BIOINFORMATICS

Vol. 18 no. 3 2002
Pages 423–433

Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis

Kurt Fellenberg[1, 2,*], Nicole C. Hauser[2], Benedikt Brors[1, 2], Jörg D. Hoheisel[2] and Martin Vingron[1,3]

[1]Department of Theoretical Bioinformatics and [2]Department of Functional Genome Analysis, German Cancer Research Center, PO Box 101949, D-69009 Heidelberg, Germany

Received on April 18, 2001; revised on August 22, 2001; accepted on October 29, 2001

ABSTRACT

Motivation: Microarray technology provides access to expression levels of thousands of genes at once, producing large amounts of data. 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.

Results: The Multi-Conditional Hybridization Intensity Processing System (M-CHIPS), a data warehousing concept, focuses on providing both structure and algorithms suitable for statistical analysis of a microarray database's entire contents including the experiment annotations. It addresses the rapid growth of the amount of hybridization data, more detailed experimental descriptions, and new kinds of experiments in the future. We have developed a *storage concept*, a particular instance of which is an organism-specific database. Although these databases may contain different ontologies of experiment annotations, they share the same structure and therefore can be accessed by the very same statistical algorithms. Experiment ontologies have not yet reached their final shape, and standards are reduced to minimal conventions that do not yet warrant extensive description. An ontology-independent structure enables updates of annotation hierarchies during normal database operation without altering the structure.

Availability and supplementary information: http://www.dkfz.de/tbi/services/mchips

Contact: k.fellenberg@dkfz.de

INTRODUCTION

Microarray analysis provides insight into the transcriptional state of the cell (transcriptome), measuring RNA levels for thousands of genes simultaneously (DeRisi *et al.*, 1996; Khan *et al.*, 1999; Brown and Botstein, 1999; Lockhart and Winzeler, 2000). This is done by hybridizing a labelled RNA sample to an array of either 'spotted' cDNA fragments or of oligonucleotides synthesized 'on chip' (Lennon and Lehrach, 1991; Schena *et al.*, 1995; Schena, 1996; Shalon *et al.*, 1996; Lockhart *et al.*, 1996). Ongoing sequencing projects promise to yield complete gene sets of most model organisms in the near future which can then be mounted on DNA chips. However, the data produced need to be stored in a proper way to allow for global comparison (Basset *et al.*, 1999). This applies not only to the signal intensities for each item in an array but also to all available descriptions of the sample the RNA has been derived from, and all details of its treatment.

Several database projects are currently addressing these questions. While ExpressDB (Harvard, Aach *et al.*, 2000) aims at storing data from nearly all available platforms, i.e. cDNA and oligonucleotide chips as well as SAGE, 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 (http://www.ncgr.org/research/genex/; NCGR), GEO (http://www.ncbi.nlm.nih.gov/geo/; NCBI), ArrayDB (NHGRI; Ermolaeva *et al.*, 1998), ArrayExpress (EBI; Brazma *et al.*, 2000), and RAD (http://www.cbil.upenn.edu/RAD2; UPenn; Stoeckert *et al.*, 2001), the last one combining both objectives.

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 multivariate statistical methods. Frequencies of annotation values, e.g. within a set of experiments clustered by their expression patterns, ought to be computable. Misspellings, different textual representations of semantically identical items, and, *vice versa*, ambiguous words the meaning of which depends on the context, interfere with counting such values. For this reason we have developed a system that both ensures consistency of annotations and circumvents the difficulties of free-text parsing. With the exception

*To whom correspondence should be addressed.
[3]Present address: Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, Germany.

of numerical values, our annotation system is entirely categorical, allowing to choose only among predefined enumeration-type values which can be readily analyzed in an automated fashion.

In order to keep our annotation concept flexible enough to include easily new attributes as well as new values, without the need to alter the analyzing algorithms, we store the definitions for the annotations and their allowed values as separate tables in the database linked to the data tables, thus avoiding a fixed, 'hard-wired' structure that would be difficult to extend.

Here we present a storage and analysis concept called Multi-Conditional Hybridization Intensity Processing System (M-CHIPS). It has been implemented as a set of organism-specific databases, namely for *Saccharomyces cerevisiae, Arabidopsis thaliana, Trypanosoma brucei, Neurospora 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 kinds of intensity data obtained from cDNA microarrays. It has been tailored for the need of the collaborating groups which use cDNA microarrays with either single-channel radioactive or multichannel fluorescence readout.

