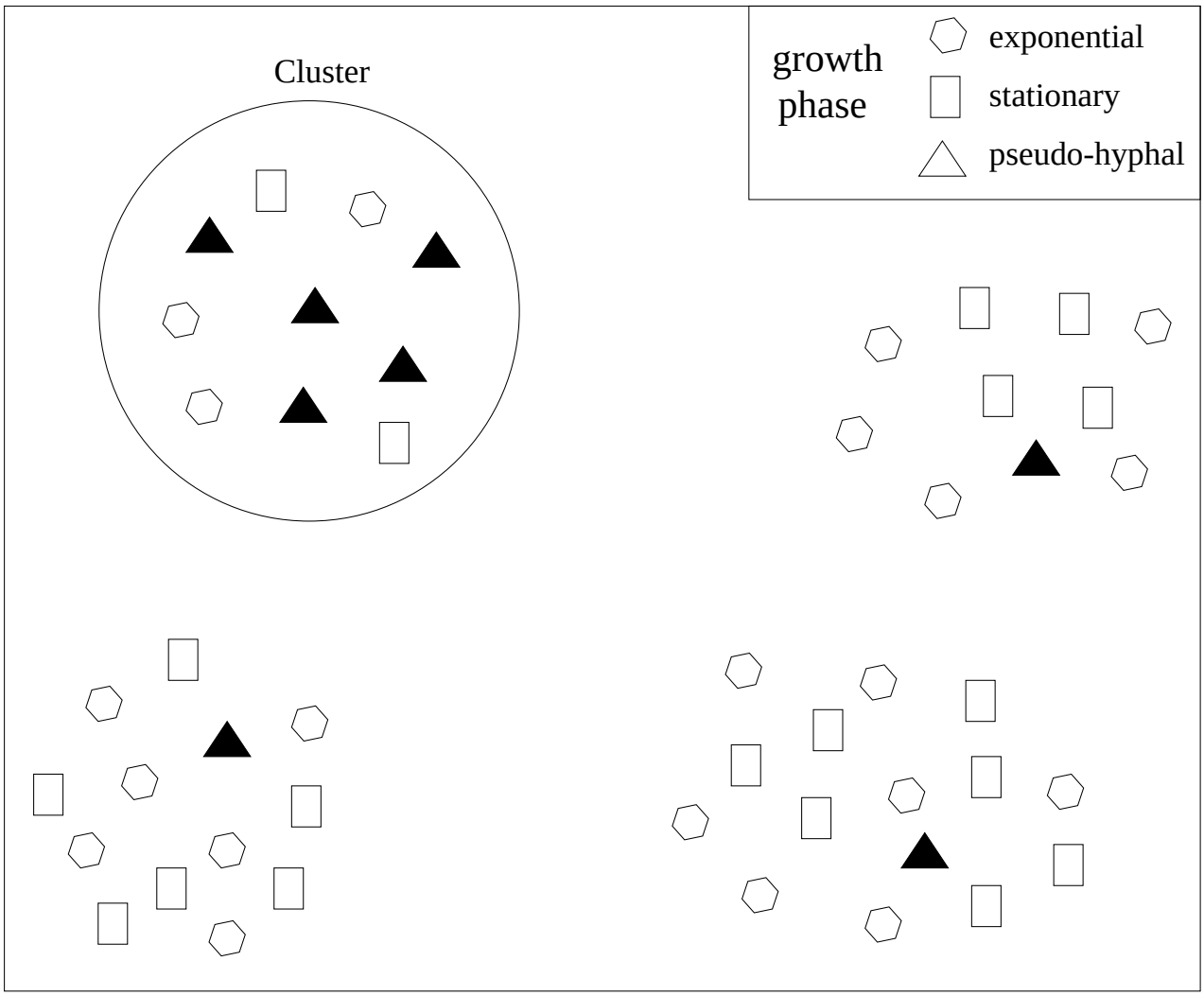

Figure 4: **Correlation between transcription pattern and experimental parameter.** The scheme — a planar projection (Fig. 3) of virtual data — sketches a simple method to determine frequencies of annotation values for a cluster of hybridizations under study. It is explained in the text. The dimensions of the axes correspond to the dimensions of the axes of the original space. Triangles are filled to facilitate counting. All values of all annotations can be scanned for being characteristic, i.e. over- or underrepresented in a hybridization cluster, thus enabling computer-based analysis of large and complex data sets. However, these values have to be countable in order to determine their frequencies inside and outside the cluster.

and important to annotate the composition of culture medium for yeast, this annotation has no meaning for cancer biopsies. Vice versa, parameters such as tumor stage or grade are meaningless in the context of yeast experiments. Public data repositories have to take experiments from very different fields of biological research. They focus on minimal standards such as MIAME[4], i.e. sets of annotations that are relevant for many different kinds of experiments. However, mimimal-standard annotation inevitably records an experiment at a low level of detail. Experimental settings may not be sufficiently described to detect artifacts or biologically relevant but unexpected transcriptional responses to subtle experimental variations.

A third obstacle to the integrated analysis of hybridization data and its annotation is the fact that most of the recently applied statistical methods are not able to properly visualize more than one variable at the same time. Hierarchical clustering is widely accepted for the analysis of microarray data. Genes and experiments are often displayed as a color-coded table of hierarchically clustered rows and columns. This kind of visualization takes a lot of space — often filling whole pages to display no more than 800 genes [47] — making it difficult to trace columns. While it reflects associations among the genes and also among hybridizations, it is easily outcompeted in revealing associations between genes and hybridizations by a method designed for analysis of interdependencies between two variables, as I will show in example.

Contributions

I have developed a relational storage concept capable of keeping the entirety of available information about microarray experiments in a form ready for statistical analysis. The concept both ensures consistency of annotations and circumvents the difficulties of free-text parsing. It is designed to permit annotation at any level of detail chosen by the collaborating biologists. With the exception of numerical values, my annotation system is entirely categorical, allowing a choice only between predefined enumeration-type values which can then be readily analyzed in an automated fashion.

Annotation storage should be flexible enough to allow easy inclusion of new attributes as well as new values. I store the definitions for the annotations and their allowed values as separate tables in the database, thus avoiding a fixed, 'hard-wired' structure that would be difficult to extend. Storing these definitions as table content rather than attributes enables them to be extended without changing the database structure and without adjustment of the analyzing algorithms.

Here I present a storage and analysis concept called M-CHIPS (multi-conditional hybridization intensity processing system). It has been implemented as a set of organism-specific databases, namely for *Saccharomyces cerevisiae*, *Arabidopsis thaliana, Trypanosoma bruceï, Neurospora*

[4] `http://www.mged.org/Workgroups/MIAME/miame.html`

crassa and human tumor samples. While differing in the annotations used to describe the samples, these databases share a common structure and thus are accessed by the very same analysis algorithms. The concept is able to integrate all types of intensity data obtained from cDNA microarrays. It has been tailored to the needs of the collaborating groups which use cDNA microarrays with either single-channel radioactive or multichannel fluorescence readout.

Furthermore I have incorporated a method particularily suitable for the detection of interdependencies not only among one set of variables such as the genes but also between any gene and hybridization under study. Correspondence analysis (CA) is an exploratory technique, allowing to visualize structures within the data and thus revealing which questions could be asked or which hypotheses could be put forward. Unlike many other methods, CA does not require any prior choice of parameters.

Like other projection methods, CA represents variables such as transcription intensities of genes as vectors in a high dimensional space. In our case the dimensionality of the space would be the number of hybridizations involved. Both PCA and CA reveal the principal axes of this high dimensional space, enabling projection into a subspace of low dimensionality that accounts for the main variance in the data. Unlike PCA, CA is able to account for the genes in hybridization-dimensional space and the hybridizations in gene-dimensional space at the same time. Both representations of the data matrix will be projected into the same low dimensional subspace, for example a plane (yielding a so-called 'biplot'), revealing associations both within and between these two variables.

I will show the general applicability of CA to microarray data analysis in examples, both with radioactive and multichannel data.

Data Storage

Data analysis is intimately linked with data storage. Terms like 'data warehousing' reflect this. Data warehousing aims at providing data in a format suitable for analysis. M-CHIPS slightly differs from a classical data warehouse solution, in which data are held in one or several so-called 'operational' databases. A warehouse then collects data from these databases mainly used for storage and makes them fit into a unified data model [48, 49]. Typically, a warehouse will collect only a few, 'important' attributes from each dataset. Operations such as extractions and transformations are recorded as meta data. The warehouse may be denormalized, i.e. it allows for redundancy in order to avoid frequent joining from distinct tables.

Designed to assist analytical tasks rather than pure data storage, M-CHIPS may be considered a data warehouse. It integrates different data sources and data formats into a denormalized structure, records meta data and enables unified access for analysis algorithms. However, there are no underlying operational databases, and data are directly entered into M-CHIPS. Thus, analysis may be carried out immediately, enabling instant decisions about follow-up experiments. There is also no loss of information in experiment description. Annotations are not extracted by compliance to minimal standards, but entered directly at a level of detail chosen by the experimenter defining the annotations. All annotations are in an analyzable form that avoids text mining, which frequently results in a loss of information.

Database model

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.

A **database query** retrieves from a specified relation a set of tuples that fulfill a certain condition (also specified in the query). The purpose of an **index** is to allow rapid access to specified values within a relation. Without an index, the server process executing a query has to read from the beginning of the table to the end, looking for relevant tuples. An index is computed on a relation for one (or several) of its attributes, ordering the contained values and storing pointers into the relation in an appropriate data structure (e.g. a b-tree or hash table). A query is directed to the matching rows by the index, and thus carried out much faster than without an index.

The PostgreSQL DBMS allows for construction of relation hierarchies by **table inheritance**[5]. Let the table B be created inheriting from table A, then

- B inherits all attributes of A.
- All tuples of B are accessible by querying the parental table A
- by anyone having read permission on A regardless of permissions on B.

Another important issue is **data integrity** or **data consistency**. Suppose a valid alteration of the data, defined by a block of sequentially permformed 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.

[5]The relational database model does not comprise table inheritance. This feature represents an object-relational extension provided by the DBMS. In the world of objects, database relations are represented as **classes**, attributes are also referred to as **slots** (AI term), tuples as **instances**, **individuals** or **objects**.

Design requirements

The data to store can be divided into raw transcription intensities, gene annotations and experiment annotations. The last have the most complex structure among these.

Gene annotations

Gene annotations may consist of clone numbers, accession numbers and heterogeneous information such as chromosomal location, enzyme categorization number or structure of the encoded protein. Since the only unique identifier for a spotted DNA fragment is its sequence, the most important information is a link to a sequence database which also holds the additional information. Furthermore the possibility of dividing the gene set into partitions should be provided. This information is necessary for separate normalization of certain sets of spots, e.g. when they have been hybridized separately.

Transcription intensities

Raw intensity data rather than processed values should be stored because processing algorithms change rapidly. Currently, image analysis itself cannot be carried out without human interaction, thus separating image data from automated analysis. Therefore, analysis should start with raw signal intensities performing processing steps like normalization and filtering on the fly. A hybridization yields a simple although huge list of intensities and background values for every spot on the array. These could, in principle, be stored in records or in so-called 'binary large objects' inside, or even in flat file format outside the database. However, it would not be possible to select subsets of data fulfilling criteria such as intensity thresholds or to perform simple calculations on database level. Such calculations may be necessary in future in order to normalize huge datasets and to extract from the normalized data when they do not fit into computer memory, suggesting storage of intensity data in database tables. The system should be flexible enough to store intensities stemming from both monochannel (radioactive label) or multichannel (fluorescent label) hybridizations. Signal intensities obtained by radioactive labeling do not represent the same quantities as those reflecting competition of differently labeled hybridizing cDNA populations. For the former, absolute signal intensities should be proportional to the amount of mRNA molecules in the target. For the latter, low intensity for a particular channel may result either from low mRNA concentration for this channel or from the fact that the binding sites on the array are occupied by high amounts of differently labeled mRNA, to give an example. Preprocessing algorithms should be able to recognize the difference and automatically apply suitable methods, e.g. for normalization.

Experiment annotations

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. They fall into two classes: First, there are common annotations that are useful for all fields of interest. These are technical annotations such as array characteristics, descriptions of labeling, hybridization or washing conditions, and of signal detection. This set of annotations should be the same for all kinds of microarray experiments. Second, there are organism-specific annotations that meet the differential requirements of the specific research areas such as 'transgene' and 'growth phase' for yeast or 'tumor type' and 'metastasis location' for tumor samples. Both common annotations and multiple organism-specific annotation sets should be stored in a unified structure so that they can be annotated and queried by the same algorithms. Otherwise, algorithmic efforts would not be feasible for many different kinds of microarray experiments.

All experiment descriptions should be directly accessible to statistical analysis. This can easily be achieved when data are not entered as free text but in a categorized, queryable form. This allows for application of multivariate procedures for correlating expression data and annotations.

To make all experiment descriptions directly accessible to statistical analysis, I permit only two types of experiment annotations, either numbers of predefined unit or values from predefined lists. For example, if we let an annotation 'growth phase' be an enumeration-type variable comprising the defined values 'exponential', 'stationary' and 'pseudo-hyphal' (see Fig. 4), the occurrence of any value can be counted within a set of hybridizations clustered by their expression profiles and compared with its overall frequency to determine whether it is characteristic, i.e. either over- or underrepresented in the cluster.

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.

Database implementation

Here I will sketch how these concepts have been implemented in our databases. A technical report on the associated web page[6] lists all technical details. Figure 5 shows the main components and the way in which they are related.

The data categories mentioned above, namely gene annotations, raw signal intensities, and experiment annotations, were taken as a basis for implementation. The corresponding items are displayed in blue, yellow and red, respectively, in Figs. 5 and 6. The gene annotations are linked both with the expression intensities and with public external gene databases in order to enable explicit characterization of genes showing a particular expression behavior. The expression intensities are stored as measurements. A measurement 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. The annotation relations themselves contain values that have been defined in the definition relations.

The following subsections will go into more detail about the implementation, dealing with the main components one by one. Fig. 6 maintains the arrangement and color code of Fig. 5, dissolving the overview into database relations and their attributes. According to the Unified Modeling Language (UML) specifications[7] 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 [51], 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

[6] http://www.dkfz.de/tbi/services/mchips
[7] http://www.omg.org/technology/documents/formal/uml.htm

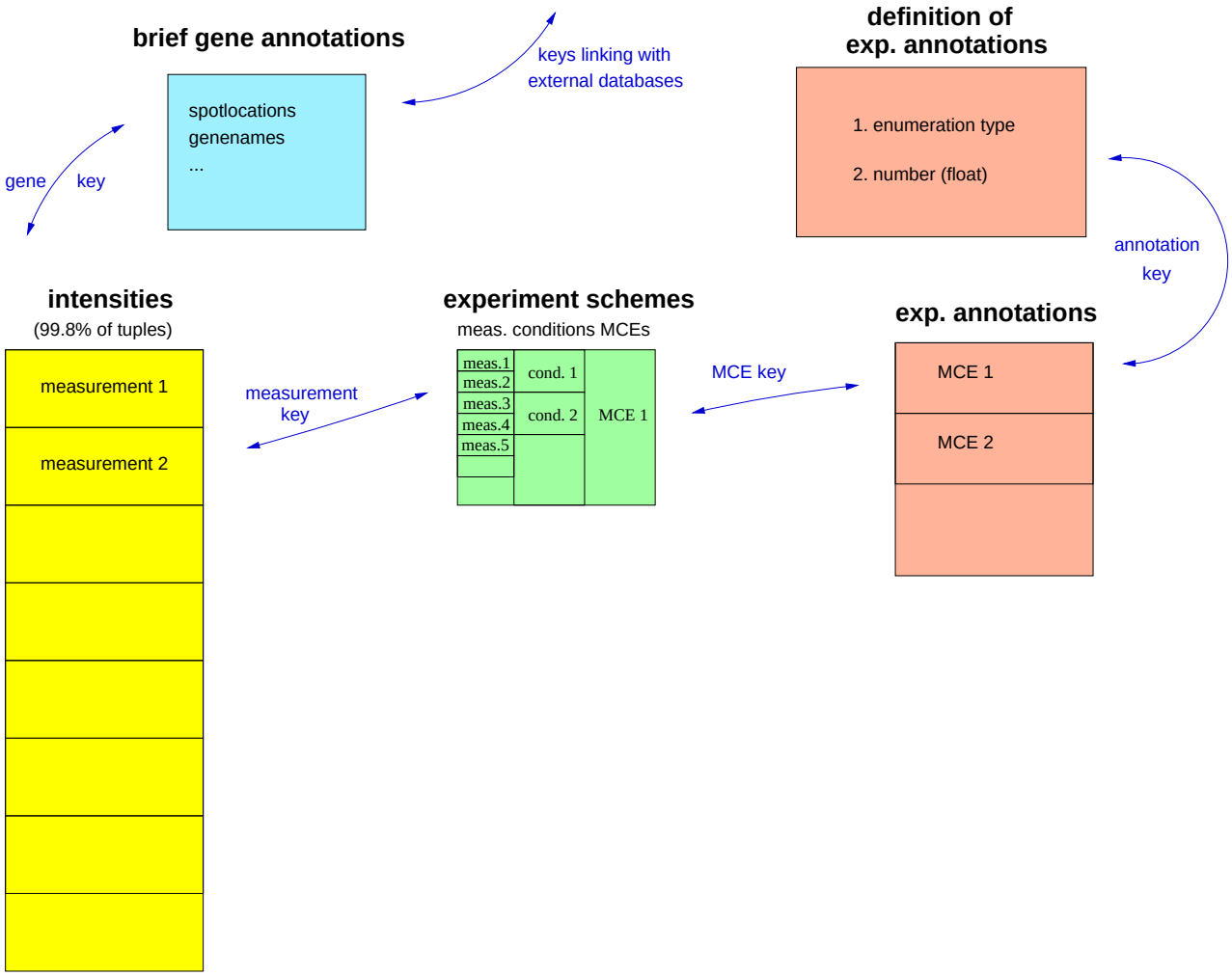

Figure 5: **Overview.** The majority of the data (in terms of both storage space and number of tuples) consist of transcription intensities (yellow). The tuples storing these data are related to the tuples of the gene annotation table ('brief gene annotations', blue), which link them to external gene databases. They are also related to measurements (meas.), experimental conditions (cond.) and multiconditional experiments (MCEs). These relations are further characterized by experiment schemes (green). One MCE is described by many experiment annotations, allowed values of which are stored as definition of experiment annotations (red).

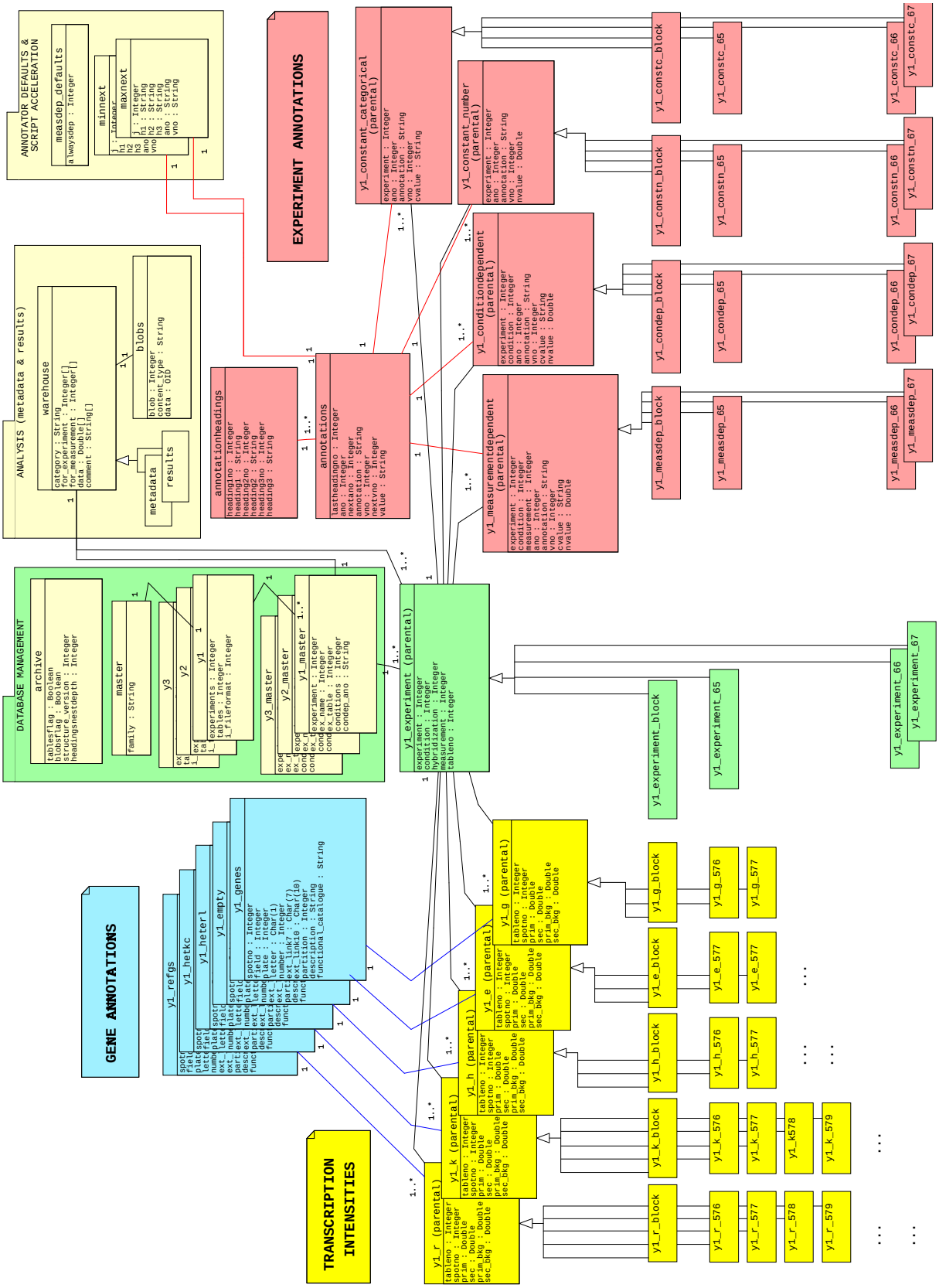

Figure 6: **UML scheme of an M-CHIPS database.** The tables are arranged and color coded according to the categories introduced in Fig. 5 and explained in the text. Overlapping tables show identical structure. Arrows indicate table inheritance. The scheme is further explained in the text. From [50].

Table = y1_genes

Field	Type	Length
spotno	int4	4
field	int4	4
plate	int4	4
letter	char()	1
number	int4	4
ext_link7	char()	7
ext_link10	char()	10
partition	int4	4
description	text	var
functional_catalogue	text	var

Table 2: Gene annotations (table structure)

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.

Gene annotations

A database may comprise several types of microarrays differing in their sets of immobilized DNA fragments. I refer to the data stemming from one type of microarray as an 'array family'. Each family thus consists of all multiconditional experiments carried out with the same microarray spotting scheme. A microarray family is endowed with its own set of gene annotations that reflect this spotting scheme. The gene annotations are linked both with the expression intensities and with public external gene databases in order to enable explicit characterization of genes showing a particular expression behavior.

Table 2 shows the structure of a gene annotation table (see also Fig. 6). The attribute 'spotno' serves as a key connecting to the tables which contain hybridization intensities. 'Field', 'plate', 'letter' and 'number' correspond to the spot location on the array as well as to the DNA stock kept in microtiter plates. Two fields of fixed length ('ext_link7' and 'ext_link10') are reserved for keys linking to external databases. Certain spotsets may have to be normalized separately, as explained in 'Design requirements'. The partition of the spots is recorded by the attribute 'partition'. 'Description' and 'functional_catalogue' are of variable format and size. They contain a brief description of the encoded protein and its functional category.

Explicit categorization of genes is available in the form of e.g., organism-specific gene ontology databases (GO consortium[8], [52, 53]) . They assign a gene to one or more categories ('terms'), which themselves are hierarchically ordered in a tree. In order to make this information available to statistical analysis, I cut the tree at a level where the subtrees rooted by the according nodes each contain a sufficient number of genes. If the number of genes in a chosen category was low (say two), a notable effect, e.g. that a considerable share (all) of the comprised genes

[8] http://www.geneontology.org

are highly associated with a certain experimental condition, could occur by chance with high probability. For statistical analysis, the information is required in a form enabling counting of instances of occurence for each category rather than reflecting the hierarchy. I therefore extract the gene categories at the chosen level of detail into the column 'functional_catalogue' of the gene annotation table, recording for each gene the associated category. One gene may also belong to more than one category. In that case the categories listed are separated by semicolons.

Transcription intensities and query performance

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 (see according one-to-many relations in Fig. 6). A measurement may comprise a hybridization in case of monochannel experiments, or a single channel of a multichannel hybridization. Experiment schemes (Fig. 6, green tables) record for each measurement to which hybridization and experimental condition it belongs, and in which multiconditional experiment (MCE) this condition is contained. 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 I use different tables to store these kinds of intensities, thus minimizing query space (Fig. 6, yellow tables). Their names include 'g' for genes and 'e' for empty spots (i.e. no DNA was spotted on this location). The controls are marked by 'h' for heterologous DNA in contrast to 'k' for heterologous DNA with *known* concentration. While both spot categories comprise heterologous DNA - that is DNA stemming from species other than the one under study - and therefore in both cases the target hybridizing to these probes has to be explicitly added to the hybridization, the difference lies in the target concentration. For the first category this concentration is not known. Therefore, such spots, e.g. so-called 'guide spots' or 'landing lights' marking anchor points of the grid and thus guiding the imaging process, cannot be used for normalization. In contrast, for heterologous DNA with known target concentration, also called 'external controls', defined concentrations of artificially made RNA are added to the sample before reverse transcription. This measure, called 'spiking' accounts for different label incorporation rates and other technically caused systematic errors by providing standard signals for normalization. The last spot category, marked 'r' for reference, is reserved for future use with different and novel kinds of controls.

The tables of all categories show the same structure (Tab. 3). The first attribute holds the ID of the stored measurement in family context. 'Spotno' identifies the spot, corresponding

Table = y1_g_589

Field	Type	Length
tableno	int4	4
spotno	int4	4
prim	float8	8
sec	float8	8
prim_bkg	float8	8
sec_bkg	float8	8

Indices: y1_g_589_ipr
y1_g_589_ise
y1_g_589_isn

Table 3: Transcription intensities (table structure)

to the identically named attribute of the gene annotation table 'y1_genes' (Tab. 2). In the tables 'y1_e_589', 'y1_h_589', 'y1_k_589' and 'y1_r_589' this attribute corresponds to 'spotno' in 'y1_empty', 'y1_hetrl', 'y1_hetkc' and 'y1_refgs', respectively. The remaining attributes contain the hybridization intensities. Each gene or EST has been spotted in duplicate resulting in two intensities ('prim' and 'sec') per hybridisation. The last two attributes are intended to take a local background value which is determined by many imaging software packages.

M-CHIPS so far contains data obtained with Bioimage^{TM}, Xdigitize^{TM}, AIS^{TM} and Genepix^{TM}. Some imaging software packages (e.g. Genepix) yield more than one intensity and background value per spot such as differently calculated intensities (e.g. pixel mean, median), different background intensities and various kinds of quality or reliability measures. From these, the contents of the above tables are either chosen or calculated as a starting point for standardized analysis in the process of database upload. The objective is to store a number proportional to the amount of hybridized label. This would be reflected by the integrated amount of signal measured for one spot. Delivered with median and mean pixel intensity as well as with the number of pixels per spot, I pick mean $\times$ number of pixels for database storage to resemble this amount as closely as possible. The intensity is often not evenly distributed within the spot area. It tends to be low in the spot center with a concentric rim of high intensity between center and margin. The second problem is contamination by highly fluorescent dust. Favoring the mean over the median in the product, I account for the first problem at the expense of intensity levels elevated by dust. The reason for this is that non-evenly distributed staining is much more frequent and thus more severe than dust contaminations.

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. In Tab. 3 they are listed below the attributes of the table. The three indices belong to, i.e. direct search within, the attributes 'prim', 'sec' and 'spotno', respectively, as abbreviated by the last two characters of their names. 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, I 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 (Fig. 6, 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*.

Experiment annotations — manifold variables under constant extension

To achieve direct access for statistical methods, all experiment descriptions have been dissected into atomic items that can be represented by either numbers of predefined unit or values from predefined lists. To meet the flexibility requirements described above, the annotations are contained in tables rather than being implemented in the database structure itself. The web-based annotation process involves reading these definition tables and recording the entered numbers or selected values in annotation tables.

Definition tables. A separate database is maintained for each organism or field of research which contains particular definitions of experiment annotations appropriate for the investigated samples. Annotation definitions for *S. cerevisiae*, *A. thaliana*, human tumor biopsies, *T. bruceï* and *N. crassa* are provided on the M-CHIPS web page[9]. Each database comes with a certain set of experiment-annotation definitions that are 'organism-specific'. However, some, mostly hybridization-protocol related, 'common' annotations are used in all databases. To facilitate inter-field analyses for the future, this portion has to be kept as large as possible. New

[9]http://www.dkfz.de/tbi/services/mchips

Table = annotationheadings

Field	Type	Length
heading1no	int4	4
heading1	text	var
heading2no	int4	4
heading2	text	var
heading3no	int4	4
heading3	text	var

Table = annotations

Field	Type	Length
lastheadingno	int4	4
ano	int4	4
nextano	int4	4
annotation	text	var
vno	int4	4
nextvno	int4	4
value	text	var

Table 4: Experiment annotations (table structure)

common annotations are added to all databases automatically by means of administration scripts. Each annotation has a unique identification number. They are stored as a linked list including an attribute pointing to the ID of the annotation next in sequence. This structure enables adding of annotations at arbitrary positions by linking the desired ancestor to a new element that points to the ID of the element following in that list. In a similar manner the whole set of defined values is stored by a second linked list within the same relation. Hierarchical structure of annotation sets used for experiment description is recorded by the content of a second relation. Although the system has not been tested with other than three levels, the nesting depth has been implemented to be arbitrary. Table 4 provides the structure for both relations. As explained above, it does not show any annotations. These can be found exclusively in the contents of such relations. Figure 7 provides an example, listing the first part of the common annotations.

The structure of both tables is denormalized for visual clarity. Since the normalized form, consisting of separated tables, would exclusively be queried by joining them, I have directly implemented the joins as database tables. Such redundancies, though not common for databases, are frequently used in data warehousing. The contents of these tables are used as meta data by the web-based user interface to compile multiple-choice forms (Fig. 8). The results of the annotation process are stored in annotation tables.

Annotation tables. A multiconditional experiment consists of at least two different experimental conditions, differing e.g. in growth conditions, tissue type or genotype of the biological material under study. Each of these conditions comprises several repeated measurements. Such a measurement may represent a hybridization in case of radioactively labeled targets or a single channel of a multichannel fluorescence signal. The experiment schemes storing which

```
yeast=> select * from annotationheadings order by heading1no, heading2no, heading3no;
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|-
         1|common_annotations           |         4|submission       |         8|-
         2|organism_specific_annotations|         5|genotype         |         9|-
(...)


yeast=> select * from annotations order by lastheadingno, ano, vno;
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
           1|   5|      6|spotted_material                   |  19|     20|DNA-oligo
           1|   5|      6|spotted_material                   |  20|     21|PNA-oligo
           1|   6|      7|readfile                           |   0|      0|[]
           1|   7|      8|array_hybridisation                |   0|      0|[]
           2|   8|      9|material_source                    |  21|     22|fresh
           2|   8|      9|material_source                    |  22|     23|frozen
(...)
```

Figure 7: **Example for experiment annotation definitions (table content).** Two SQL statements are listed along with the first few rows of their results. The first one shows the content of a table named annotationheadings (topmost red in Fig. 6). These headings serve to hierarchically structure the annotations into sections. This table is linked to the second one through 'heading3no', here named 'lastheadingno', because the nesting depth is arbitrary and may be decreased or increased in other databases. The annotations are stored in the second table (Fig. 6, 'annotations') along with their allowed values. The attributes 'ano' and 'vno' are used as IDs to reference annotations or their values, respectively, as described above. The attributes 'nextano' and 'nextvno' point to the next entry, thus implementing the 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. From [50].

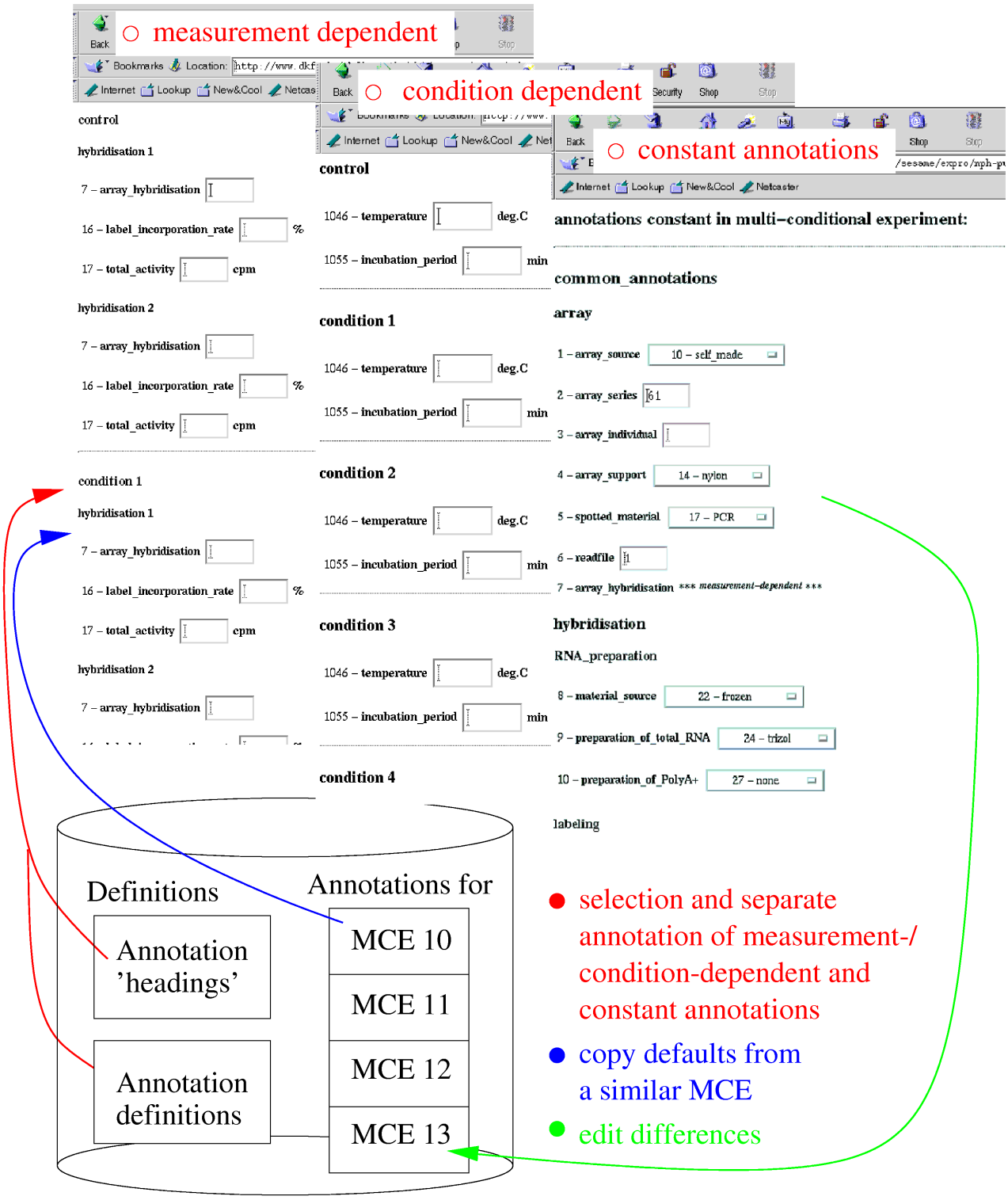

Figure 8: **Experiment annotation process.** The annotation process may start with copying default values from the most similar multiconditional experiment (MCE). Secondly, from the complete list of defined annotations the measurement-dependent ones are selected and then annotated for each single measurement. Afterwards, from the remaining annotations, those being condition-dependent 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. The HTML form for the constant annotations was compiled from the annotation definitions shown in Fig. 7. From [50].

Table = y1_constant_categoricalvalue_65

Field	Type	Length
experiment	int4	4
ano	int4	4
annotation	text	var
vno	int4	4
cvalue	text	var

Table = y1_constant_number_65

Field	Type	Length
experiment	int4	4
ano	int4	4
annotation	text	var
vno	int4	4
nvalue	float8	8

Table 5: Constant experiment annotations (table structure)

of these measurements belong to which experimental condition also record which of them were performed simultaneously on the same array. Most of the experiment conditions are *constant* for an entire experiment, some are *condition-dependent* or *measurement-dependent*, i.e. they can take different values for each condition or measurement. This gives the designer a choice of storing the annotations either according to these three categories or measurement-wise. While data import by the user is easier when following the first scheme, the latter is preferable for statistical analysis. Following the first scheme, I decided to store the three sets in separate annotation tables for convenient algorithmical handling (Fig. 6, red tables, names beginning with 'y1'). Merging the tables for each measurement is easy, whereas splitting up measurement-wise stored annotations would require repeated value comparison. Table 5 displays the structure of annotation tables storing the annotated values that are constant throughout the experiment. Categorical, i.e. enumeration type annotations and numbers are stored in separate tables. The latter may be of either categorical or of continous range. The two types of annotation are reflected by the type of the attributes 'cvalue' and 'nvalue'. This is the only difference in the structure of the two tables.

As a representative of intended redundancy both number ('ano') and name ('annotation') are listed for an annotation as well as for its value. For the small data volume of the annotations (see 8 above) this does not have major consequences for storage space or for performance. However the redundancy might serve to reconstruct experimental annotations if an error occurs e.g. in numbering of annotations or values. Redundant storage appears advisable here because annotation definitions are under constant change as new kinds of experiments evolve.

The tables storing condition- and measurement dependent annotations comprise condition and measurement as additional attributes and integrate categorical values and numbers. Recorded content - sorted by categories condition-dependent, measurement-dependent and constant -

can be found e.g. at `http://mips.gsf.de/proj/eurofan/eurofan_2/b2`, linked in the 'experimental condition' - column.

The experiment annotations share the feature of writing separate tables, that are assembled into one large block later as discussed for the transcription intensities. Fig. 6 shows four table sets differing in the structure inherited from their parental tables, two containing constant and one containing condition- and measurement dependent annotations, each. In contrast to the transcription intensities, which are written by separate tables for each measurement, the annotations are stored experiment-wise.

Database management

Having introduced the parts of M-CHIPS, I will now complete the picture. The tables in the 'ANALYSIS' box of Fig. 6 (beige) hold metadata (e.g. normalization parameters) and analysis results, which can be stored back into the database. At this step, the results are automatically made available via WWW, however protected by individual passwords for each array family (i.e. type of array) within a certain (organism-specific) database. Links to those results that are publicly available are given in Appendix C, which also discusses the presentation format. In the database, those results are stored as binary large objects (BLOBs). These are unstructured bitstreams as opposed to the structured tables. The topmost table in the green 'DATABASE MANAGEMENT' box is named 'archive'. Its first two attributes flag whether any table or binary large object (BLOB) in the database has been altered. This information is used by an overnight backup mentioned below. The fourth table holds the nesting depth of the headings, that hierarchically structure the annotations, the third refers to the structure version of the database. Altering the database structure is always time-consuming, because all accessing algorithms have to be adapted. Apart from version no. 2, which is described here, there are older databases of version 1, which can deal with single-channel data only. Because data transformation into the new structure is error-prone, data have been kept unchanged. Storage and analysis algorithms have been adapted such that they can handle both versions.

It should be noticed, that for each of the relations discussed above the relation name starts with 'y1'. Figure 6 exemplifies M-CHIPS storage rather than providing a complete formal representation. The example database comprises three different families, i.e. experiments carried out on three different types of microarrays. The remaining database management tables (in green box) show the array families y1, y2 and y3 (e.g. referring to three different types of yeast arrays). Outside the green box, all tables belonging to a particular array family are represented by those belonging to the first one, names starting with y1. The other families comprise identically structured tables which are not shown here. In fact, of the tables already discussed, only 'annotationheadings' and 'annotations' do not belong to a particular array family. The contained annotation definitions are valid for any experiment of the particular

field of research, i.e. throughout the database. The number of tables within an M-CHIPS database thus varies with the number of comprised array families. It also depends on how many uploaded measurements have already been assembled into a block table. M-CHIPS operates on a unified storage concept for standardized algorithmical access rather than on a fixed database structure.

Methods for database operation

M-CHIPS allows for unified analytical access to microarray experiments from different fields of research. An instance of the above described concept of storage is a field-specific (i.e. organism-specific) database. Such a database is charged and queried by the experimenters themselves using algorithms which mediate upload and annotation of experiments, as well as data analysis. Although different databases adopt different parameter sets for experiment annotation, they are accessed by the very same algorithms.

M-CHIPS consists of C, Perl and MATLAB functions. The system is currently available for the SUN-Solaris and HPUX platform. Gene annotations are stored during the process of generating a new database or a new microarray family within an existing database. While up to now this is done by the database administrator, upload of both transcription intensities and annotation of experiments are done by users.

Experiment Annotation

Experiments can be annotated remotely by the experimenters using a web-based interface. Annotation appears to be a time-consuming process, if hundreds of experimental parameters have to be entered for each single measurement. For this reason, I provide the possibility to select annotations that are *constant* or *condition-dependent* as defined above and that have to be entered only once, in contrast to *measurement-dependent* annotations. Furthermore, it is possible to copy the whole set of annotations from a similar experiment and edit only those that differ. Few parameters should be varied per condition, so the majority of the annotations are constant throughout the experiment. Among these, the majority are constant not only for one particular experiment, reflecting more or less constant execution of the same protocols for e.g. hybridization and washing. The annotation process is sketched in Fig. 8. It avoids redundant data entry, such that it is possible to enter detailed descriptions (111 annotations) of large multiconditional experiments (24 measurements) in less than 15 minutes.

Upload of transcription intensities

After image processing, the output files of the imaging software are uploaded. At first the MCE is defined by entering information about

- how many conditions are comprised by the MCE,
- how many measurements are available for each condition,
- whether or not the MCE to upload is a multichannel experiment,
- which of these measurements are channels belonging to the same hybridization (only for multichannel experiments), and
- path and filenames of the files containing the intensities of the measurements to upload.

If the experiment has been already annotated before upload, the first two items are already present. After selecting the annotated MCE, hybridizations and imaging output files are assigned to the corresponding conditions and measurements already defined in the database.

To avoid wrongly assigned input, all files are thoroughly tested by the routine before import. Every spot provided in such a file is checked for being a member of the particular array type which has been chosen for that experiment.

Safety aspects of database operation

Algorithms for storage and analysis are designed to be used by the people who generate the data. To meet the requirements of the users, they have to allow for multiuser access including safe management of simultaneous write access, short waiting periods and protection against unauthorized access. In principle, consistency and security of the data are threatened by global accidents, private errors and unauthorized access.