In a classical data warehouse, data are held in one or several databases. A warehouse then collects data from their storage databases and makes them fit into a unified data model (Ballard *et al.*, 1998; Schönbach *et al.*, 2000). Typically, a warehouse will collect only a few, 'important' attributes from each dataset. Such operations like transformations and extractions are recorded as meta data. It 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, we consider M-CHIPS 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 so-called '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 information loss.

DESIGN REQUIREMENTS

The contained data can be divided into raw signal intensities, gene annotations and experiment annotations. The last display the most complex structure among these.

Signal intensities

Since processing algorithms change rapidly, raw intensity data rather than processed values should be stored. Currently, image analysis itself cannot be carried out without human interaction. 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 passing criteria like intensity thresholds or to perform simple calculations on the database level. Such calculations may be necessary in the 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 because the binding sites on the array are taken 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.

Gene annotations

Gene annotations may consist of clone numbers, accession numbers and heterogeneous information like 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 to divide the gene set into partitions should be provided. This information turns out to be necessary to normalize separately certain sets of spots, e.g. when they have been hybridized separately.

Experiment annotations

Experiment annotations may comprise, e.g. 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 contained, too. They fall

into two classes: first, there are common annotations that are useful for all fields of interest. These are technical annotations like 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 such 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 be achieved easily 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, we permit only two types of experiment annotations, either numbers of predefined unit or values from predefined lists. If we, e.g. let an annotation 'growth phase' be an enumeration-type variable comprising the defined values 'exponential,' 'stationary' and 'pseudo-hyphal,' the occurrence of the value 'exponential' 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 under-represented in the cluster.

While in free text descriptions the number of occurences of a value are 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. Absence of highly flexible free text annotations has to be compensated for by increased flexibility in database storage.

DATABASE IMPLEMENTATION

Here we sketch how these concepts have been implemented in our databases. A detailed description can be found in a technical report on the associated web page (http://www.dkfz.de/tbi/services/mchips). The data categories mentioned above, namely raw signal intensities, gene annotations, and experiment annotations, were taken as a basis for implementation. Figure 1 shows the corresponding tables in yellow, blue and red, respectively.

Different array types and gene annotations

A database may comprise several microarray 'families' that each consist of the entirety of Multiconditional Experiments (MCEs) 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 and include a key recording the above mentioned partitions. 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 (see Figure 1).

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 one-to-many relations in Figure 1). A measurement may comprise a hybridization in case of monochannel experiments, or a single channel of a multichannel hybridization. Experiment schemes (see Figure 1, green tables) record for each measurement, to which hybridization and experimental condition it belongs, and in which 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 gene-representing spots as opposed to empty spots and various kinds of controls. For this reason we use different tables to store these kinds of intensities. In Figure 1, these tables are shown in yellow.

Fast querying of tuples is mediated by so-called indices, which immediately guide the search to the specified tuples. If all measurements were stored in one big table per category, adding a new measurement would be slow because of the time necessary for recomputing the indices. Therefore, new measurements are inserted as separate tables, computing indices only for the new tuples.