Global accidents are avoided by the ability of the DBMS to handle transactions, or they may be reverted using global backups. A transaction gives the database an all-or-nothing capability when making modifications. It can comprise one or multiple queries with each of the performed changes becoming valid upon successful execution of the whole transaction and none of them in case of an error. At the same time all other users are prevented from seeing the partially committed transaction until it has been successfully finished, preventing database consistency from being damaged by simultaneous write access. Although transaction-based database management slows down performance, it would be unwise not to use a transaction-based DBMS in a multiuser scenario. M-CHIPS has been operational for two and a half years

without any global error. Nevertheless, the data are constantly recorded by over-night tape backup to prepare for such an accident.

Private errors differ from global accidents in effect and measures taken for repair. In case of accidentally deleting hybridizations from a single database it would be inappropriate to reset the whole system to the state of the night before. To be prepared for such a case, SQL (structured query language) dumps are performed separately for each database as part of a nightly performed process. SQL is a common standardized language for database queries. Such dumps consist of SQL statements that can be used to restore data subsets from a whole database down to a single tuple of a particular table. The over-night process involved also tests and reports important consistency and status parameters of all databases by e-mail.

Unauthorized access is prevented by password authentication. To ensure that data (which may be unpublished) cannot be altered or read by unauthorized individuals, update and/or read permissions can be granted on any database table to a particular user. Granting such permissions to user groups rather than separately to each user is a common procedure to circumvent the necessity of changing permissions for each database table upon registration of a new user. In my implementation nearly all the relations inherit from few parental tables and are accessed via their parental table only. Permission inheritance enables the administrator to quickly grant e.g. read access to a new user by changing permissions for a few parental tables in place of dealing with many tables or user groups. However, the main reason for access via parental tables is to enable pooling of tuples from hybridisation tables into large blocks without syntax alteration of accessing queries (compare 'Transcription intensities' above). Authentication of the user is mostly taken care of by the operating system (OS) that is used for running M-CHIPS, namely Solaris or HPUX. However, if access is conducted via WWW, the OS account of any accessing user is the one of the webserver. In these cases, the user is asked to provide an external username along with two passwords. The first password is requested by the APACHE webserver and authenticates the user. The second one authorizes access to either experiment annotations or results in a certain database. It is requested by the web-interface itself and verified by comparison to an encrypted password file entry. Both webserver and M-CHIPS web-interface use the UNIX System encryption method which is based on the NBS DES[10] algorithm. The size of the key space depends upon the randomness of the password which is selected. Where the webserver uses only one global password file (for the directory), access to any annotations or results section of any database can be granted separately via separate password files by the web-interface. The web-interface consists of CGI scripts. After successful login, it passes on authentication as a 10-digit random number in a hidden tag on each successive HTML page until the user logs out. Therefore the user keeps being authentified when the particular script ends. No reauthentication or permanent client-server-connections are needed.

[10]**N**ational **B**ureau of **S**tandards (USA) **D**ata **E**ncryption **S**tandard

Data Anlaysis

Preprocessing of hybridization intensities

Normalization

Prior to high-level analysis, data have to be normalized and filtered. In M-CHIPS, preprocessing starts with normalization of raw signal intensities. 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 hybridizations. The normalization is based on robust affine-linear regression, i.e. it corrects for additive offsets and multiplicative distortions at the same time. The algorithms fit one measurement versus a control measurement. The performance may be judged from the scatterplot of the raw data (measurement versus control measurement, Fig. 9). 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.

M-CHIPS implements two algorithms as described in [54] and [55]. From amongst the options given in [54] I use the 5% quantile[11] of each hybridization as the additive offset to subtract initially. The original algorithm results in a shift of the original data to a lower intensity level, making it necessary to ignore values below a certain threshold. For correspondence analysis it is advisable to use another normalization method because low intensity signals have to be kept in order to avoid missing data. Instead, I shift all hybridizations additively to a higher range, in order to prevent overly biasing CA by the large relative error common to low intensities. This shifting is done such that the 5% quantiles coincide with that of the control measurement. For both algorithms, the set of trusted spots of unvaried expression taken into account for

[11] Ranking the genes by signal intensity for a given hybridization, the maximum intensity of the lowest 5% of the genes is referred to as 5% quantile of the hybridization.

fitting can be specified. In most cases, the share of differentially transcribed genes is low, enabling to use the entire set for normalization. Otherwise, external control spots reporting defined mRNA concentrations or trusted housekeeping genes have to be used. For the former, defined amounts of complementary mRNA are added to the samples prior to the labelling step. For the latter, a text file is imported into M-CHIPS, listing genes that are trusted to be constitutively transcribed under the investigated experimental conditions.

In order to normalize a whole multiconditional experiment, the above algorithms are 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 (Fig. 9). 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

Prior to high-level analysis, M-CHIPS provides a means to select genes which fulfill the following criteria: considerable absolute expression level in at least one of the conditions; substantial change relative to the control condition in at least one of the other conditions; and reproducibility in the separation from the control condition (Fig. 10) in at least one of the other conditions.

Intensity. For many arrays and experiments, the majority of genes spotted on the array are not expressed to a measurable amount. While displaying notable ratios due to measurement fluctuations, they can be eliminated by means of an intensity filter. For monochannel experiments, meaningful intensity levels are obtained by the normalization procedure. For more than one channel, apart from reflecting a low concentration of the corresponding mRNA, a low signal can be caused e.g. by high concentrations of differently labeled mRNA taking the majority of the binding sites of the spot. Therefore, multichannel intensity values are not valid as such but only in conjunction with the other channel(s) of the same hybridization. This establishes the above requirement of one and the same control condition on each hybridization for comparability. For the same reason, normalized multichannel intensities cannot be used for high-level analysis nor for intensity filtering. However, they can serve to compute ratios reflecting the relative abundance of a certain mRNA sequence under a specific condition compared to a control condition. To compute intensity levels from multichannel ratios for filtering purposes,

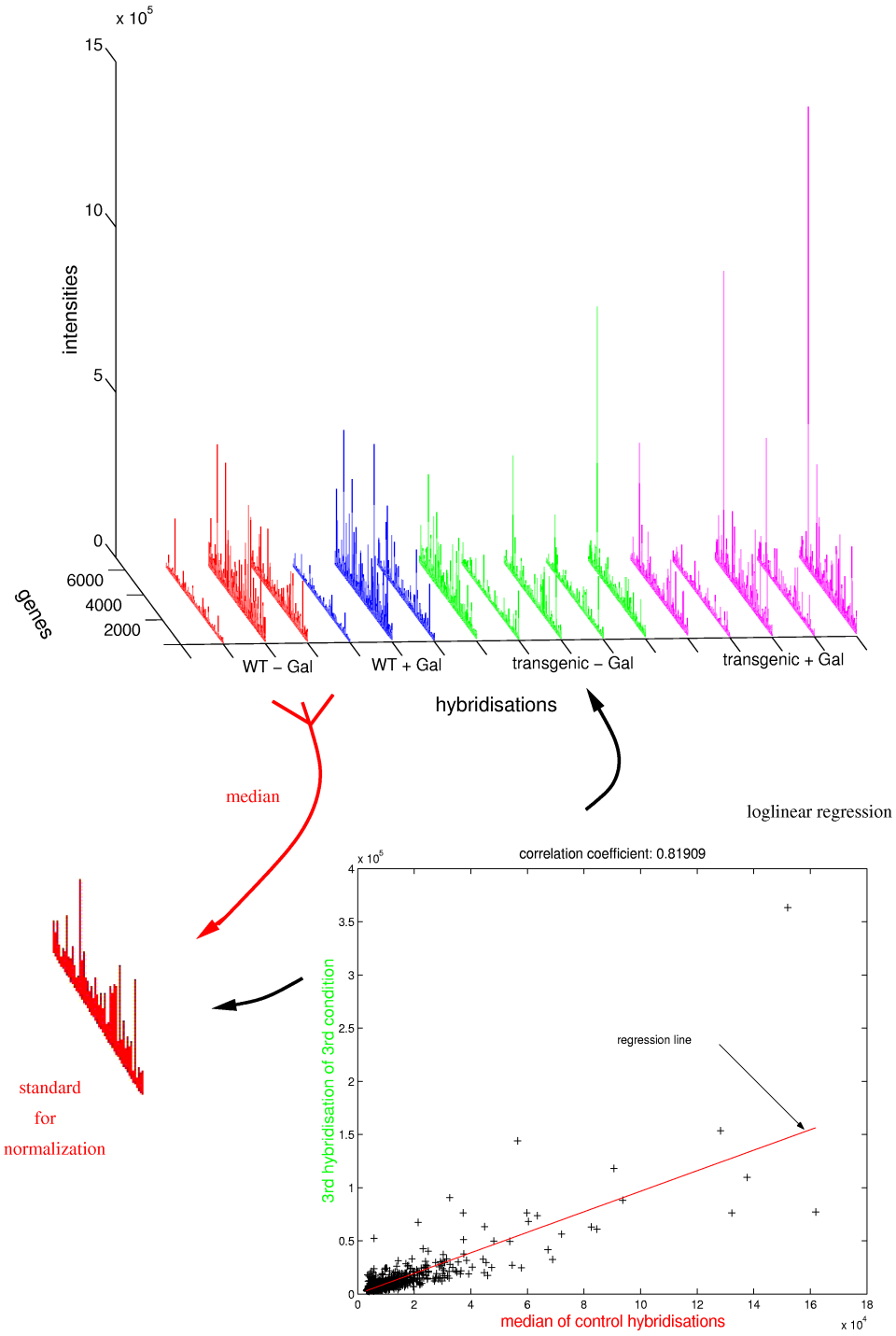

Figure 9: **Normalization of monochannel data.** Original intensity levels are shown for each hybridization and gene of a data set explicitly discussed in Fig. 15. It comprises four experimental conditions, each of which has been studied by three to five repetitively performed hybridizations. Necessity for normalization is stressed by apparently different intensity levels for hybridizations representing the same experimental condition. Each single hybridization is adapted to the gene-wise median of the hybridizations belonging to the control condition (red). Thus, the normalization algorithm is iterated once for each hybridization including the control hybridizations (that are adapted to their gene-wise median). The adaption is carried out by log-linear regression as shown for the third hybridization of the green condition. The scatterplot axes show arbitrary (machine dependent) intensity units.

these ratios are multiplied with an average control measurement, being the gene-wise median of the absolute intensities of the control channels. This average is a more stable basis for the determination of intensity levels. Apart from eliminating outliers by averaging repeated measurements, this procedure accounts for the above example of highly abundant differently labeled mRNA. Provided that less than 50% of the non-control conditions under study show such a high abundance of a specific mRNA, the intensity level of the control condition for that mRNA will not be low due to lack of binding sites.

Ratio. For multichannel data, ratios for each measurement are computed by dividing each normalized non-control channel gene-wise by the control channel of the same hybridization. For monochannel data, each hybridization is divided by the gene-wise median of all control hybridizations.

Separation. Apart from intensity and ratio filters, reproducibility measures [54] are applied to extract genes that are reproducibly up- or down-regulated. These measures integrate repeatedly performed measurements for the same experimental condition by providing the separation from a control condition (Fig. 10). Apart from being filter criteria for the set of genes, they are plotted versus the average intensity level and ratio as a measure for quality control. Moreover, they have been successfully applied directly as high-level analysis input (not shown). For this, all negative separation values are set to zero and attached with a positive sign for upregulation or a negative one for downregulation, instead. Thus they can be viewed as log ratio signal, which is suppressed by imperfect reproducibility.

Filtering. To filter out genes displaying intensities clearly above the detection limit, significant relative change and good reproducibility of this change, intensity-, ratio- and separation-thresholds can be applied. Genes not satisfying these constraints are discarded. One can also discard genes above rather than below a threshold which proved to be useful to account for saturation effects occuring e.g. if radioactively labeled arrays were exposed too long. In general, M-CHIPS provides AND-combination of three independent constraints, each of which can be defined as

- selecting genes above or below a certain threshold;
- that threshold applies to raw or normalized intensities or ratios, condition-medians of normalized intensities, ratios of condition-medians of intensities, min-max separations, or standard-deviation separations;
- it may be operational only within a specified set of conditions, e.g. all the conditions under study.

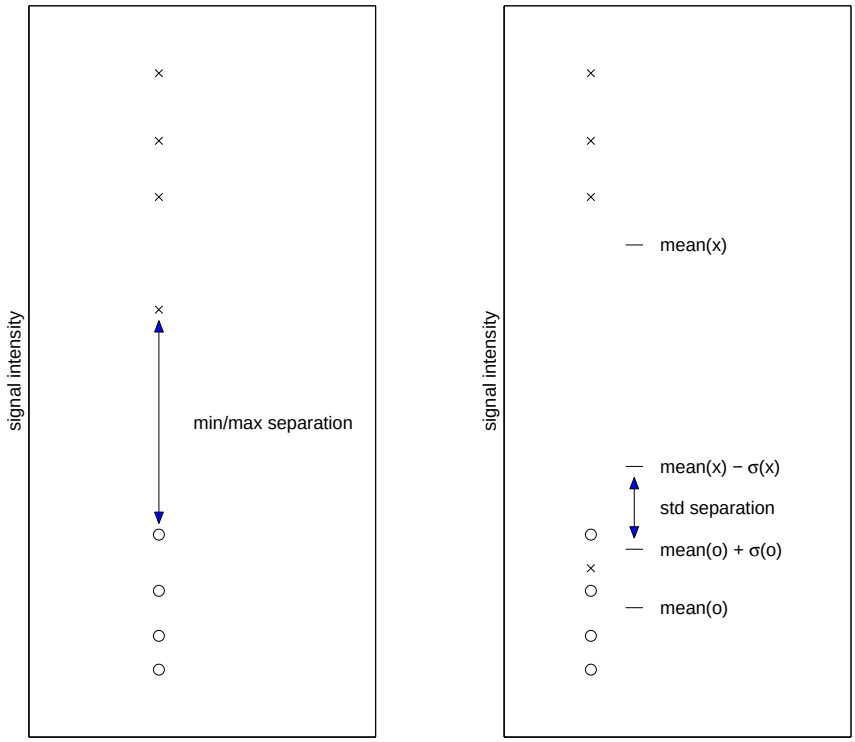

Figure 10: **Minmax- and standard-deviation separation.** Distributions of repeated measurements are differential among the genes, depending on the intensity level. Usually, there are not more than three to five values per gene and condition available for averaging. Here they are denoted as circles and 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 [54]. 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. From [54].

Correspondence analysis applied to mono- and multi-channel microarray data

For analysis of transcription intensities I implemented hierarchical clustering [28] and correspondence analysis [55]. While the former yields easy to interpret output and is widely used for microarray data analysis, it poorly visualizes important details as shown below. Moreover, it is of limited use for integrated analysis of more than one variable and therefore not useful for the investigation of interdependencies among genes and measurements intended here. In contrast, CA provides the capability to simultaneously visualize more than one variable.

Correspondence Analysis

I provide here a concise summary of the technique, see refs. [56] and [57] for a thorough exposition. An informal, intuitive description will be given below. The aim is to embed both rows (genes) and columns (hybridizations) of a matrix in the same space, the first two or three coordinates of which contain most of the information. Let I genes and J hybridizations be collected into the $I \times J$ matrix $\mathbf{N}$ with elements n_{ij} . Let n_{i+} and n_{+j} denote the sum of the ith row and jth column, respectively. By n_{++} I denote the grand total of $\mathbf{N}$. The mass of the jth column is defined as $c_j = n_{+j}/n_{++}$, and likewise the mass of the ith row is $r_i = n_{i+}/n_{++}$. Basis for the calculation is the *correspondence matrix* $\mathbf{P}$ with elements $p_{ij} = n_{ij}/n_{++}$ from which the matrix $\mathbf{S}$ with elements $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$ is derived. $\mathbf{S}$ is submitted to singular value decomposition [58], i.e. it is decomposed into the product of three matrices: $\mathbf{S}=\mathbf{U}\Lambda\mathbf{V}^T$. Λ is a diagonal matrix, and its diagonal elements are referred to as the singular values of $\mathbf{S}$. I think of them as sorted from the largest to the smallest and denote them by λ_k. The coordinates for gene i in the new space are then given by $f_{ik} = \lambda_k u_{ik}/\sqrt{r_i}$, for $k = 1, ..., J$. Hybridizations are viewed in the same space with hybridization j given coordinates $g_{jk} = \lambda_k v_{jk}/\sqrt{c_j}$, for $k = 1, ..., J$. These coordinates are called principal coordinates.

To reduce dimensionality, only the first two or three coordinates of the new space are plotted. The loss of information associated with this dimension reduction is quantified in terms of the proportion of the so-called total inertia $\sum_k \lambda_k^2$ that is explained by the axis displayed. Total inertia is proportional to the value of the χ^2 statistic, and thus the amount of information represented in, e.g., a planar embedding $(\lambda_1^2 + \lambda_2^2)/\sum_k \lambda_k^2$, corresponds to the proportion of the χ^2 statistic explained by the embedding.

The above summary is aimed to provide all the information needed to implement a simple CA algorithm. This 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 (Appendix B).

Standard coordinates as an aid in visualization

Correspondence analysis attempts to separate dissimilar objects (genes or hybridizations) from each other; similar objects are clustered together resulting in small distances. In contrast, the distance between a gene and a hybridization cannot be directly interpreted. For visualization of between-variable association in the plot one includes virtual genes which have all their intensity focused in one hybridization [57]. The coordinates of such a gene are called standard coordinates of the hybridization where this gene is expressed. Likewise, one could introduce standard coordinates for genes. The standard coordinates for the genes are computed as $u_{ik}/\sqrt{r_i}$ and for the hybridizations as $v_{jk}/\sqrt{c_j}$. In practice, the spread of the set of real genes and hybridizations is much smaller than the spread introduced when including these virtual genes and hybridizations via their standard coordinates. As a consequence, the real points would shrink to a tiny area, so I rather depict the direction from the centroid of the data to the standard coordinates instead of the standard coordinates themselves.

Medians and replicate hybridizations in correspondence analysis

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

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

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. The intensity unit cancels out when calculating the correspondence matrix $\mathbf{P}$.

Interpretation of a correspondence analysis biplot

Correspondence analysis was originally developed for contingency tables and is intimately connected with the χ^2 test for homogeneity in a contingency table. The question asked by this is whether the differences among the rows (or columns) of the table are large enough to reject the hypothesis that the rows (columns) are homogenous. In other words, it is asked whether the discrepancies between the observed rows and an average row profile expected for the homogenous case, are so large that they are unlikely to occur by chance alone. The question is answered by computing a measure of discrepancy between all the observed and expected values: The difference between observed and expected value is squared and subsequently divided by the expected value. The result is referred to as the χ^2 distance between observed and expected value. These distances, calculated for all elements of the table, sum up to the χ^2 statistic. The value of the χ^2 statistic is high when there is an association between rows and columns of the table. In CA, points are depicted such that the sum of the distances of the points to their centroid (called "total inertia") is proportional to the value of the χ^2 statistic of the data table. The farther a point is away from the centroid, the higher is its row's contribution to the value of the statistic. In this sense, CA decomposes the overall χ^2 statistic. Distances among points are not meant to approximate Euclidean distances but rather the χ^2 distance. This distance is low when the profiles of two vectors show similar shape, irrespective of their absolute values.

Together with the row-points, CA displays points representing columns and does so using the same χ^2 criterion. This also establishes the link between row and column points. If a column determines an outstanding entry of a row (and vice versa), then the corresponding row and column points tend to lie on a common line through the centroid. For a positive association the two points will lie on the same side of the centroid, with the distance to it larger the stronger the association is. A negative association will cause the column-point and the row-point to lie on opposite sides of the centroid. The following example demonstates the practical use for microarray data analysis. For clarity, a fictitious example has been constructed. Fig. 11 shows virtual profiles for 24 genes, transcription intensity plotted versus experimental condition. Let each condition comprise two repeated hybridizations. CA will decompose these profiles corresponding to their association to the hybridizations under study. Fig. 12 shows the expected planar projection. The genes are depicted as black dots, the

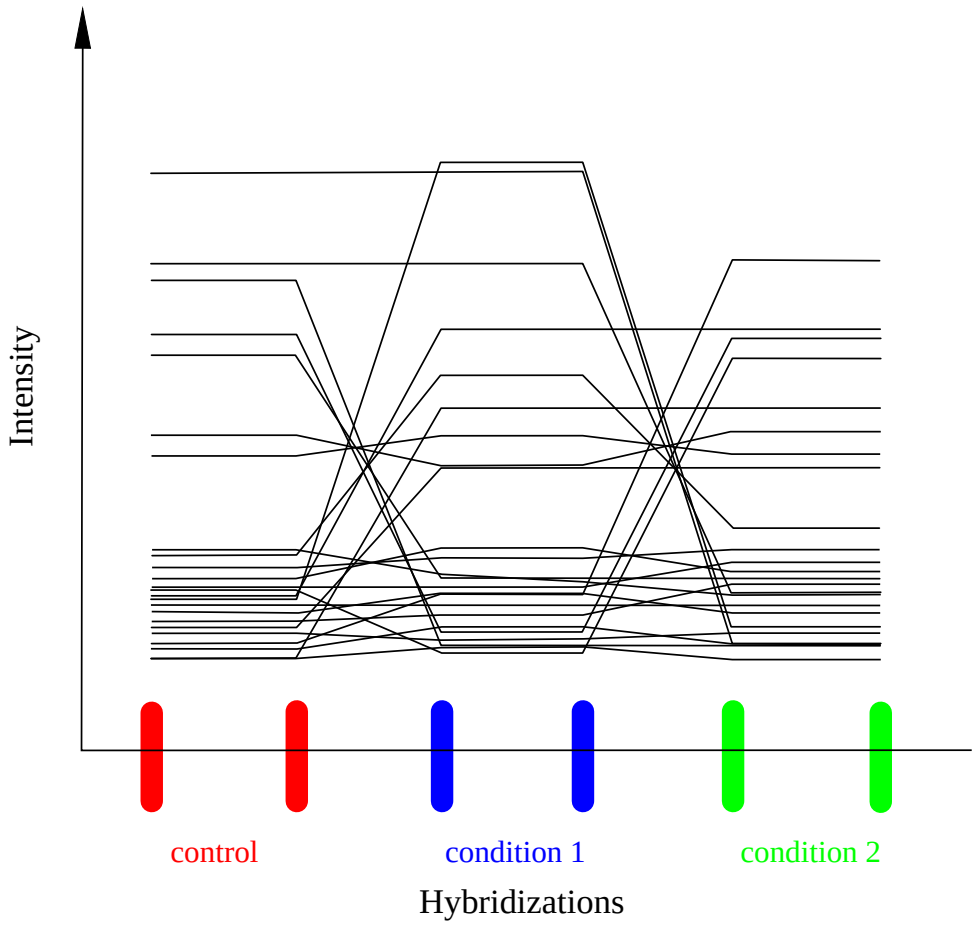

Figure 11: **Virtual data example.** To demonstrate interpretation of a CA biplot, a data example has been constructed. It resembles real data in that the majority of the genes is lowly or not transcribed to a measureable amount. It comprises only 24 genes and differs from the real world in perfect reproducibility among the two hybridizations of each experimental condition.

hybridizations as boxes, color-coded according to the experimental condition they belong to. To illustrate the locations of the genes, each gene cluster is shown together with the according gene profiles. 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 [59]. 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.

The above items are sufficient for correct CA plot-interpretation of most real data examples. However, the second item, claiming common directions for conditions and associated genes is not mathematically correct and has to be seen more as a 'rule of thumb'. It can be proven that the directions of a condition and its highest possible associated gene never coincide exactly. To properly visualize associations between conditions and genes, virtual genes will be used that are fully concentrated on one condition. They serve as representatives of the hybridizations in gene-space (see standard coordinates above). Because they show highest possible association to the according conditions, they are located far away from the centroid. Plotting them directly would cause all other points representing genes and hybridizations to be located in a small centric area. I will represent them as lines from the centroid to their coordinates (see Fig. 14), with the latter lying outside the plot margins after zooming into the area of conditions and 'real' genes.

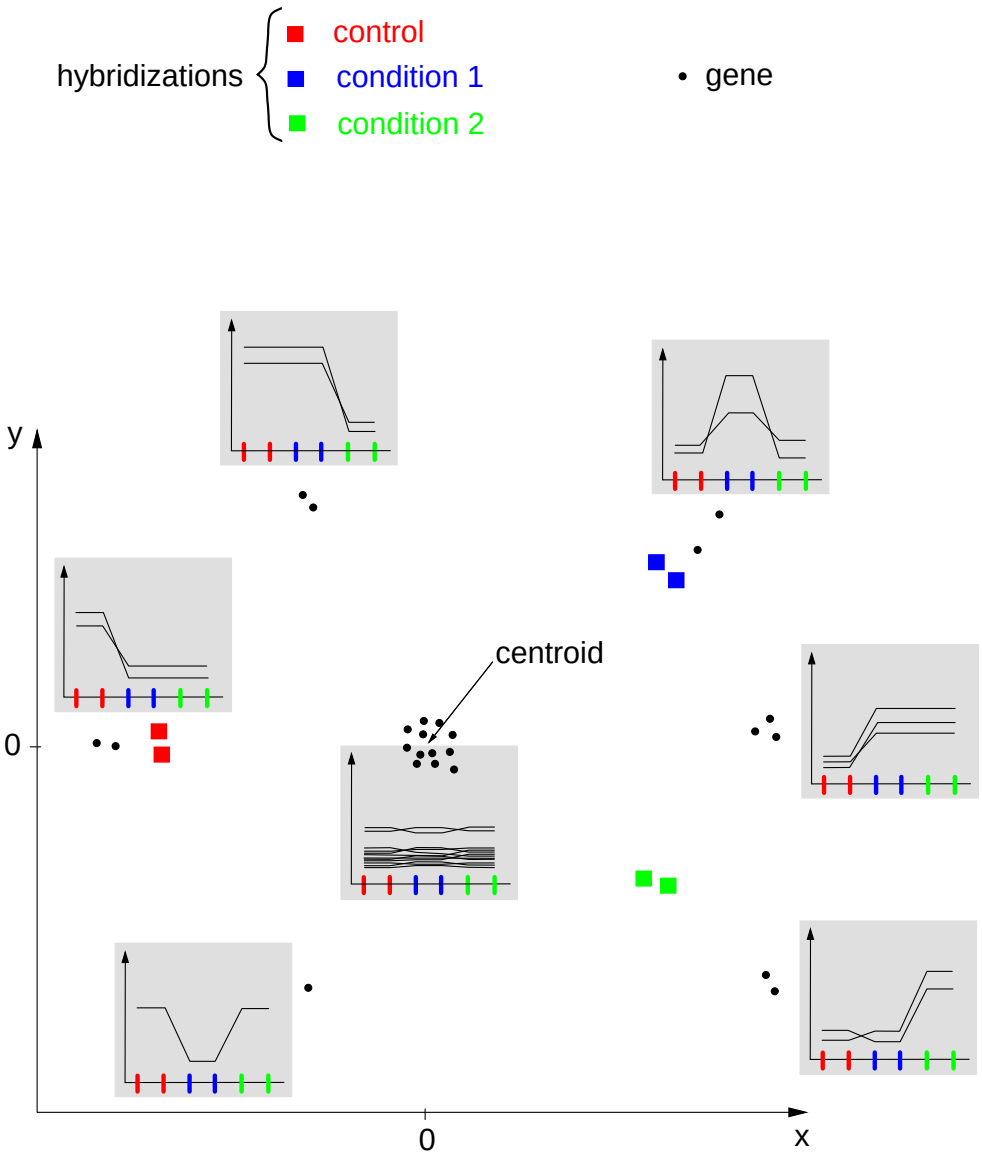

Figure 12: **Planar projection by CA.** The plot resembles output that can be expected for the virtual data example shown in Fig. 11. It was constructed to demonstrate properties of such projections more clearly than possible by showing a single plot of real data. Gene-clusters are shown together with the according gene profiles, which are discussed in the text. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless.

Multichannel data example — analysis of a well-known data set

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.* [47] which are publicly available[12]. Spellman *et al* arrested the *S. cerevisiae* cell cycle by four different methods, namely α factor-, *CDC15*- and *CDC28*-based blocking, and elutrition. Here and in the legend to Fig. 14 I will refer to these four methods as 'alpha', 'cdc15', 'cdc28', and 'elu', respectively. 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 data consist of two-channel fluorescence signals. Following the authors I based my analysis on the log-ratio of the intensities of the two channels. To make the data analyzable by CA, they were additively shifted to a positive range. In my analysis I gave mass to all hybridizations instead of applying HMS. The standard coordinates of the hybridization medians, on the other hand, were computed 'without mass' and are depicted as lines emanating from the centroid. I analyzed the 800 cell-cycle associated genes depicted in Fig. 1 in ref. [47] over all 73 hybridizations. To allow for direct comparison, part B of the original figure, which also marks the histone gene-cluster, is shown here as Fig. 13. The CA biplot is shown in Fig 14 with hybridizations colored according to their phase assignment and following the color code of the original figure.

The planar embedding produced by CA (Fig. 14*a*) shows the hybridzations 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 in Fig. 14*b* which is further subdivided according to the method of cell cycle arrest that had been used. The red-encircled genes will be discussed in the next section 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

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

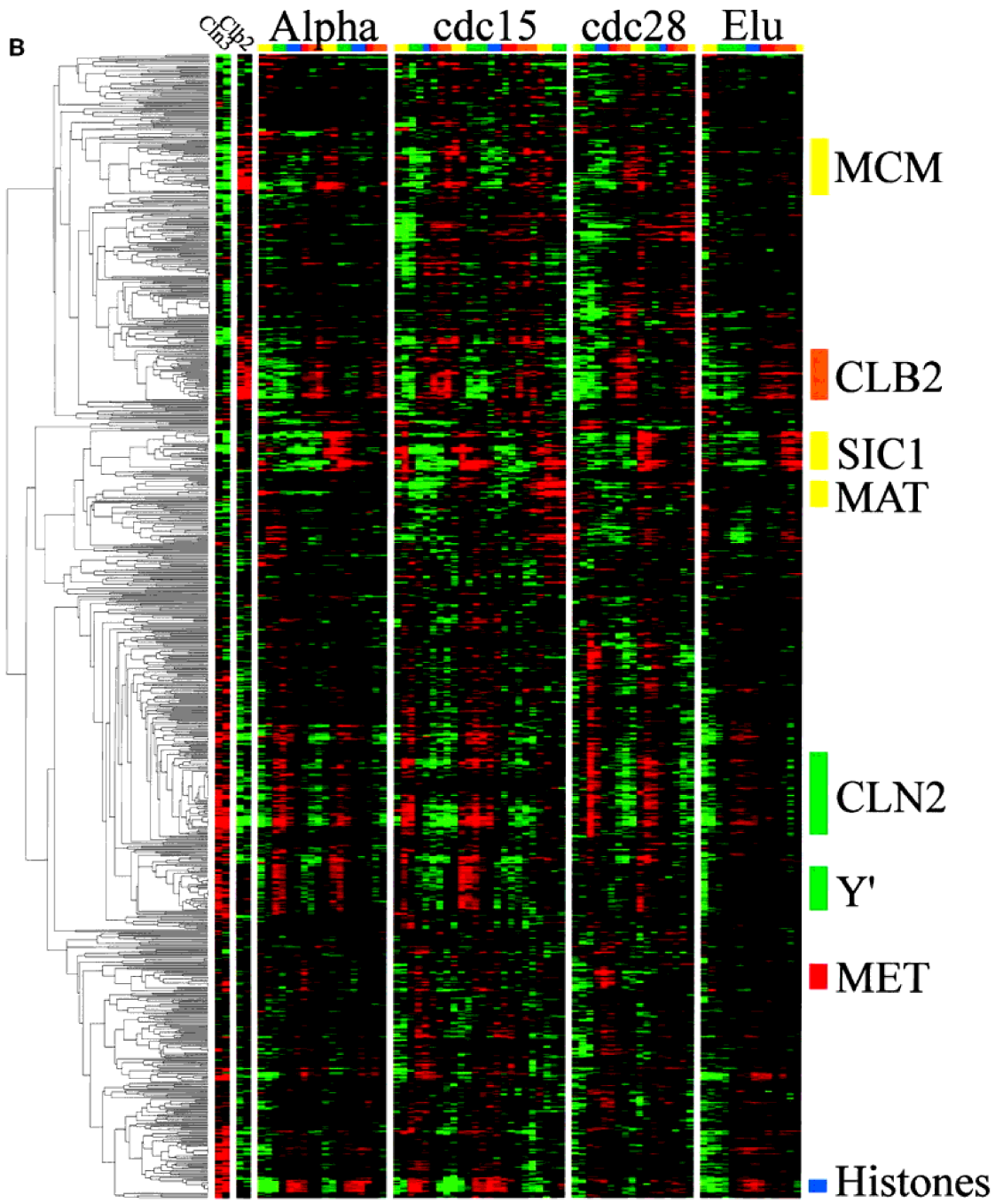

Figure 13: **Original figure 1B [47].** This figure has been reproduced from ref. [47]. It shows gene expression during the yeast cell cycle. Genes correspond to rows, and the time points of each experiment are the columns. The ratio of induction/repression is shown for each gene such that the magnitude 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.

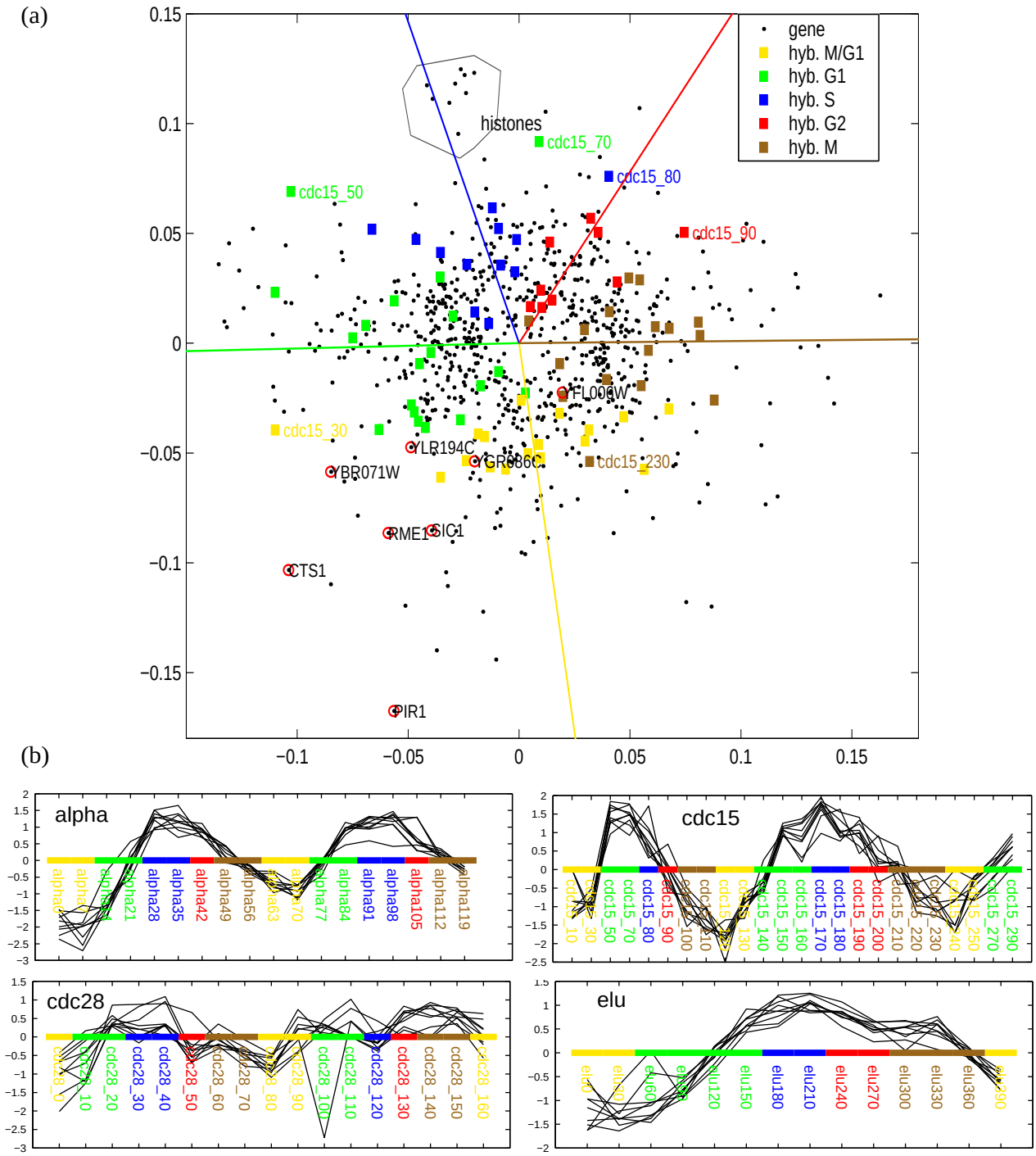

Figure 14: **Cell cycle synchronization data by Spellman *et al.*** The data set composed of 800 cell-cycle associated genes has been projected by CA as is. No HMS has been employed in order not to bias the resulting plot in terms of separation of the cell-cycle phases. The outlying hybridizations have been identified to be caused by a slight phase shift of the cdc15-based synchronization visible in the upper right panel of Fig. 14*b*, which shows the profiles of the nine histone genes *HHF1, HHF2, HHT1, HTB2, HHT2, HTB1, HTA1, HTA2* and *HHO1* encircled in black in Fig. 14*a*. Further explanation is given in the text. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless. From [55].

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 of Fig 14*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. Fig 14*a* 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.* [47] (printed here as Fig. 13 for direct comparison) does not properly show this shift.

Monochannel data example — overexpression of *CDC14*

Instead of following the cell cycle through S phase, G2, mitosis and G1, this experiment focuses on the transition from mitosis to G1. In late mitosis, mitotic cyclin-dependent protein kinases have to be inactivated in order to exit mitosis. Cdc14p plays a major role in this transition to G1 being a dual specific phosphatase. In this experiment cells were arrested in mitotic meta phase, and *CDC14* was overexpressed by inducing the controlling *GAL1*-promoter. Thus, one cannot directly observe the effect of *CDC14* overexpression because it will be overlaid with the gene expression changes due to the presence of galactose. To subtract for these effects, wild type and transgenic strain were grown under repressing and inducing conditions leading to four samples which were subjected to array hybridization: wild type without galactose; wild type with galactose; transgenic yeast without galactose (no induction of *CDC14*); and transgenic yeast with galactose.

In contrast to the previous example, CA was used to analyze transcription intensities instead of two-channel ratios. The data, consisting of three to four hybridizations per condition were normalized. Genes (1,400 of of 6,100) were extracted for being reproducibly differential (positive minmax separation) from the control condition (WT strain without induction) in at least one of the other conditions. Measurement noise was further reduced by HMS. Planar

embedding by HMS (Fig. 15) explains 80.8% of the total inertia, compared to 52.5% in case of embedding all the hybridizations separately (not shown). Hybridization medians are represented both in principal coordinates and as lines to their standard coordinates. The actual hybridizations, each separated into primary and secondary spot sets, are drawn as supplementary points. They are represented only in principal coordinates as are the genes.

The biplot (Fig. 15) clearly shows four directions corresponding to the four conditions. 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 [60] and *CTS1* which belongs to the cluster of *SIC1* co-regulated genes [47]. *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.* (see Fig. 14, 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.

Integration of gene and experiment annotations with transcription profiling

With correspondence analysis, as discussed above, it is possible to visualize associations both among and between genes and hybridizations. However, these associations extracted from gene×measurement tables need biological interpretation to become meaningful results. Both genes and measurements can be annotated by biologically relevant information. Much like the intensity or ratio data themselves, these annotations are not suited for visual inspection due to their high number. While they are usually reduced to a few putatively important parameters at the risk of overlooking something unexpected, I want to keep all of them. Previously described efforts in storing complex and detailed experiment annotations in a statistically accessible form have to be seen as prerequisites for integrating as much information into analysis as possible.

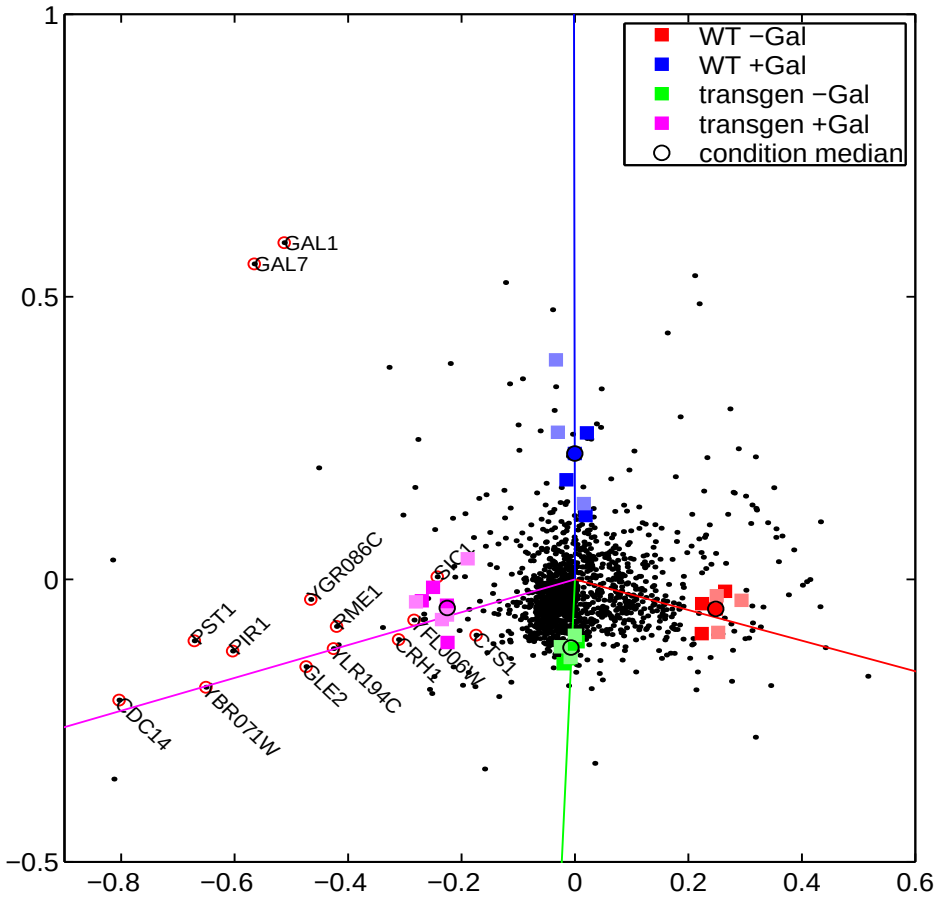

Figure 15: **Overexpression of Cdc14p.** In arrested yeast cultures Cdc14p expression was induced under control of the *GAL1* promoter and investigated in comparison to uninduced transgenic, uninduced WT and inductor-exposed WT cells. Three to four hybridizations have been performed for each experimental condition. Both spot sets of the array are drawn separately for each hybridization, primary, and secondary spot set depicted in light and dark colors, respectively. The conditions are colored, and their hybridization medians are marked according to the legend in the upper right corner. Lines are drawn in the direction of the standard coordinates of the condition medians in appropriate colors: genes like *GAL7* and *GAL1* are associated with both the WT and the transgenic strain grown in the presence of galactose to an equal share. *CDC14* is associated with the induced transgenic strain only. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless. From [55].