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

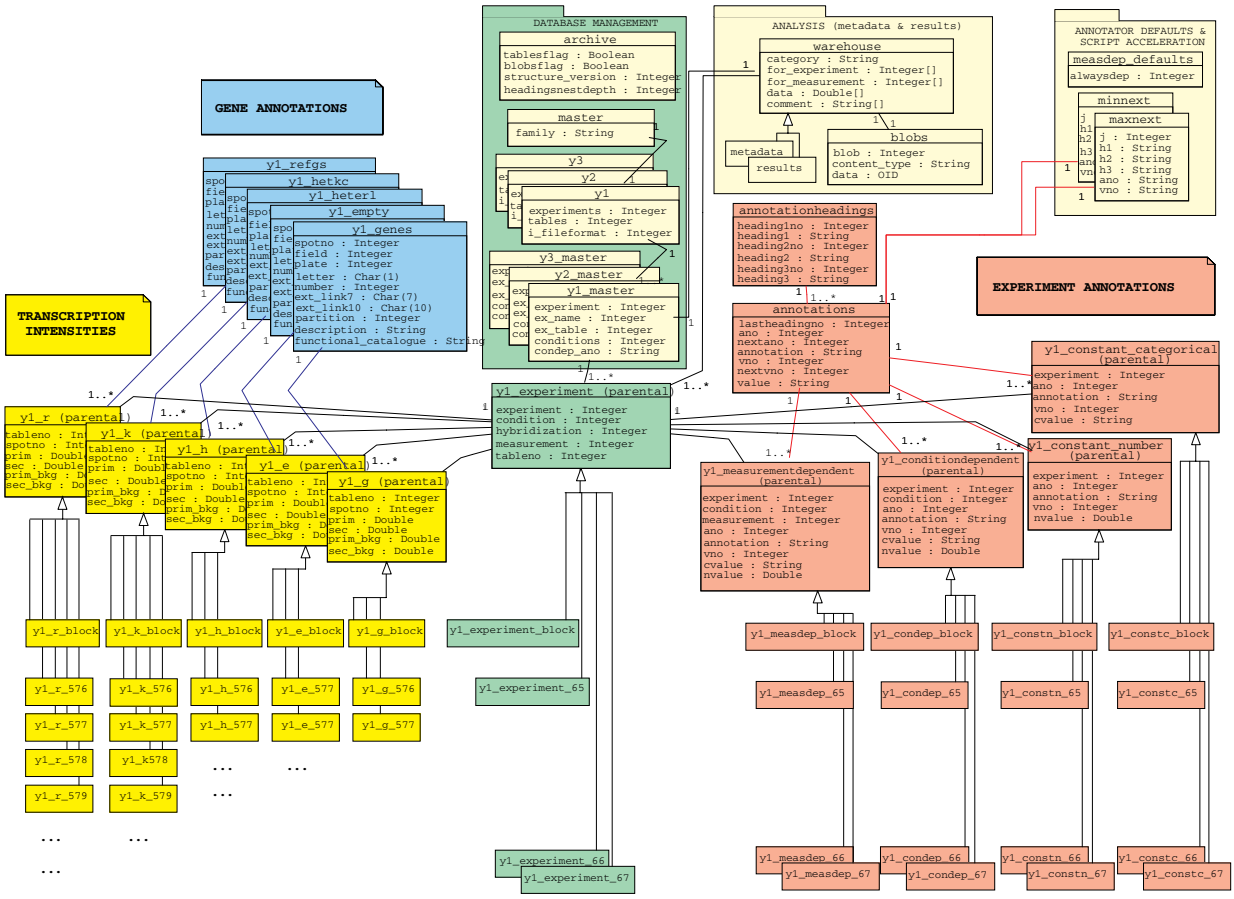

Fig. 1. UML scheme of an M-CHIPS database. The tables are arranged and color coded according to the categories mentioned in the text. Overlapping tables show identical structure. Arrows indicate table inheritance. All child tables in our databases have the same structure as their parents. The database management tables (in green box) show the array families y1, y2 and y3 (e.g. 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. Other families comprise identically structured tables. The number of tables within an M-CHIPS database thus varies with the number of comprised array families. It also depends on how may uploaded measurements already have been assembled into a block table. M-CHIPS operates on a unified storage concept for standardized algorithmical access rather than on a fix database structure.

'block' tables that are filled by overnight jobs collecting measurements that are not to be altered or deleted. Thus, computation of large indices is performed at times of low traffic as an investment in query performance. We use table inheritance as an elegant aid in keeping track of both single and block tables. Since every access to the intensity tables is directed via one of the parental tables, query syntax does not change with merging a set of tables into one block. This block will be a child of a specific parental table as are the tables to collect (see Figure 1, small yellow tables).

On a SUN E450 server under Solaris 2.7, a PostgreSQL 6.5.3 server process retrieves 2 consecutively uploaded hybridizations (comprising 6103 yeast genes in double spotting) out of 686 ones stored in separate tables on average in 85 s. The same query performs in 2.3 s, if the 686 hybridizations are assembled into one big table. Even to retrieve two out of 2251 hybridizations takes only 2.8 s when all hybridizations are *en bloc*.

Experiment annotations

To achieve direct access for statistical methods to all experiment descriptions, they 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 in the database structure itself. The web-based annotation pro-

cess (described below) 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 which contains particular definitions of experiment annotations appropriate for the attended samples. We provide annotation definitions for *S. cerevisiae*, *A. thaliana*, human tumor biopsies, *T. brucei* and *N. crassa* on our web page (http://www.dkfz.de/tbi/services/mchips). 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, we try to keep this share as large as possible. New 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 table. Hierarchical structure of annotation ontology is recorded by the content of a second table. Table 1 gives 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 be queried exclusively by joining them, we 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 like those in Figure 2. The results of the annotation process are stored in annotation tables.