General interconnectivity among different visualization plots

Analysis techniques provided by M-CHIPS include hierarchical clustering [28], CA [55], and statistical analysis of experiment annotations for arbitrary sets of hybridizations, e.g. those clustered by similar expression profiles. Comparison of different visualizations of a dataset are facilitated by highlighting data points which have been selected in another plot. It is also possible to mark all genes bearing a certain keyword like 'cell cycle' in their gene annotation or to import multiple sets of gene tags from text lists. In the CA plot, several disjoint gene sets can be visualized by different color, e.g. to highlight different functional categories or to mark interesting clusters of genes. For the latter, gene sets can be selected by encircling them by mouse clicks. Expression profiles of marked genes can be displayed in a parallel coordinate plot.

In the same manner clusters of measurements can be selected. They are automatically scanned for significant experiment annotation values. For each value of every annotation, instances of occurence are counted. For a particular value, its frequency in the cluster is determined as the number of its occurences in the cluster divided by the number of measurements in the cluster. Comparison to its frequency in the whole set of measurements under study reveals whether it is over- or underrepresented in the cluster as sketched in Fig. 4. An example is presented in the following subsection.

Characterization of measurement clusters by experiment annotation scan

A time course has been recorded for wild type *S. cerevisiae* cells under oxidative stress by 0.2 M hydrogen peroxide. Data have been preprocessed and visualized by CA (Fig. 16). Experimental and computational details are given in Fig. 17 and Appendix A. The plot comprises both genes and measurements. The genes are depicted as black dots. Measurements are shown as squares, color-coded according to the experimental condition they belong to. There is one outlying cluster of measurements belonging to the 30 min timepoint (pink), whereas other measurements of the very same condition are located in a distant area, clustering with other timepoints. Selecting these outliers, searching for at least 2-fold over- or underrepresented annotation values results in values belonging to only 8 out of 111 annotations (Tab. 6). 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

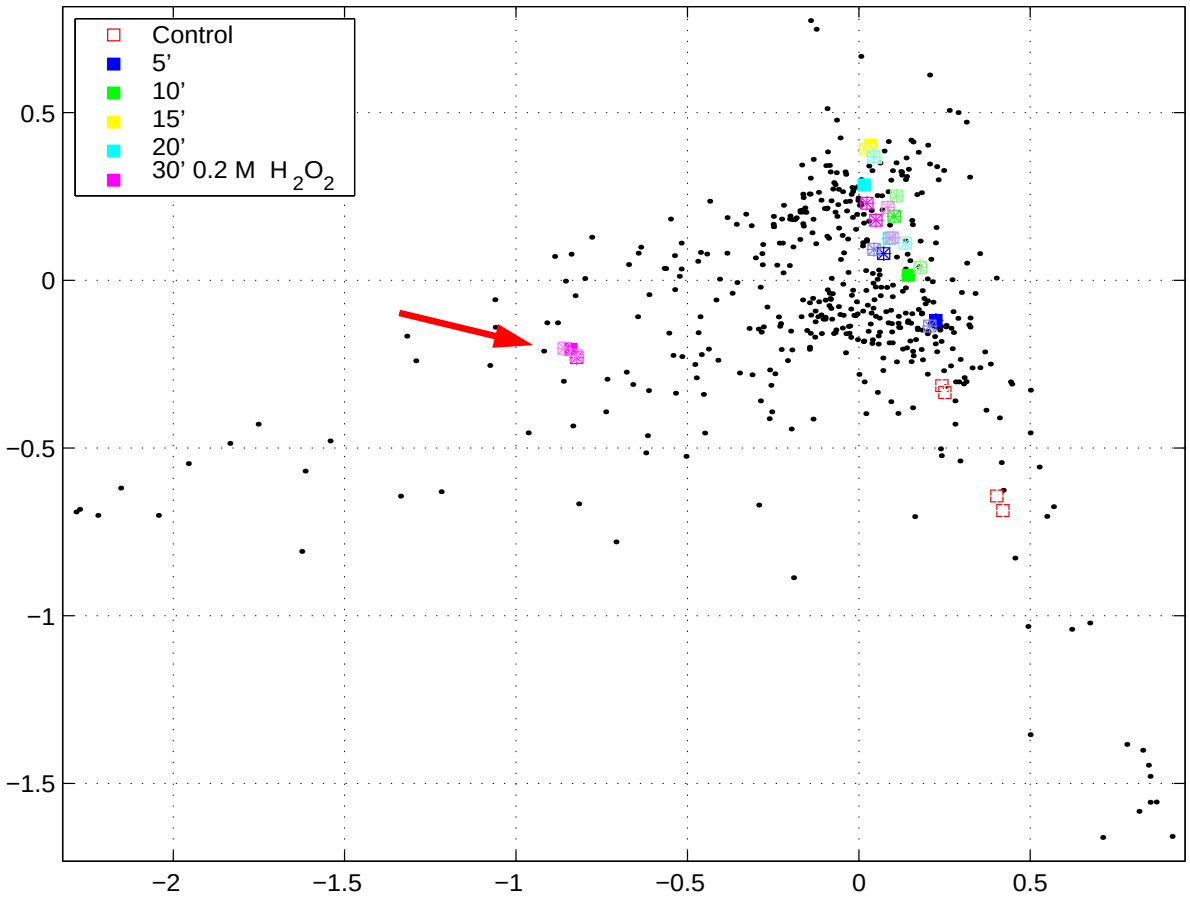

Figure 16: **Oxidative stress.** The correspondence analysis plot shows a dataset recorded from wild type yeast cells responding to 0.2 M hydrogen peroxide in the medium. Genes are depicted as black dots, measurements (in this case monochannel hybridizations) are shown as squares, color-coded according to the experimental condition (here time point) they belong to. Further explanation is given in the text. The outlying cluster of pink labeled measurements (red arrow) is further characterized by experiment annotation values over- or underrepresented in this cluster as shown in Tab. 6. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless. From ref. [50], modified.

More than or exactly 2x over- or 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 16:  label_incorporation_rate
value 44 is absent (0/2 in cluster :  1/14 in total)
value 46 is absent (0/2 in cluster :  1/14 in total)
value 51 is absent (0/2 in cluster :  2/14 in total)
value 52 is 7x overrepresented (1/2 in cluster :  1/14 in total)
value 56 is 7x overrepresented (1/2 in cluster :  1/14 in total)
value 59 is absent (0/2 in cluster :  1/14 in total)
value 68 is absent (0/2 in cluster :  1/14 in total)
value 84 is absent (0/2 in cluster :  1/14 in total)
value 87 is absent (0/2 in cluster :  2/14 in total)
value 88 is absent (0/2 in cluster :  2/14 in total)
value 93 is absent (0/2 in cluster :  1/14 in total)

annotation 17:  total_activity
value 26000000 is absent (0/2 in cluster :  1/14 in total)
value 34000000 is absent (0/2 in cluster :  1/14 in total)
value 35000000 is absent (0/2 in cluster :  1/14 in total)
value 36000000 is absent (0/2 in cluster :  1/14 in total)
value 38000000 is 7x overrepresented (1/2 in cluster :  1/14 in total)
value 39000000 is absent (0/2 in cluster :  1/14 in total)
value 43000000 is 7x overrepresented (1/2 in cluster :  1/14 in total)
value 46000000 is absent (0/2 in cluster :  1/14 in total)
value 56000000 is absent (0/2 in cluster :  1/14 in total)
value 61000000 is absent (0/2 in cluster :  2/14 in total)
value 65000000 is absent (0/2 in cluster :  1/14 in total)
value 71000000 is absent (0/2 in cluster :  1/14 in total)
value 80000000 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)
```

Table 6: Frequencies of characteristic annotation values

show up in mid-range for both annotations in table 6.

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 in Figure 16 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 CA plot shown in Fig. 17.

Disregarding the measurements hybridized on a different array series, it is possible to differentiate response phases and to identify the genes involved. Figure 17 reveals a leap in transcriptional status of the cells between 15 and 20 minutes after start of treatment, consistent with the work of Godon and co-workers [61]. This includes both genes with increased and decreased transcription levels. As an example of a marked expression pattern I selected a cluster of four genes being activated in the initial phase of the response but down-regulated after 20 minutes (Fig. 18).

Scanning annotations of continuous range

For the example above, all annotations were treated as categorical. Sometimes, especially with higher numbers of measurements, it is desirable to aggregate values for annotations of continuous range (see no. 16 and 17 in Tab. 6). 'Label incorporation rate' may thus be discretized into e.g. low, medium and high values. M-CHIPS provides methods enabling discretization of annotation ranges into a chosen number of bins due to their particular distribution or by expert knowledge. After selecting a cluster of measurements in a CA plot, a loop is entered for repeated extraction of characteristic annotation values. At the beginning of each loop step, a menu enables alteration of the representation factor threshold as well as discretization of an annotation (Fig 19*a*). The discretization routine displays a histogram of all values annotated for the selected annotation in the whole database (Fig. 19*b*). The number of intervals to display as well as linear or logarithmic scale can be selected for the histogram. The routine allows to discretize into equally spaced intervals. Alternatively, arbitrary interval centers can be manually selected by mouse click. When the histogram shows the desired segmentation of

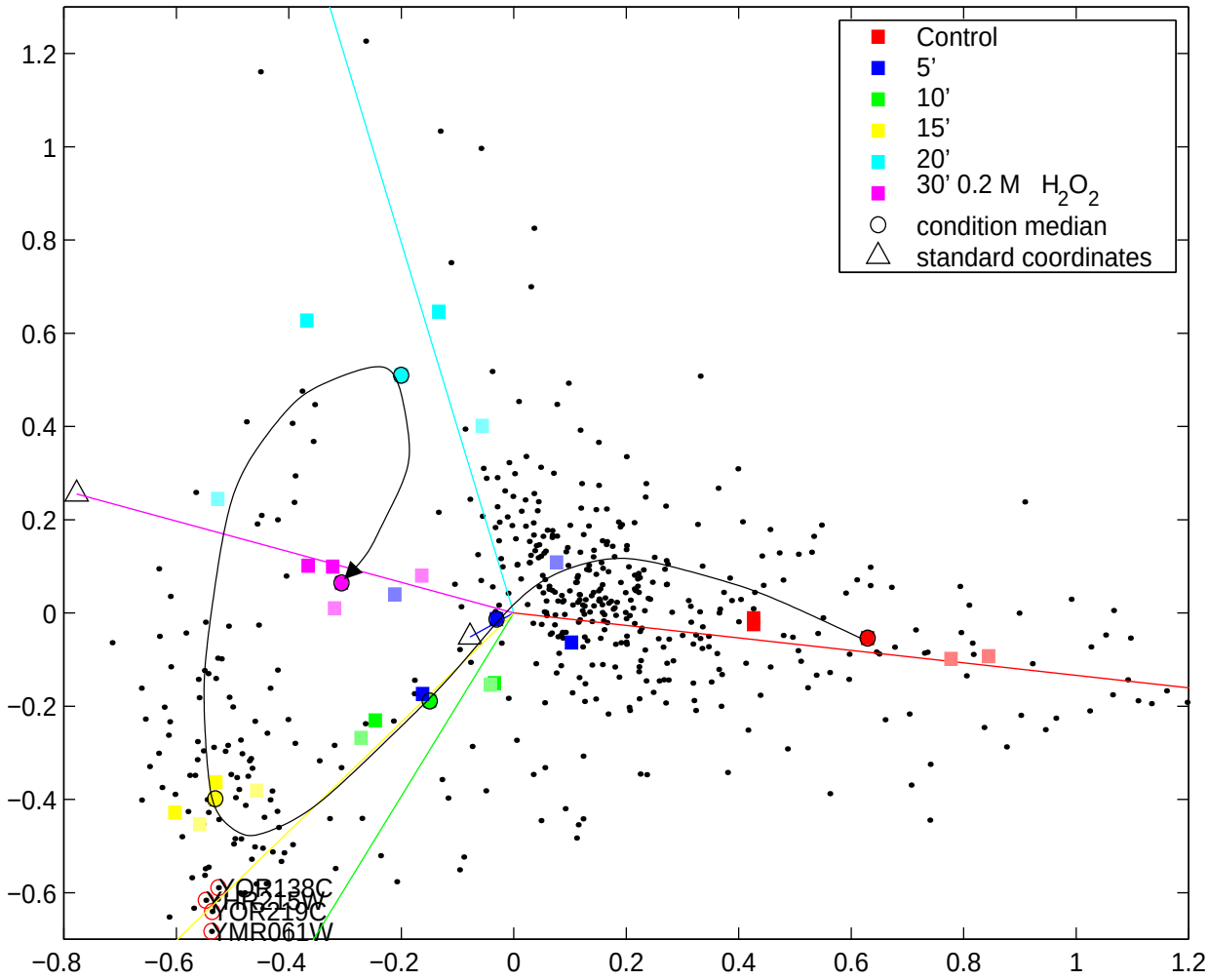

Figure 17: **Oxidative stress.** Wild type yeast was grown in the presence of 0.2 M hydrogen peroxide and sampled in 5 min-intervals (10 min at end). The experiment was performed twice yielding two hybridizations per experimental condition (timepoint). These conditions are color-coded according to the legend, double-spotting in dark and light colors. A black arrow shows the course of the experiment. Four genes are tagged that are almost exclusively associated with the 15 min timepoint. Their profiles are given in Fig. 18. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless. From the supplemental material to ref. [50], modified.

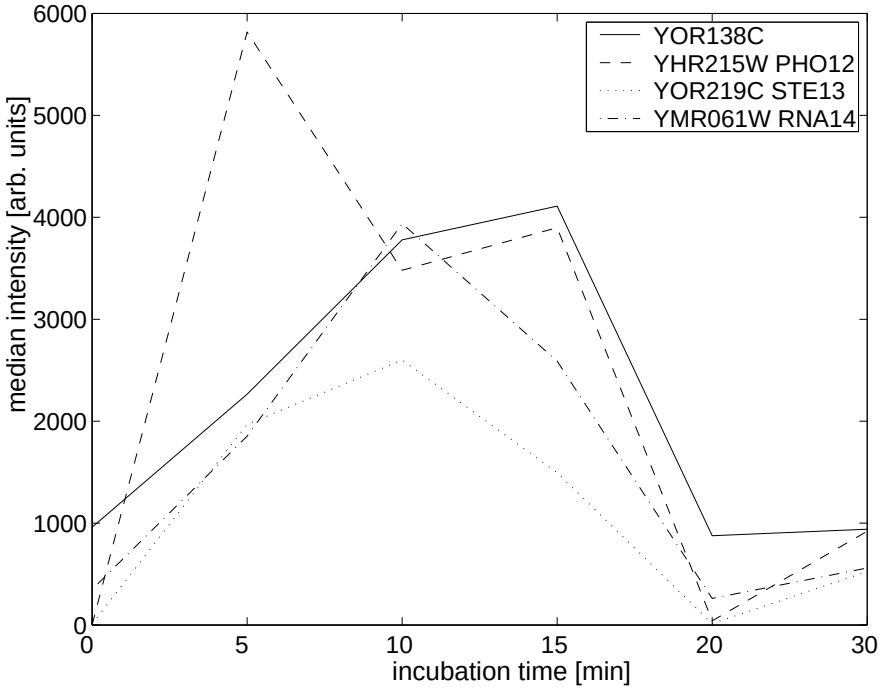

Figure 18: **Profile of the four genes tagged in Fig. 17.** For the four genes tagged in Fig. 17, the median transcription intensity is plotted against the response time. Their ORF (open reading frame) identifiers and their names are provided in the legend. The ordinate shows arbitrary (machine dependent) intensity units.

the value range, the values can be replaced by the displayed interval centers. The effect of this measure in conjunction with the choosen representation threshold is immediately visible in the displayed list of characteristic values (see Tab. 6) and can be revoked, revised, improved or augmented by discretization of other annotations in the next round.

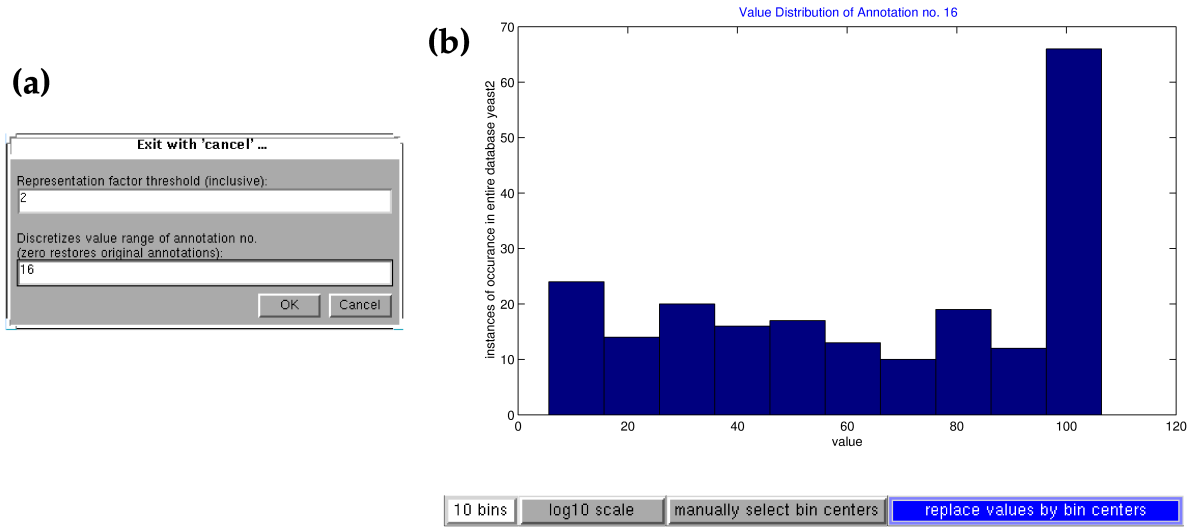

Figure 19: **Processing annotations of continous range.** The values of continuously ranged annotations can be identified with a certain interval containing them. After selection of appropriate intervals (equally spaced or customized in linear or logarithmic scale), the values are assigned the center of the according interval and the scanning for characteristic scores if performed with the new values. The procedure can be iterated (or reversed) until an informative binning has been achieved.

Discussion

The present work introduces a storage system and methods to study interdependencies among large-scale microarray data. I applied correspondence analysis as an explorative and powerful statistical tool to study interdependencies both between and among sets of variables, i.e. genes and hybridizations obtained by expression profiling [55]. It is necessary to carefully preprocess the data and to perform correspondence analysis in a way that is capable of dealing with measurement noise and outliers, thus adapting the method to the particular requirements of microarray data.

To achieve a wide analytic scope, it is not sufficient to put the hybridization intensities into a common format as a platform for preprocessing and analysis. With more and more hybridizations at hand, the necessity rises of providing this platform also for the experiment annotations for biological interpretation. For the statictical analysis of numerous and large datasets stemming from multiconditional microarray experiments, the data need to be properly annotated and stored in a manner such that both the datasets and their annotations are accessible to statistical analysis. Including annotated experimental parameters into statistical analysis offers the opportunity to identify the global players behind transcription patterns.

Free text annotation of recent microarray databases impairs its direct statistical access. Parameter sets used for experiment discription have not yet reached their final shape, and standards are reduced to minimal conventions that do not yet 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 the 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 requirements above and to integrate CA into a larger framework of data platform and supplemental methods [50]. M-CHIPS (Multi-Conditional Hybridization Intensity Processing System) allows for data analysis of all of its components including the experimental annotations. It addresses the rapid growth in the amount of hybridization data, more detailed experimental descriptions, and new kinds of

experiments in the future.

Data platform

M-CHIPS in the context of recent microarray data platforms

Previously published microarray database concepts have focused on the ability to include intensity data from different platforms and to make these comparable [19, 21]. These datasets are valuable only if they are annotated by sufficiently detailed experiment descriptions.

However, in many databases a substantial number of these annotations is in free-text format and not readily accessible to computer-aided analysis. Some projects have started to develop controlled vocabulary for experiment description (e.g., ArrayExpress [21], RAD [22] and GEO [62]). However, so far little effort has been made to categorize the descriptions down to minute detail and make them amenable to analysis. To my knowledge, M-CHIPS is the first micorarray repository without any free-text in experiment annotation.

Moreover, the concept of categorizing experiment description down to atomic annotations stored as content of definition tables is unique to M-CHIPS. GeneX[13] (NCGR), ArrayExpress[14] (EBI), SMD[15] (Standford, [63]), ArrayDB[16] (NHGRI, [20]), ExpressDB[17] (Harvard, [19]), EpoDB[18] (Upenn, [64,65]), RAD[19] (UPenn), and GEO[20] (NCBI) all show a number of biological or protocol-related keywords in their UML structure representations or attribute listings, indicating a more or less static, 'hard-wired' implementation.

As my concept is not meant to be implemented in a large public gene expression database, I have not dealt with the inclusion of additional platforms such as oligonucleotide chips or SAGE, but have concentrated on mining the wealth of information contained in the experiment annotations. However, I have been able to serve several collaborating groups in providing databases and analysis tools for data from different areas of research (i.e. experiments with *S. cerevisiae*, *A. thaliana*, *T. bruceï*, *N. crassa* and human cancer samples), obtained by different platforms (radioactive hybridization to nylon or polypropylene membranes and fluorescent hybridization to glass slides), and by means of different imaging software.

[13] GeneX `http://www.ncgr.org/genex/doc/GeneXSchema.pdf`
[14] ArrayExpress `http://www.ebi.ac.uk/arrayexpress/Full_Schema.gif`
[15] SMD `http://genome-www4.stanford.edu/MicroArray/SMD/doc/db_specifications.html`
[16] ArrayDB `http://genome.nhgri.nih.gov/arraydb/schema.html`
[17] ExpressDB `http://arep.med.harvard.edu/ExpressDB/ExpressDB.v200.help.htm`
[18] EpoDB `http://www.cbil.upenn.edu/EpoDB/release/schema/schema.html`
[19] RAD `http://www.cbil.upenn.edu/cgi-bin/RAD2/schemaBrowserRAD.pl`
[20] GEO `http://www.ncbi.nlm.nih.gov/geo/info/geo.tbl`

Analytical scope

The storage system provides an unprecedented level of detail for experiment description captured in categorical and continuous variables. For data entry, this ensures completeness of experiment annotation, i.e. a level of completeness exceeding minimal standards. For analysis, it provides the opportunity to include this complete experiment information as additional variables, i.e. to study it by means of multivariate statistics. Additional attributes or additional allowed values for existing attributes can easily be added without changing the database structure or any algorithm. Previously annotated experiments can be revisited to annotate newly defined annotations. However, it is not necessary to do so. In a set of hybridizations lacking an annotation for some cases, this annotation will be ignored, enabling correct analysis of the others. Thus, no pressure is put on the defining and annotating experimenter to revisit hundreds of already stored experiments when he or she wishes to define a useful new annotation.

In microarray research and other fields, biologists and bioinformaticians have to work together. A way to share the work would be that the biologist does the practical work up to obtaining data, which are then given to the bioinformatician, who then does the analysis. Statistical implications may lead to a low value of the resulting data if the bioinformatician had not been consulted for experimental design. Analysis carried out by the bioinformatician alone, might be statistically sound, but has its weaknesses in biological interpretation.

I regard M-CHIPS to be also a communication tool. It supplies complete experiment information to the bioinformatician. A technical parameter in the protocol, such as the concentration of a particular buffer component, may be identified to be associated to the downregulation of a set of genes — even by someone not knowing its purpose. On the other hand, I implemented analysis algorithms in a way accessible to the biologist. A graphical user interface completely replaces input of function names as well as any command line parameter, enabling to perform the entire analysis by mouse clicks.

Sharing both the complete data and all analysis methods facilitates communication between experimenter and bioinformatician, resulting in analysis methods that are tailor-made for the special requirements of microarray data. Modular structure of the system and a MATLAB^{TM} environment enable quick implementation of new tools. Likewise, the experimenter is able to analyze the data herself or himself instantly after upload for direct and immediate feedback into follow-up experiments.

Furthermore, M-CHIPS has already successfully served as a communication tool between collaborating workgroups. The EUROFAN II, B2 data set[21] was aquired by 5 European groups. The raw intensities were collected and uploaded by Nicole Hauser. However, only

[21] `http://mips.gsf.de/proj/eurofan/eurofan_2/b2`

the experimenters were able to provide detailed information about technical protocols and experimental conditions. In M-CHIPS, experiment annotation is web-based to ensure that any experiment can be annotated from remote by the experimenters themselves. The effort involved in annotating experiments is minimized. After preprocessing, raw and preprocessed data as well as gene- and experiment annotations were given to the Munich Information Center for Protein Sequences (MIPS[22]) for web publication.

Reliability and universal applicability

The M-CHIPS concept allows information from heterogeneous experiments to be stored in databases of similar structure so that the same algorithms for analysis can be applied. The system has been used by collaborating groups since June 1999. Thus, all algorithms described above have been extensively tested. Currently we have 33 yeast specific (MIAME compliant[23]), 54 human tumor specific, 71 Arabidopsis specific (MIAME compliant), 41 Trypanosome specific, 20 Neurospora specific and 78 common (technical, MIAME compliant) experiment annotations. Compliance with standards such as those proposed by EBI (MIAME) is independent from my storage schema. The experimenter defining the annotations decides on standard compliance and level of detail. The sets of hierarchically ordered annotations are listed on the associated web page[24]. The entire descriptions of all hybridizations stored in M-CHIPS databases can be analyzed statistically. Apart from reliably working for two and a half years, the system has proven to be stable and performant with high amounts of data. I currently perform adiministration for 1765 hybridizations in 12 databases. They belong to the above five fields of research and comprise both radioactive-label and multichannel experiments. Although these databases may contain different parameter sets used for experiment description, they share the same structure concept[25] and therefore can be accessed by the very same algorithms for statistical analysis.

Correspondence analysis applied to microarray data

The peculiarities of microarray data

Microarray technology provides access to expression levels of thousands of genes at once. However, there can be systematic errors induced by, for example, one needle always spotting a little less liquid than the other needles, a microtiter plate opened for a longer time and

[22] `http://mips.gsf.de`

[23] `http://www.mged.org/Workgroups/MIAME/miame.html`

[24] `http://www.dkfz.de/tbi/services/mchips`

[25] M-CHIPS databases do not share exactly the same database structure as shown in chapter 'Data Storage', section 'Database implementation'.

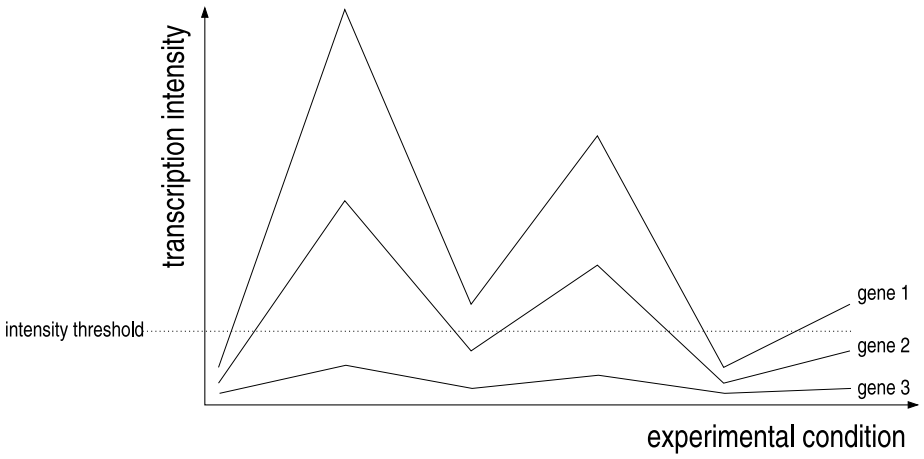

Figure 20: **Schematic gene profiles of related shape.** The plotted profiles are of similar shape, although showing different absolute values.

thus containing higher concentrations due to evaporation as well as lower or higher label incorporation rates. Additionally, these data show a substantial amount of random noise.

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 hybridizations. Therefore, the columns of the data table (hybridizations) have to undergo a normalization process, correcting for affine-linear transformation among the columns.

Another problem is that ratios are unreliable at low levels of intensity. Figure 20 shows three similar profiles. For biological interpretation it would be useful to cluster genes one and two together because they show the same expression behaviour although being expressed at different absolute levels. In this sense, the curve of gene three shows the same shape. However, commonly the majority of genes are not expressed to a measureable amount and their signals flicker with measurement noise. Let curve three resemble such a gene profile determined by chance rather than reflecting mRNA abundance. The set of untranscribed genes is often large enough to hold several genes for any profile under study. Therefore they have to be discarded by intensity filtering. It is also advisable to disregard all genes the values of which do not change in the entirety of conditions.

In spite of these measures, single gene profiles obtained by microarray hybridization are commonly verified by northern blot prior to further investigation of the genes. Let the expression profiles be prices in a supermarket. Thus a certain profile — e.g. a high expression level in the first and no abundance in the two remaining conditions — is represented by 2.49 Euro. This price will reliably identify the shampoo if only four articles have been bought. However, a microarray experiment is more like buying the entire store. Resulting data tables normally are only a few hybridizations wide, but maybe tens of thousands of genes long, having the shape of very long supermarket bill. Given a certain amount of noise, any particular gene profile will have a high probability of occuring somewhere in the list by chance.

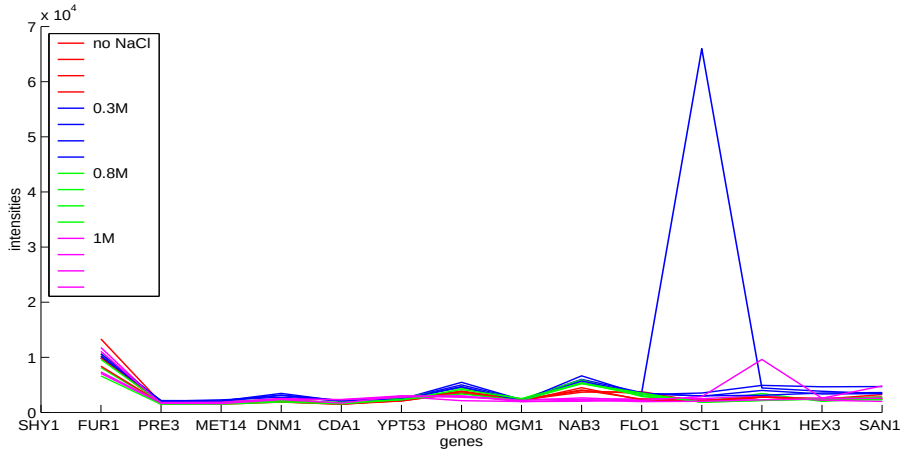

Figure 21: **Reproducibility of signals in a multiconditional experiment.** Wild type yeast was exposed to different concentrations of sodium chloride in the medium (see legend). Normalized transcription intensities of 14 genes are shown in a parallel coordinates plot, lines representing measurements (here hybridizations) 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. The corresponding image is provided in Appendix A (Fig. 22). 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.

Therefore, any microarray experiment is of low value if it does not comprise several repetitions. Repeated measurements allow the signals to be filtered according to reproducibility measures before analysis. Investigating the reproducibility of single signals by means of repeated measurements will lead to the identification and subsequent elemination of outliers (Fig. 21).

Adaption of correspondence analysis

For the above reasons, 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 by Fig. 21 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.

I prevent this by choosing the principal axes according to the condition medians only (HMS). My HMS technique furthermore allows the original data to be still visible in the plot thus combining the noise reduction capabiliy of HMS with the quality control aspect of retaining the original data. 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 [55].

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 [54] 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.

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. 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.

Applicability to microarray data

Traditionally, correspondence analysis has been used predominantly on categorical data in the social sciences [66,67], but its application has been extended also to (positive) physical quantities [56] and to proteomics [68,69]. I have shown that CA applied to microarray data provides an informative and concise means of visualizing these data, being capable of uncovering relationships both among either genes or hybridizations and between genes and hybridizations. The method has proved to be generally applicable to microarray data, regardless of whether

they have been obtained by radioactive labeling or by two-channel fluorescent-tag labeling. Normalization procedures lead to intensity values that can be interpreted as being proportional to a cell's content of mRNA molecules per gene and per condition. For two-channel intensities, the log ratios of red versus green channel appear to work just as well.

The CA analysis algorithms are embedded into a larger sofware package named M-CHIPS. The system enables preprocessing, e.g. different methods for normalization, the performance of which can be visually checked, quality control plots, and gene extraction by intensity, ratio and reproducibility thresholds. These thresholds can be applied to raw data, normalized data, ratios, minmax-separation, standard-deviation separation, condition-medians of fitted intensities or to ratios of these condition-medians. These categories can also be used for sorting the genes. Comparison of different visualizations of a dataset are facilitated by shared gene tags. It is also possible to mark all genes bearing a certain keyword such as 'cell cycle' in their gene annotation or to import multiple sets of gene tags from text lists. High-level analysis can be carried out for raw or fitted intensities, ratios, distances or ranks. In addition to CA, hierarchical clustering can be performed for comparison, permitting arbitrary combination of five distance measures with five linkage methods to cluster genes, measurements or condition-medians. Thus M-CHIPS not only integrates data stemming from different platforms for access by CA but also enables a highly flexible use of the CA routines as well as interaction with other algorithms.

Moreover, the platform enables integrated analysis of both transcription intensities and complex experiment annotations. Statistical analysis of experiment annotations can be applied for arbitrary sets of hybridizations in any CA plot by mouse click, e.g. for those clustered by similar expression profiles. This provides a means to reveal both experimental artifacts and biologically meaningful correlations from huge sets of experiment descriptions in an automated way. The resulting experimental parameters are candidates for being the active players which drive the cells to the expression pattern observed in the corresponding hybridization cluster. While this is a fairly simple method, it already provides good analytical access to long lists of annotations and huge sets of hybridizations which could not be thoroughly evaluated by visual inspection. More sophisticated statistical methods can also be directly applied because, unlike with free text annotation, instances of occurrence are readily countable for all annotation values. Future plans include integrated visualization of both transcription intensities and experiment annotations by correspondence analysis.

This work demonstrates the applicability of correspondence analysis to and high value for the analysis of microarray data, displaying associations between genes and experiments. It has been adapted to deal with the particular difficulties of microarray data, and its applicability has been further enhanced by incorporation into a framework of annotation data and supplemental algorithms. To introduce the method, I have shown its application to the well-known *S. cerevisiae* cell-cycle synchronization data by Spellman *et al.* [47], allowing for comparison