Annotation tables. A MCE 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 repeatedly performed 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 of these measurements belong to which experimental condition also record which of them were performed simultaneously onto 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. We decided to store the three sets in separate annotation tables for convenient algorithmical handling (see Figure 1, 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.

ALGORITHMS

The database was designed to be charged and queried by the experimenters themselves using algorithms which mediate upload and annotation of experiments, as well as data analysis. M-CHIPS consists of C, Perl and MATLAB functions. We intend to make the complete M-CHIPS source code available, as well as SQL statements creating a sample database.

Experiment annotation

Experiments can be annotated from remote by the experimenters themselves using a web 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, we 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 the differing ones. Few parameters should be varied per condition, so the majority of the annotations is constant throughout the experiment. Among these, the majority is 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 Figure 2. It is possible to enter detailed descriptions (111 annotations) of large MCEs (24 measurements) in less than 15 min.

Preprocessing

In M-CHIPS, preprocessing starts with normalization of raw signal intensities. The normalization is based on robust affine-linear regression of one measurement versus a control measurement (see below). The performance may be judged from the scatterplot of the raw data (measurement versus control measurement). In this plot, a regression line represents the multiplicative distortion and additive offset determined by the fitting algorithm. The performance of the fit is visible in how well the regression line matches the central dense parts of the cloud. Furthermore it can be observed, which properties of the raw data led to an eventual suboptimal result. The scale

Table 1. Example for experiment annotation definitions

```
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
(...)
```

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 Figure 1). 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 (named 'annotations,' see Figure 1) 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 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.

of the plot can be switched between linear and double-logarithmic. In log scale the regression line appears as a curve the curvature of which depends on the additive offset between the two measurements. We use two algorithms as described in Beißbarth *et al.* (2000) and Fellenberg *et al.* (2001). For both, the set of trusted spots of unvaried expression taken into account for fitting can be specified (housekeeping genes, external controls, or entire set).

M-CHIPS discriminates between mono- and multi-channel experiments. For the former, each measurement is normalized versus the genewise median of the hybridizations for the control condition, resulting in absolute intensities. For the latter, the channel belonging to the control condition serves to normalize the other channel(s) of the same hybridization, resulting in intensity ratios. For many arrays and experiments, the majority of genes spotted on the array is not expressed to a measurable amount. While displaying notable ratios due to measurement fluctuations, they can be eliminated by means of an intensity filter. To compute intensity levels from