with their visualization of this data set. Furthermore I have applied correspondence analysis 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). This example also shows that CA is capable of subtracting particular effects, such as the influence of galactose in the medium.

Correspondence analysis versus other methods

Correspondence analysis is an explorative computational method for the study of associations between variables. For many other methods, unequal choice of parameters such as the number of clusters will focus on different properties of the transcription profiles and may therefore produce unequal results for the same data. CA does not require any choice of parameters. Similar to principal component analysis it displays a low dimensional projection of the data, e.g. into a plane. Thus it visualizes clusters of data points, also showing how fuzzy or defined cluster borders are. However, it does this for two variables simultaneously thus revealing associations between them.

Visualization using CA is based on representing the χ^2 distance among genes and among hybridizations, thus representing a decomposition of the value of the χ^2 statistic. Emphasis is placed on the genes and hybridizations that contribute to this value through their association. In this respect it resembles the doubly sorted hierarchical clusterings [70], although our examples demonstrate that CA is capable of revealing intricate detail, e.g. subtle discrepancies between phase classification and transcription pattern of hybridizations. The emphasis on association between genes and hybridizations distinguishes CA from other embedding methods such as principal component analysis or multidimensional scaling although these methods share the idea of representing objects in a two or three dimensional space that can be visualized. While CA and PCA use the same mathematical machinery for dimension reduction and visualization, namely singular value decomposition, their difference stems from the different distance measures used.

Alter *et al.* [43] successfully applied singular value decomposition to the analysis of the same data set that I used as my first example. In my 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.*[26] shows that this is not necessarily the case in the singular value decomposition alone. Moreover, CA does not depend on model assumptions as demonstrated in my second example. There genes were identified whose transcripts are specifically up-regulated in an overexpressing mutant yeast strain that is induced by galactose, whereas they are at normal levels (or even undetectable) in this mutant strain without galactose and likewise in wild type,

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

irrespective of the additive. Thus I find CA to be generally applicable to and particularly well-suited for gene expression data because of its ability to display simultaneously genes and hybridizations as well as the strength of their association.

Perspectives

M-CHIPS focuses on experiment annotations. Gene annotations are restricted to spot location, functional categories and external links. In the example given, an outlying cluster of hybridizations was successfully determined to be caused by a different array batch. It is not known whether this is due to different numbers of PCR cycles, different batches of polymerase or array support, spotting humidity or template contamination. Currently, efforts are being made to extend the M-CHIPS concept also to array production. Besides being a central laboratory information management system (LIMS) recording e.g. primers and clones, it is meant to further improve on the amount of parameters already amenable to statistical analysis. Future plans comprise compiling the MATLAB code to provide a homogeneous and easy to install software package and implementing an XML interface for data exchange with a public microarray data repository. Also, it would be interesting to compare the transcriptional information acquired to data on actual gene expression, for example obtained with protein arrays. M-CHIPS should be equally applicable to both kinds of data, since they share the same data structure.

CA is an intuitive means of explorative microarray data analysis. It shows which questions could be asked or which hypotheses could be put forward. It decomposes the overall χ^2-statistic, distances among points in the plots approximate χ^2-distances. Thus they resemble the statistical significance of an observed Euclidean distance with respect to all other distances measured, rather than the absolute distance itself. However, statistical significance is not a crucial requirement for justifying an inspection of the maps. Correspondence analysis can be regarded as a way of re-expressing the data in pictorial form for ease of interpretation [71]. In the present work, I used CA to transform data tables of genes versus measurements into two-dimensional maps. Alternatively, one could cluster genes or measurements into discrete groups of similar profiles [72]. Moreover, CA can be used in discriminant analysis and classification [73]. As already mentioned, CA has been applied prevalently to survey data in social sciences. To grasp the full potential of the technique, one should consider that using CA to simultaneously display genes and hybridizations is analogous to investigating a questionnaire comprising two questions. More than two variables can be integrated by applicative arrangement in a two-way table or by multiple or joint CA of multiway tables. For microarray data analysis this enables straightforward inclusion of additional information such as gene and experiment annotations, if stored in a format accessible to statistical analysis. Integration of gene- and experiment annotation data may enable to identifiy active players among

these parameters that drive the cells to the observed expression patterns. Sequence patterns in regulatory elements also show great promise for linking with microarray data [74, 75, 76]. Discovery of complex interdependencies between experimental parameters, gene properties, promotor sequences, mRNA abundance, and protein levels might be possible by integrated analysis of all these variables.

Acknowledgements

First and foremost, I would like to thank Nicole Hauser, Judith Boer, Marcel Scheideler, Frank Diehl, Susanne Diehl, Verena Aign, Albert Neutzner, Helene Tournu, Arno Meijer, Luis Lombardia, Manuel Beccera, Andy Hayes, Nikolaus Schlaich, José Pérez-Ortín, Arnaud Lagorce, Melanie Bier, Sonja Bastuck, Tamara Korica, Andrea Bauer, and Matthias Nees who entrusted the outcome of their hard work to me for data administration. Many thanks also for suggestions for improvement, critical discussions and for reported bugs. Especially, I am indebted to Nicole Hauser who did a great job as a beta-tester, attending M-CHIPS from the very beginning. I am further grateful to Tim Beißbarth and Dieter Finkenzeller for contributing algorithms.

I want to thank the supervisors of this work, Jörg Hoheisel and Martin Vingron, for their guidance and help and for their tolerance. I am also grateful to Prof. Dietmar Schomburg and Prof. Heinz Saedler, for correcting and co-correcting this thesis.

I want thank all members of the divisions "Functional Genome Analysis" and "Theoretical Bioinformatics", past and present, for a wonderful co-operative atmosphere, nice discussions and a lot of fun. Special thanks to Rainer Spang and Tobias Müller for mathematical enlightenment and Tim Beißbarth and Benedikt Brors for bioinformatical discussions. Antje Krause helped me in my first steps with PostgreSQL. I also enjoyed discussions with Helen Parkinson, Jaak Vilo, Ugis Sarkans and Alvis Brazma from the European Bioinformatics Institute.

The work relied on permanently running computer systems, which were administered by Tobias Reber, Heiko Schmidt and Karlheinz Groß. I am especially grateful to Karlheinz for being available also at the weekend in cases of emergency.

Many thanks to Leonie Ringrose, Nicole Hauser, Katharina Engbruch, Brigitte Engbruch, Marcus Frohme, Benedikt Brors, Jörg Hoheisel and Martin Vingron for critically reading the manuscript.

Katharina Engbruch and my parents Gerhild and Kurt Arthur Fellenberg were a constant source of support and encouragement.

Bibliography

[1] J. DeRisi, L. Penland, P. O. Brown, M. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.*, 14:457–460, 1996.

[2] J. Khan, M. Bittner, Y. Chen, P. S. Meltzer, and J. M. Trent. DNA microarray technology: the anticipated impact on the study of human disease. *Biochim. Biophys. Acta*, 1423:M17–M28, 1999.

[3] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.*, 21(Suppl.):33–37, 1999.

[4] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405:827–836, 2000.

[5] D. J. Duggan, M. Bittner, Y. Chen, P. Metzler, and J. M. Trent. Expression profiling using cDNA microarrays. *Nat. Genet.*, 21(Suppl.):10–14, 1999.

[6] 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.

[7] S. Braxton and T. Bedilion. The integration of microarray information in the drug development process. *Curr. Opin. Biotechnol.*, 9(6):643–9, Dec 1998.

[8] C. Debouck and P. N. Goodfellow. DNA microarrays in drug discovery and development. *Nat. Genet.*, 21(1 Suppl):48–50, Jan 1999.

[9] S. F. Dobrowolski, R. A. Banas, E. W. Naylor, T. Powdrill, and D. Thakkar. DNA microarray technology for neonatal screening. *Acta Paediatr. Suppl.*, 88(432):61–4, Dec 1999.

[10] K. Tang, D. J. Fu, D. Julien, A. Braun, C. R. Cantor, and H. Koster. Chip-based genotyping by mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, 96(18):10016–20, Aug 1999.

[11] 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.

[12] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, Jan 2002.

[13] G. G. Lennon and H. Lehrach. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.*, 7:314–317, 1991.

[14] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[15] M. Schena. Genome analysis with gene expression microarrays. *BioEssays*, 18:427–431, 1996.

[16] D. Shalon, S. J. Smith, and P. O. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, 6:639–645, 1996.

[17] D. J. Lockhart, M. Dong, M. C. Byrne, M. T. Folletie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, 14:1675–1680, 1996.

[18] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. Technical Report 584, Dept. of Statistics, UC Berkeley, CA, 2000.

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

[20] 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.

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

[22] 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.

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

[24] 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.

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

[26] 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.

[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] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] M. Vingron. Bioinformatics needs to adopt statistical thinking. *Bioinformatics*, 17(5):389–90, May 2001.

[47] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

[48] C. Ballard, D. Herreman, D. Schau, R. Bell, E. Kim, and A. Valencic. *Data modeling techniques for data warehousing.* IBM International Technical Support Organization, www.redbooks.ibm.com, San Jose, CA, 1998.

[49] C. Schönbach, P. Kowalski-Saunders, and V. Brusic. Data warehousing in molecular biology. *Briefings in Bioinformatics*, 1:190–198, 2000.

[50] K. Fellenberg, N. C. Hauser, B. Brors, J. D. Hoheisel, and M. Vingron. Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. *Bioinformatics*, in press.

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

[52] GOC. Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.

[53] GOC. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11(8):1425–1433, 2001.

[54] T. Beißbarth, K. Fellenberg, B. Brors, R. Arribas-Prat, J. M. Boer, N. C. Hauser, M. Scheideler, J. D. Hoheisel, G. Schütz, A. Poustka, and M. Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16:1014–1022, 2000.

[55] K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel, and M. Vingron. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, 98:10781–10786, 2001.

[56] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*, page 223. Academic Press, London, 1st edition, 1984.

[57] M. J. Greenacre. *Correspondence Analysis in Practice*, pages 181–183 and 36. Academic Press, London, 1st edition, 1993.

[58] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numer. Math.*, 14:403–420, 1970.

[59] V. E. Velculescu, S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, C. J. Wang, G. A. Beaudry, K. M. Ciriello, B. P. Cook, M. R. Dufault, A. T. Ferguson, Y. Gao, T.-C. He, H. Hermeking, S. K. Hiraldo, P. M. Hwang, M. A. Lopez, H. F. Luderer, B. Mathews, J. M. Petroziello, K. Polyak, L. Zawel, W. Zhang, X. Zhang, F. G. Zhou, W. Haluska, J. Jen, S. Sukumar, G. M. Landes, G. J. Riggins, B. Vogelstein, and K. W. Kinzler. Analysis of human transcriptomes. *Nat. Genet.*, 23:387–388, 1999.

[60] D. O. Morgan. Regulation of the apc and the exit from mitosis. *Nat. Cell Biol.*, 1(2):E47–53, Jun 1999.

[61] C. Godon, G. Lagniel, J. Lee, J. M. Buhler, S. Kieffer, M. Perrot, H. Boucherie, M. B. Toledano, and J. Labarre. The H2O2 stimulon in saccharomyces cerevisiae. *J. Biol. Chem.*, 273(35):22480–9, Aug 1998.

[62] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucl. Acids Res.*, 30(1):207–10, Jan 2002.

[63] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, and J. M. Cherry. The stanford microarray database. *Nucl. Acids Res.*, 29(1):152–5, Jan 2001.

[64] F. Salas, J. Haas, B. Brunk, C. J. Stoeckert Jr, and G. C. Overton. Epodb: a database of genes expressed during vertebrate erythropoiesis. *Nucl. Acids Res.*, 26(1):288–9, Jan 1998.

[65] C. J. Stoeckert Jr, F. Salas, B. Brunk, and G. C. Overton. Epodb: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucl. Acids Res.*, 27(1):200–3, Jan 1999.

[66] J . Blasius and M. J. Greenacre, editors. *Visualization of Categorical Data.* Academic Press, London, 1st edition, 1998.

[67] M. J. Greenacre and J. Blasius, editors. *Correspondence Analysis in the Social Sciences.* Academic Press, London, 1st edition, 1994.

[68] T. Pun, D. F. Hochstrasser, R. D. Appel, M. Funk, V. Villars-Augsburger, and C. Pellegrini. Computerized classification of two-dimensional gel electrophoretograms by correspondence analysis and ascendant hierarchical clustering. *Appl. Theoret. Electrophoresis*, 1:3–9, 1988.

[69] K. P. Pleissner, V. Regitz-Zagrosek, B. Krudewagen, J. Trenkner, B. Hocher, and E. Fleck. Effects of renovascular hypertension on myocardial protein patterns: analysis by computer-assisted two-dimensional gel electrophoresis. *Electrophoresis*, 19(11):2043–50, Aug 1998.

[70] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U.S.A.*, 96:9212–9217, 1999.

[71] M. J. Greenacre. *Correspondence Analysis in Practice*, chapter 10, pages 74–85. Academic Press, London, 1st edition, 1993.

[72] M. J. Greenacre. *Correspondence Analysis in Practice*, chapter 14, pages 111–118. Academic Press, London, 1st edition, 1993.

[73] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*, chapter 7, pages 185–206. Academic Press, London, 1st edition, 1984.

[74] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281(5):827–42, Sep 1998.

[75] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8(11):1202–15, Nov 1998.

[76] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:384–94, 2000.

[77] N. C. Hauser, M. Vingron, M. Scheideler, B. Krems, K. Hellmuth, K. D. Entian, and J. D. Hoheisel. Transcriptional profiling of all open reading frames of *Saccharomyces cerevisiae.* *Yeast*, 14:1209–1221, 1998.

[78] N. C. Hauser, K. Fellenberg, R. Gil, S. Bastuck, J. D. Hoheisel, and J. E. Perez-Ortin. Whole genome analysis of a wine yeast strain. *Comp. Funct. Genome.*, 2:69–79, 2001.

Appendix

A - Data examples

Data examples discussed in this thesis have been acquired by Spellman *et al.* [47], Nicole Hauser, Sonja Bastuck, Melanie Bier and Albert Neutzner.

Spellman *et al.* cell-cycle data

For direct comparability I used the list of cell-cycle regulated genes as displayed in [47] Fig. 1A (webversion incl. gene names[27]) as well as the displayed values from the associated web page[28]. Missing entries were treated as unchanged. The data were shifted to a positive range by adding the minimum value + 1. The resulting data table[29] was submitted to CA without further processing.

CDC14 overexpression

Sampling and hybridization. The yeast strains used were derivatives of W303 (*ade2-1, his3-11, 15, leu2-3, 112, trp1-1, ura3, ssd1*Δ*, can1-100, [psi+],ho*) which were either WT or overproducing Cdc14p after induction. The strains are referred to as WT or *CDC14* transgenic (*ura3::GAL1-MycCDC14-URA3 CLB2HA3*). Yeast cultures of both strains were grown in complete medium plus 2% raffinose to mid-logarithmic growth phase (OD_{600}=0.5) at which point nocodazole was added to a final concentration of 15 μg/ml. Samples were taken before addition of nocodazole and when synchronization of the cell culture was verified by microscopy. For overexpression of Cdc14p, cells were induced by 2% galactose and samples were taken after 1 h. Harvesting of cells for RNA preparation, radioactive labeling by reverse transcription, and hybridization onto the PCR-based whole genome DNA-array were performed as described [77].

[27] `http://genome-www.stanford.edu/cellcycle/figures/figure1Anames.html`
[28] `http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt`
[29] `http://www.dkfz-heidelberg.de/tbi/people/fellenberg/fig1.asc`

Normalization and filtering. The raw intensity data were obtained from AIS imaging software. Normalization was performed as described [54,55]. The normalized data matrix was filtered for genes with positive minmax separation for at least one of the conditions under study [54].

Data. A tab delimited table (comprehensive list) on the associated web page[30] lists both raw and normalized data for the primary and secondary spot of each gene on the filters. It also contains reproducibility measures and genewise medians used for filtering and correspondence analysis, respectively. The filtered set of 1402 genes are marked by an asterisk in the first column. The associated web page also provides the complete experiment annotations as well as detailed information about the elements spotted on the array.

Oxidative stress

Sampling and hybridization. Yeast strain FY1679 (*MATa/MATa ura3-52/ura3-52 trpD63/TRP1 leu2D1/LEU2 his3D200/HIS3 GAL2/GAL2*) was grown to mid-logarithmic growth phase, at which point the culture was split and hydrogen peroxide added to a final concentration of 200 mM. Samples were taken 5, 10, 15, 20 and 30 min after treatment. Cells were harvested for RNA preparation as described [77]. Radioactive labeling by reverse transcription and hybridization on the PCR-based whole genome DNA-array were also performed according to [77].

Normalization and filtering. The raw intensity data as obtained from AIS imaging software were normalized as described [54, 55]. After normalization the data were filtered for genes fulfilling the following criteria:

- Significant absolute intensity, i.e. normalized intensity of at least 5000 in at least one of the hybridizations.
- Significant relative change, i.e. normalized intensity divided by the median of normalized intensities for the control hybridizations of at least 4 or $<= 1/4$.
- Significant reproducibility of this maximum relative change, i.e. minmax separation of at least 1 for at least one of the conditions under study [54].

508 out of 6103 genes were extracted by the above constraints. In the data table[31] they are marked by an asterisk in the first column. The data have been subjected to correspondence analysis, further reducing measurement noise by HMS [55]. Planar embedding explains only

[30] `http://www.dkfz-heidelberg.de/tbi/services/mchips/cdc14.html`
[31] `http://www.dkfz-heidelberg.de/tbi/services/mchips/ox_stress.asc`

76.4% (50.3% without HMS) of the total variance within this dataset, demonstrating the ability of CA to show the major variances among the data and overlook minor changes.

Data. A tab delimited table[31] both provides the raw intensities as computed by AIS imaging software and the preprocessed data for high-level analysis. The table lists raw and normalized data for the primary and secondary spot of each gene on the filters. It also contains reproducibility measures and gene-wise medians used for filtering and correspondence analysis, respectively. The filtered set of 508 genes is marked by an asterisk in the first column. A HTML document[32] holds the experiment annotations for the described experiment. It is divided into condition-dependent, measurement-dependent and constant annotations and comprises also the measurement-dependent annotations for measurement no. 60 and 61 (different array series) disregarded in the analysis shown by Fig. 17 but included in Fig. 16.

Sodium chloride concentration series

These data were used to exemplify an outlying intensity signal (Discussion, Fig. 21). Experiment annotations are listed on an HTML page[33]. The raw intensity data as obtained from AIS imaging software were normalized as described [54, 55]. Figure 22 shows an area of low quality of the scanned image. The outlier shown in Fig. 21 is marked by a blue arrow. The corresponding replicate spot, expected to show about the same signal intensity, is marked by a green arrow.

[31] `http://www.dkfz-heidelberg.de/tbi/services/mchips/ox_stress.asc`
[32] `http://www.dkfz-heidelberg.de/tbi/services/mchips/ox_stress.html`
[33] `http://www.dkfz-heidelberg.de/tbi/people/fellenberg/diss/NaCl_conc_series.html`

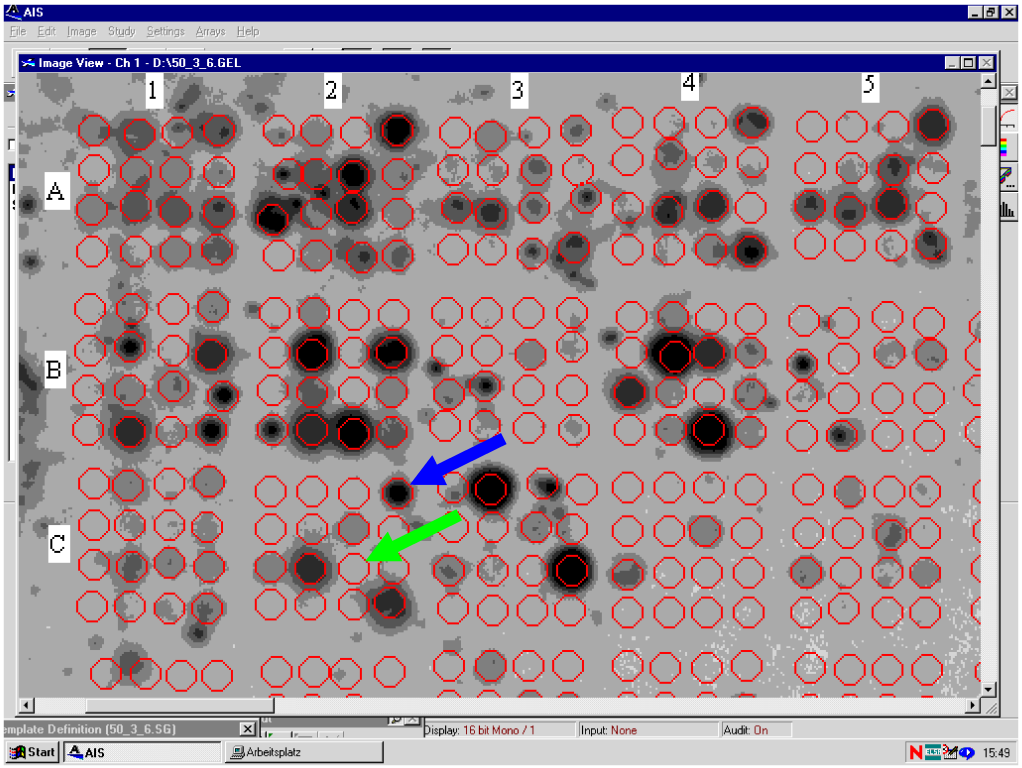

Figure 22: **Outlying measurement using radioactive label.** This particular area of the filter shows a high amount of unspecifically bound aggregated label. These 'stars' differ from the mostly larger spots in that they show a much higher intensity than spots of comparable size. At first sight the outlier shown in Fig. 21 in the discussion (blue arrow) appears as a spot. Its large size and location almost perfectly in a grid position lead to this impression. However, comparison with real spots show that the staining is much too intense for the corresponding area. The corresponding secondary spot (green arrow) explains the immense difference of the primary spot intensity to all other values for the gene *SCT1* shown in Fig. 21.

B - MATLAB^{TM} implementation

Algorithm 1 lists a short implementation of simple correspondence analysis. More complex ones, adapted to the requirements of microarray data analysis, are at the core of M-CHIPS.

```
 P=N./sum(sum(N));                % correspondence matrix P
 q=sum(p')'*sum(p);               % q=r_i c_j
 [U,D,V]=svd((P-q)./sqrt(q),0);   % ``,0'' produces economy size decomposition,
                                  % can be omitted
 F=diag(1./sqrt(sum(P')))*U*D;    % gene coordinates
 G=diag(1./sqrt(sum(P)))*V*D;     % hybridization coordinates

% The following statements plot out the above computed coordinates:
 plot(F(:,1),F(:,2),'.k'); hold on;            % plots ...
 text(F(:,1),F(:,2),num2str([1:size(F,1)]'));  % & annotates genes
 plot(G(:,1),G(:,2),'s');                      % and hybridizations
 text(G(:,1),G(:,2),num2str([1:size(G,1)]'));

% In order to correctly display χ²-distances, the two axes have to be scaled
% identically:
 from=min([get(gca,'xlim') get(gca,'ylim')]);           % determines ...
 to=max([get(gca,'xlim') get(gca,'ylim')]);             % ...  max.  range
 set(gca,'xlim',[from to]); set(gca,'ylim',[from to]);  % fixes range
```

Algorithm 1: A simple correspondence analysis program in MATLAB

C - WWW presentation of analysis results

Results can be stored in the database by the analyst (e.g. the experimenter), which automatically makes them available via Internet. Here they are, however, protected by passwords. Access can be granted on individual sets e.g. to collaborating persons. A password is valid in conjunction with a user name for accessing a particular array family (i.e. type of array) within a certain (organism-specific) database only. Some results, however, have been copied to a conventional web page to make them publicly available:

- ***CDC14* induction**. The dataset comprises one multiconditional yeast experiment (13 hybridizations, complete array). It is discussed in this thesis and in [55].
  `http://www.dkfz-heidelberg.de/tbi/services/mchips/cdc14.html`

- **Oxidative stress (timecourse).** The dataset consists of one multiconditional yeast experiment (14 hybridizations, complete array). It is discussed in this thesis and in the supplemental material to ref. [50].
  `http://www.dkfz-heidelberg.de/tbi/services/mchips/supplement_bioinformatics.html`

- **Eurofan II, B2.** The dataset comprises 36 multiconditional yeast experiments (253 hybridizations, complete array) dealing with filamentous growth, ceramide signalling, stresses, hypoxia, PKC pathway, C and N-limited growth in continous culture and other conditions. It was aquired by Alistair Brown and Helene Tournu (Aberdeen, GB), Rudi Planta and Arno Meijer (Amsterdam, NL), Esperanza Cerdan and Manuel Becerra and Luis Lombarda (La Coruna, ES), Joerg Hoheisel and Nicole Hauser (Heidelberg, DE), and Steve Oliver and Andy Hayes (Manchester, GB). Nicole Hauser collected and uploaded the intensity data. They were annotated by the experimenters themselves via web annotator, processed in M-CHIPS and presented by the Munich Information Center for Protein Sequences (MIPS).
  `http://mips.gsf.de/proj/eurofan/eurofan_2/b2`

- **Heatshock-timecourse.** The dataset comprises one multiconditional yeast experiment (12 hybridzations, complete array, [54])
  `http://www.dkfz-heidelberg.de/funct_genome/yeast-data.html\#heat-shock`

- **Wine-yeast versus laboratory strain.** A transcriptional and genomic comparison was carried out on all open reading frames of the wine yeast strain T73 and a standard laboratory strain (MYC730) as described [78]:
  `http://www.dkfz-heidelberg.de/funct_genome/yeast-data.html\#wine`

Such a result may either represent a single multiconditional experiment or is a recombination of conditions and/or measurements stemming from different multiconditional experiments. Experiment annotations can be viewed both in the original form, i.e. for the experiments and recombined for the results. For each result, the data comprised have been at least normalized and filtered. Visually supervised normalization as well as choice of filter parameters have been performed by the analyst named in the result header, who stored it in the database. There may be two different types of results, namely a 'color coded list' or a 'complete list'.

Color coded list

These lists comprise factors of relative changes with respect to the control condition and the median value of normalized signal intensities. Significance levels were assessed by two reproducibility criteria as described [54]. The highly stringent 'minmax separation' is calculated by taking the minimum pairwise distance between data points of one condition and data points of the control condition. The less stringent criteria, called 'standard-deviation separation', is defined as the difference of the means of the two data sets diminished by one standard deviation. Nicole Hauser developed the idea of color-coding the factors in these lists according to the two stringency measures. Data are classified as being of high or medium significance (Fig. 23).

Complete list

A complete list shows the complete data as tab-delimited ASCII file. The tables include raw and fitted (i.e. normalized) intensities for each individual measurement as well as ratios which were calculated by division of each individual hybridization by the gene wise median of all control hybridizations for monochannel data and by division of each channel by the corresponding control channel of the same hybridization for multichannel data. In addition the medians and reproducibility measures [54] for repeat measurements are provided.

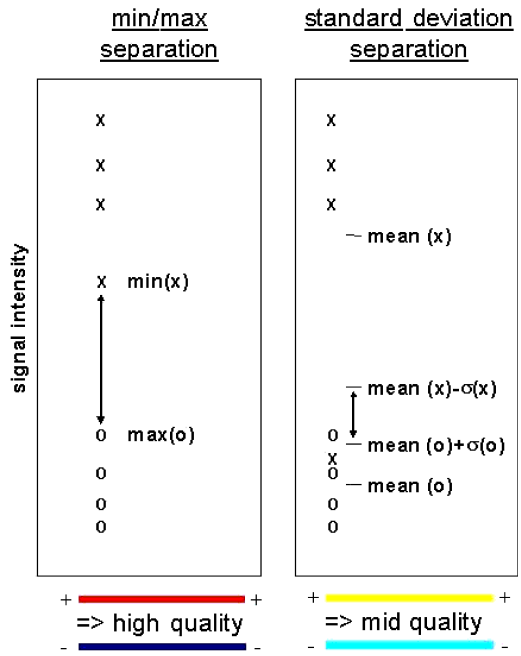

Figure 23: **Reproducibilty color-coding.** Upregulations are marked by warm (red and yellow), downregulations by cold colors (blue and light blue); high reproducibility is shown by dark (red and blue), medium reproducibility by light colors (yellow and light blue). If none of the two criteria is fulfilled the signal should by regarded as not reliable - even if a high ratio is displayed - and is tagged white in the color coded list. From [54], modified.

Software used

The following software packages were used. The programs were run on a Sun Ultra 10 workstation or Sun E 450 server machine under the Solaris operating system or on a HP Superdome mainframe computer[34] under HPUX:

- PostgreSQL (Version 6.5.3): Open source object-relational DBMS (database management system) suited for transaction-based multi-user database operation. Publicly available at `http://www.postgresql.org`

- Apache (Version 1.3): Open source WWW server program. In the context of the present work, both HTML (hypertext markup language, Version 4.0 specifications) documents and CGI (common gateway interface) programs (C and Perl code) have been created. Publicly Available at `http://www.apache.org`

- Matlab (Version 6.0) incl. Statistics Toolbox: Interpreted numerical programming environment suited for matrix algebra. MathWorks Inc. MA, U.S.A.

- Perl (Version 5.004_04): Interpreted programming language suited e.g. for regular expression search in texts developed by Larry Wall. Publicly available at `http://www.perl.com`

- I am grateful to Tim Beissbarth for permitting me to incorporate fdiffs, a scatterplot-based tool suited for **f**inding **diff**erentially transcribed gene**s** into the M-CHIPS data platform.

- Futhermore, I thank Dieter Finkenzeller for interfacing cl-vishn, his OpenGL-accelerated 3D-visualization tool, to show M-CHIPS-generated three-dimentional CA projection plots. The interactive 3D plots allow for seamless rotation and return sets of genes that have been selected by the user.

[34] Recently among the 200 fastest computers in the world (`http://www.top500.org`, February 2002). More information about the Superdome installation at the German Cancer Research Center can be found at `http://www.dkfz.de/zdv`

Our data examples that are discussed in this thesis were acquired by Nicole Hauser, Sonja Bastuck, Melanie Bier and Albert Neutzner using the AIS imaging software (Version 3.0, Array Vision Module, Imaging Research Inc., St. Catherines, Canada) running on a Pentium PC under Windows NT. The program automatically detects spots and calculates intensity values.

Zusammenfassung

Die Zelle bewirkt Stoffwechselvorgänge und die genaue Einhaltung ihres Wachstumsprogramms, Anpassungen an veränderte Umgebungsbedingungen oder Kommunikation mit anderen Zellen im Organverbund durch feinregulierte Ausprägung eines Portfolios von Proteinen. Dieses ist genau auf die jeweiligen Einflüsse abgestimmt. Hierzu wird die im Zellkern als DNS gespeicherte Sequenzinformation kopiert und gelangt als Boten-RNS in das den Kern umgebende Zellplasma, wo sie die Synthese des von ihr codierten Proteins bestimmt. Dabei sind Vorhandensein bzw. Menge der Boten-RNS eine wichtige Regulationsgröße für Vorhandensein und Menge des jeweiligen Proteins.

Die in einem bestimmten Zustand der Zelle vorhandene Menge an Boten-RNS ist mittels DNS-Chip-Technologie erstmals für tausende, in manchen Fällen für alle Gene eines Organismus zur gleichen Zeit meßbar. Dazu wird aus den zu untersuchenden Zellen die RNS isoliert und unter Einbau radioaktiv- oder fluoreszenzmarkierter Bausteine in DNS umgeschrieben, welche besser handhabbar ist. Bei der sog. "Hybridisierung" wird sie auf einen DNS-Chip aufgebracht.

Dieser besteht aus einem Glasplättchen oder einer Nylonfolie als Trägermaterial, auf das DNS-Fragmente unterschiedlicher Sequenz punktförmig aufgetupft wurden. Diese DNS Punkte (sog. "Spots") dienen zur Mengenbestimmung der anschließend aufgebrachten, aus der Aufbereitung des Biomaterials stammenden DNS. Da sich DNS gleicher Sorte zu Doppelsträngen verbindet ("hybridisiert"), ist auf einem solchen Spot an der Menge hybridisierter markierter DNS Vorhandensein bzw. Menge der Boten-RNS für dieses Protein ablesbar. Auf manchen Chips sind für einen bestimmten Organismus alle proteincodierenden DNS-Fragmente als Spot vertreten.

Die für alle Gene einzeln erfaßte Menge an Boten-RNS nennt man "Transkriptionsstatus" der Zelle bzw. "Transkriptom". In einer Meßreihe kann z.B. für mehrere experimentelle Bedingungen, denen die Zellen ausgesetzt werden, einzeln der Transkriptionsstatus per DNS-Chip-Hybridisierung festgestellt werden.

Der Nutzen dieser Technik beruht neben der gleichzeitigen Erfassung vieler Gene auf ihrer breiten Anwendbarkeit. Sie wird in der Grundlagenforschung z.B. zum Studium der Funktion einzelner Gene durch Untersuchung von Organismen verwendet, in denen dieses Gen inaktiviert wurde. Weiterhin können DNS-Chips in der pharmazeutischen Forschung zum

Auffinden oder zum Design therapeutisch wirksamer Substanzen, sowie in der Medizin zum Testen genetischer Veranlagung für bestimmte Erbkrankheiten oder zur Krebsdiagnostik eingesetzt werden. Allerdings werden mit dieser Methode sehr große Datenmengen produziert, deren Interpretation bisher noch ein Problem darstellt.

Unterschiedliche Hintergrundsignalstärken bewirken einen additiven Signalwertversatz, andere technisch bedingte systematische Fehlerquellen wie Einsatz unterschiedlicher Gesamtmengen an RNS oder verschiedene Einbauraten bei der Markierung der RNS führen zu multiplikativen Unterschieden zwischen Hybridisierungen. Diese linearen Versätze der Signalstärken werden durch "Normalisierung" vor einer weiterführenden Datenanalyse korrigiert. Außerdem gehört zur Vorverarbeitung der Daten das Herausfiltern aller Gene, die in keiner der untersuchten Bedingungen signifikante (bzw. reproduzierbare, s.u.) Signaländerungen erfahren.

Neben systematischen Abweichungen zeigen die Daten einen beträchtlichen Rauschpegel, die Proportionen der Signale im unteren Intensitätsbereich sind unzuverlässig und die Datensätze enthalten gewöhnlich Ausreißer. Weiterhin ist die Anzahl der Spots auf einem Chip gegenüber der Anzahl experimenteller Bedingungen überproportional hoch. Das führt dazu, daß eine Menge von Genen, für die über eine Serie von experimentellen Bedingungen ein bestimmtes Genprofil gemessen wurde, eine erhebliche Anzahl falsch-positiver Gene enthält, die zufallsbedingt aufgrund des Meßrauschens und der großen Anzahl Gene auftreten.

Abgesehen von der Fähigkeit, sowohl Gene als auch experimentelle Bedingungen gleichzeitig zu visualisieren, müssen deshalb von einer weiterführenden Analyse zusätzliche Anforderungen erfüllt werden. Um die statistische Signifikanz von Beobachtungen zu erhöhen, wird jede experimentelle Bedingung wiederholt, d.h. durch mehrere Hybridisierungen vermessen. Die Anzahl dieser Wiederholungen ist oft klein, so daß es nicht empfehlenswert ist, diese durch ihren Durchschnitt und ihre Standardabweichung zu repräsentieren. Eine Analysemethode muß deshalb wiederholte Messungen integrieren können. Ausserdem muß sie in der Lage sein, Meßrauschen zu unterdrücken und Ausreißer zu tolerieren.

Die vorliegende Arbeit stellt ein System zur intelligenten Speicherung von DNS-Chip-Daten vor, sowie Methoden zur Datenanalyse von DNS-Chip-Messreihen mit hohem Datenvolumen.

Das System ermöglicht eine sorgfältige Vorbehandlung der Daten. Ein Kernpunkt ist die Anwendung einer speziellen statistischen Analysemethode, die das Studium von Abhängigkeiten sowohl innerhalb als auch zwischen Variablenmengen — hier Genen und DNS-Chip Hybridisierungen — erlaubt.

Dieses Verfahren, die "Korrespondenzanalyse", ist an die speziellen Anforderungen von DNS-Chip-Daten so angepasst worden, daß wiederholte Messungen bei Rauschunterdrückung und Toleranz gegen Ausreißer integriert werden können, ohne auf die Visualisierung jeder einzelnen Messung zu verzichten.

Das Verfahren ist eine Projektionsmethode. Ähnlich wie bei der Hauptkomponentenanalyse

erhält man eine niedrigdimensionale Projektion hochdimensionaler Daten. Die Korrespondenzanalyse tut dies allerdings gleichzeitig für Gene und Hybridisierungen, so daß Assoziationen zwischen einzelnen Genen und Experimenten sichtbar werden. Die vorliegende Arbeit demonstriert die Anwendbarkeit der Korrespondenzanalyse auf und den hohen Nutzen für die Analyse von DNS-Chip-Daten. Um die Methode einzuführen, wird ihre Anwendung auf einen bekannten, publizierten Datensatz, die Hefe-Zellzyklus-Synchronisation von Spellman *et al.* (Mol. Biol. Cell 9 (1998), 3273-3297), gezeigt und die Visualisierung durch Korrespondenzanalyse mit der in der Originalpublikation gewählten Darstellung verglichen. Zusätzlich wird die Korrespondenzanalyse auf ein Experiment aus unserer Arbeitsgruppe angewandt, um ihre Eignung für Daten unterschiedlicher Komplexität, Struktur und experimenteller Technik (Zweikanal-Fluoreszenz- bzw. radioaktive Markierung) zu demonstrieren.

Eine für die Analyse von Chip-Daten geeignete Methode muß uneingeschränkten Zugriff auf die Daten haben. Dazu sollte sie in eine Datenbankplattform integriert sein, die große Datensätze in einem definierten Format für Datenvorverarbeitung und Analyse bereithält.

Neben Speicherung und Analyse der Hybridisierungssignale sind für die biologische Interpretation der daraus resultierenden Ergebnisse Informationen über die auf dem Chip repräsentierten Gene sowie eine genaue Beschreibung der untersuchten experimentellen Bedingungen unerläßlich. Für die Interpretation großer Datensätze sollten diese Beschreibungen ("Annotationen") in einem für computerbasierte Analyse geeigneten Format vorliegen, da eine visuelle Auswertung am hohen Datenaufkommen scheitert. Die Einbeziehung solcher experimentellen Parameter in eine statistische Analyse eröffnet die Möglichkeit, biologisch bedeutsame, den Transkriptionsmustern zugrundeliegende Zusammenhänge oder Mechanismen zu identifizieren.

Die Freitext-Annotation heutiger DNS-Chip-Datenbanken behindert einen direkten Zugriff auf diese Daten mittels statistischer Methoden. Experiment-Ontologien haben ihre endgültige Form noch nicht erreicht und Standards sind reduziert auf Minimalkonventionen, die für ausführliche Experimentbeschreibungen ungeeignet sind. Komplexe und hochvariante experimentelle Szenarien verursachen eine hohe Komplexität und Vielgestaltigkeit experimenteller Annotationen und erfordern daher eine flexibleres Speicherkonzept als das einer Standard-Datenbanklösung. Dies gilt insbesonders, wenn die Daten in einem mit statistischen Methoden zugreifbaren Format vorliegen sollen. Eine ontologieunabhängige Datenbankstruktur erlaubt die Aktualisierung von Beschreibungshierarchien während des normalen Datenbankbetriebs. Eine Änderung der Datenbankstruktur ist hierzu nicht erforderlich.

Ich habe ein Sytem entwickelt und implementiert, welches den oben angeführten Anforderungen genügt. Es integriert die Korrespondenzanalyse in ein größeres Rahmenwerk aus Datenbank-Plattform und ergänzenden Algorithmen. Der Name M-CHIPS steht für "Multi-Conditional Hybridization Intensity Processing System". Es erlaubt eine direkte statistische Analyse aller erfaßten Daten inklusive der experimentellen Annotationen. Es berücksichtigt das schnelle

Anwachsen des Datenvolumens für die Hybridisierungsdaten, Experimentbeschreibungen beliebigen Detailgrades sowie zukünftige Experimentszenarien. Es stellt ein universelles Speicherkonzept zur Verfügung. Eine Organismus-spezifische Datenbank ist eine einzelne Instanz dieses Konzepts. Obwohl solche Datenbanken unterschiedliche Systeme experimenteller Beschreibungen enthalten, haben sie gleiche Datenbankstruktur, was den Zugriff durch dasselbe Algorithmenpaket ermöglicht.

Kurzzusammenfassung

DNS-Chips ("Microarrays") ermöglichen global ausgelegte transkriptionelle Studien durch gleichzeitige Erfassung mehrerer zehntausend Gene. Hierdurch werden große Datenmengen produziert. Diese Daten enthalten Meßrauschen, Signale im unteren Intensitätsbereich sind unzuverlässig und die Datensätze enthalten gewöhnlich Ausreißer. Darüberhinaus ist die Anzahl der Spots auf einem Chip gegenüber der Anzahl experimenteller Bedingungen überproportional hoch. Das führt dazu, daß eine Menge von Genen, für die über eine Serie von experimentellen Bedingungen ein bestimmtes Genprofil gemessen wurde, eine erhebliche Anzahl falsch-positiver Gene enthält, die zufallsbedingt aufgrund des Meßrauschens und der großen Anzahl Gene auftreten. Abgesehen von der Fähigkeit, sowohl Gene als auch experimentelle Bedingungen gleichzeitig zu visualisieren, muß eine weiterführende Analyse deshalb mehrere, für eine experimentelle Bedingung wiederholt ausgeführte Messungen integrieren können, und außerdem in der Lage sein, Meßrauschen zu unterdrücken und Ausreißer zu tolerieren.

Die vorliegende Arbeit stellt ein System zur intelligenten Speicherung von DNS-Chip-Daten vor, sowie Methoden zur Datenanalyse von DNS-Chip-Messreihen mit hohem Datenvolumen. Das System ermöglicht eine sorgfältige Vorbehandlung der Daten. Ein Kernpunkt ist die Anwendung einer speziellen statistischen Analysemethode, die das Studium von Abhängigkeiten sowohl innerhalb als auch zwischen Variablenmengen — hier Genen und DNS-Chip Hybridisierungen — erlaubt. Dieses Verfahren, die "Korrespondenzanalyse", ist an die speziellen Anforderungen von DNS-Chip-Daten so angepasst worden, daß wiederholte Messungen bei Rauschunterdrückung und Toleranz gegen Ausreißer integriert werden können, ohne auf die Visualisierung jeder einzelnen Messung zu verzichten.

Die vorliegende Arbeit demonstriert die Anwendbarkeit der Korrespondenzanalyse auf und den hohen Nutzen für die Analyse von DNS-Chip-Daten. Zur Einführung wird ihre Anwendung auf einen bekannten, publizierten Datensatz, die Hefe-Zellzyklus-Synchronisation von Spellman *et al.* (Mol. Biol. Cell 9 (1998), 3273-3297), gezeigt und die resultierende Visualisierung mit der in der Originalpublikation gewählten Darstellung verglichen. Zusätzlich wird die Korrespondenzanalyse auf ein Experiment aus unserer Arbeitsgruppe angewandt, um ihre Eignung für Daten unterschiedlicher Komplexität, Struktur und experimenteller Technik (Zweikanal-Fluoreszenz- bzw. radioaktive Markierung) zu demonstrieren.

Eine für die Analyse von Chip-Daten geeignete Methode sollte uneingeschränkten Zugriff auf die Daten haben. Ich habe ein Sytem entwickelt und implementiert, welches die Korrespondenzanalyse in ein größeres Rahmenwerk aus Datenbank-Plattform und ergänzenden Algorithmen integriert. Der Name M-CHIPS steht für "Multi-Conditional Hybridization Intensity Processing System". Es erlaubt eine direkte statistische Analyse aller erfaßten Daten inklusive der experimentellen Annotationen. Es berücksichtigt das schnelle Anwachsen des Datenvolumens für die Hybridisierungsdaten, Experimentbeschreibungen beliebigen Detailgrades sowie zukünftige Experimentszenarien. Es stellt ein universelles Speicherkonzept zur Verfügung. Obwohl einzelne, organismus-spezifische Datenbanken unterschiedliche Systeme experimenteller Beschreibungen enthalten, haben sie gleichartige Datenbankstruktur, was den Zugriff durch dasselbe Algorithmenpaket ermöglicht.

Lebenslauf

Personalien

Name	Kurt Fellenberg
Geburtsdatum	25. Februar 1970
Geburtsort	Gummersbach
Familienstand	ledig
Eltern	Gerhild (geb. Ganschinietz) und Kurt Arthur Fellenberg
Erster Wohnsitz	Tilsiter Str. 14, 51643 Gummersbach

Werdegang

Schule	1976-1980	Grundschule Gummersbach-Steinenbrück
	1980-1989	Gymnasium Moltkestraße Gummersbach
	Abitur:	19. Mai 1989
Studium	1990 - 1997	Studium der Biologie an der Universität zu Köln
	1996:	Studentische Hilfskraft am Zentrum für Paralleles Rechnen
	1996:	Laborpraktikum in der Abteilung von Prof. Dr. Klaus Rajewsky, Institut für Genetik
	1996-1998:	Diplomarbeit, Titel: "Optimierung von Cre-Östrogenrezeptor-Fusionsproteinen zur lokalen Geninaktivierung in der Haut", Betreuer: Prof. Dr. Klaus Rajewsky, Dr. Werner Müller
	Vordiplom:	30. September 1993, Universität zu Köln
	Diplom:	28. Januar 1998, Universität zu Köln
Promotion	Seit April 1998	an der Universität zu Köln, Mathematisch-Naturwissenschaftliche Fakultät, Betreuer: Prof. Dr. Dietmar Schomburg zweiter Referent: Prof. Dr. Heinz Saedler. Die experimentellen Arbeiten wurdern nach §6 Abs. 2 der Promotionsordnung der Universität zu Köln am Deutschen Krebsforschungszentrum in Heidelberg unter der praktischen Anleitung von Prof. Dr. Martin Vingron durchgeführt.

Begutachtete Publikationen

- Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis.
  Fellenberg K, Hauser NC, Brors B, Hoheisel JD, Vingron M.
  Bioinformatics, im Druck.

- Monitoring the switch from housekeeping to pathogen defence metabolism in *Arabidopsis thaliana* using cDNA arrays.
  Scheideler M, Schlaich NL, **Fellenberg K**, Beißbarth T, Hauser NC, Vingron M, Slusarenko AJ, Hoheisel JD.
  J Biol Chem, im Druck.

- ArrayExpress; establishing a public repository for DNA array-based gene expression data.
  Brazma A, Robinson A, Vilo J, Vingron M, Hoheisel JD, **Fellenberg K**, Muilu J.
  In: DNA-Chip Technology; Advances in Biochemical Engineering/Biotechnology, im Druck.

- Correspondence anaylsis applied to microarray data.
  Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M.
  Proc Natl Acad Sci USA. 2001. **98**(19): 10781-10786.

- Whole genome analysis of a wine yeast strain.
  Hauser NC, **Fellenberg K**, Gil R, Bastuck S, Hoheisel JD, Perez-Ortin JE.
  Comp Funct Genom. 2001. **2**(4): 69-7.

- Processing and quality control of DNA array hybridization data.
  Beißbarth T, **Fellenberg K**, Brors B, Arribas-Prat R, Boer JM, Hauser NC, Scheideler M, Hoheisel JD, Schütz G, Poustka A, Vingron M.
  Bioinformatics. 2000. **16**(11): 1014-1022.

- DNA microchips; transcriptional profiling and beyond.
  Beier M, Matysiak S, Hauser N, Scheideler M, Würtz S, **Fellenberg K**, Vingron M, Hoheisel JD.
  Chimia. 2000. **54**: 29-30.

Konferenzbeiträge

- Correspondence analysis applied to microarray data.
  Fellenberg K.
  Eingeladener Vortrag, 25. Jahrestagung der Gesellschaft für Klassifikation. München, 14.-16. März 2001.

- The use of bioinformatics in the interpretation of multiple comparisons.
  Fellenberg K.
  Eingeladener Vortrag, EuroBiochips Konferenz. Hamburg, 5.-7. Juni 2000.

Andere Veröffentlichungen

- Technischer Bericht: M-CHIPS storage concept.
  Fellenberg K.
  http://www.dkfz-heidelberg.de/tbi/people/fellenberg/report/report.pdf

- Poster: Transcription Profiling using DNA Array Data.
  Hauser N, **Fellenberg K**, Beißbarth, T.
  http://www.dkfz-heidelberg.de/tbi/people/fellenberg/NormalizationPoster.pdf

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit — einschließlich Tabellen, Karten und Abbildungen —, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie — abgesehen von unten angegebenen Teilpublikationen — noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Universität zu Köln sind mir bekannt.

Die von mir vorgelegte Dissertation ist von Prof. Dr. Martin Vingron angeleitet worden. Die Arbeit wurde gemäß § 6 Abs. 2 dieser Promotionsordnung am Deutschen Krebsforschungszentrum in Heidelberg durchgeführt. Betreuer nach § 5 Abs. 2 ist Prof. Dr. Dietmar Schomburg, zweiter Referent ist Prof. Dr. Heinz Saedler.

Teilpublikationen:

K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel und M. Vingron. Correspondence anaylsis applied to microarray data. *Proc. Natl. Acad. Sci. USA.*, 98(19):10781-10786, 2001.

K. Fellenberg, N. C. Hauser, B. Brors, J. D. Hoheisel und M. Vingron. Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. *Bioinformatics*, im Druck.

T. Beißbarth, K. Fellenberg, B. Brors, R. Arribas-Prat, J. M. Boer, N. C. Hauser, M. Scheideler, J. D. Hoheisel, G. Schütz, A. Poustka und M. Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16(11):1014-1022, 2000.

M. Beier, S. Matysiak, N. Hauser, M. Scheideler, S. Würtz , K. Fellenberg, M. Vingron und J. D. Hoheisel. DNA microchips; transcriptional profiling and beyond. *Chimia*, 54:29-30, 2000.

N. C. Hauser, K. Fellenberg, R. Gil, S. Bastuck, J. D. Hoheisel und J. E. Perez-Ortin. Whole genome analysis of a wine yeast strain. *Comp. Funct. Genom.*, 2(4):69-7, 2001.

M. Scheideler, N. L. Schlaich, K. Fellenberg, T. Beißbarth, N. C. Hauser, M. Vingron, A. J. Slusarenko und J. D. Hoheisel. Monitoring the switch from housekeeping to pathogen defence metabolism in *Arabidopsis thaliana* using cDNA arrays. *J. Biol. Chem.*, im Druck.

J. D. Hoheisel, K. Fellenberg, B. Brors, F. Diehl, N. C. Hauser und M. Vingron. Transkriptionelle Untersuchungen auf DNS-Microarrays; Aspekte der Daten-Evaluation und Interpretation. *BIOforum*12/01:908-910, 2001.

K. Fellenberg. M-CHIPS storage concept. Technischer Bericht:
`http://www.dkfz-heidelberg.de/tbi/people/fellenberg/report/report.pdf`

K. Fellenberg und N. C. Hauser. Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. Patentanmeldung 01 123.732.8 vom 4.10.2001 (Prioritäts-Datum).

Bewerbung für den Wissenschaftspreis 2002 des Wissenschaftszentrums Nordrhein-Westfalen und des Industrie-Clubs Düsseldorf.

A. Brazma, A. Robinson, J. Vilo, M. Vingron, J. D. Hoheisel, K. Fellenberg und J. Muilu. ArrayExpress; establishing a public repository for DNA array-based gene expression data. In *DNA-Chip Technology; Advances in Biochemical Engineering/Biotechnology*, im Druck.

K. Fellenberg, M. Vingron und J. D. Hoheisel. Correspondence analysis with microarray data. In *Perspectives in Gene Expression*. Editor: K. Appasani. Eaton Publishing/BioTechniques, in Vorbereitung.

Köln, den 26. Februar 2002