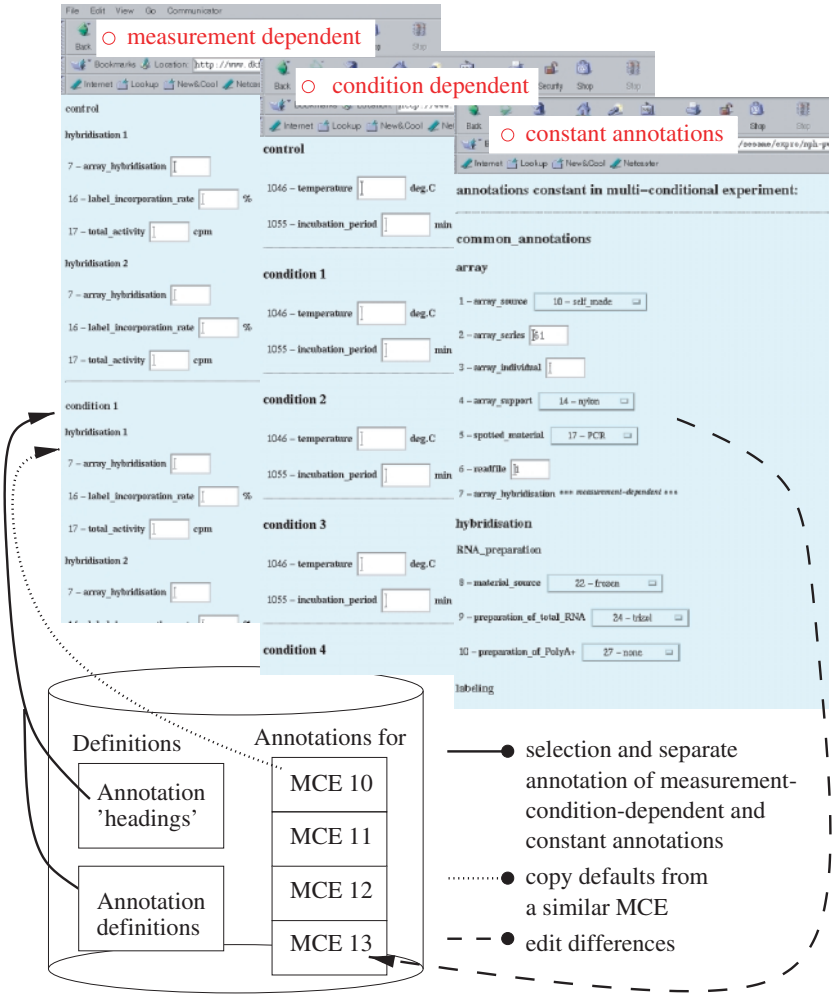

Fig. 2. Experiment annotation process. The annotation process may start with copying default values from the most similar MCE. Secondly, from the complete list of defined annotations the measurement-dependent ones are selected and then annotated for each single measurement. Afterwards, from the remaining annotations, those being condition-dependent for the particular experiment are choosen and annotated for each experimental condition. For the constant annotations, it suffices to edit few differential ones, if the questionaire is prefilled with default values copied from a similar experiment. Compare the HTML form for the constant annotations with Table 1.

multichannel ratios, these ratios are multiplied with an average control measurement, being the genewise median of the absolute values of the control channels.

Apart from intensity and ratio filters, reproducibility measures (Beißbarth *et al.*, 2000) are applied to extract genes that are reproducibly up- or down-regulated. These measures integrate repeatedly performed measurements for the same experimental condition. In addition, they are plotted versus the average intensity level and ratio as a measure for quality control.

Statistical analysis

Analysis techniques include hierarchical clustering (Eisen *et al.*, 1998), correspondence analysis (Fellenberg *et al.*, 2001), 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 correspondence analysis 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

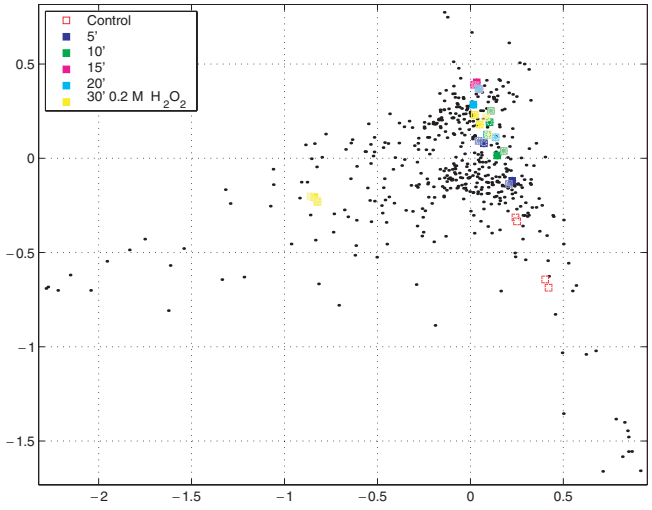

Fig. 3. Oxidative stress. The correspondence analysis plot shows a dataset recorded from wild type yeast cells responding to 0.2 M hydrogen peroxide in the medium. Genes are depicted as black dots, measurements (in this case monochannel hybridizations) are shown as squares, color-coded according to the experimental condition (here time point) they belong to. Further explanation is given in the text and in the supplemental material. The outlying cluster of yellow labeled measurements is further characterized by experiment annotation values over- or underrepresented in this cluster as shown in Table 2.

plot. In the same manner clusters of measurements can be selected and plotted. Moreover, they can be automatically scanned for significant experiment annotation values. For each value of every annotation, instances of occurrence are counted. For a particular value, its frequency in the cluster is determined as the number of its occurrences 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. An example is shown in Figure 3.

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 correspondence analysis. Experimental and computational details are given in the supplemental material on our web site (http://www.dkfz.de/tbi/services/mchips).

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 (yellow), 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 eight out of 111 annotations (see Table 2).

The first two annotations listed provide the information that the entire cluster was hybridized on array individual six which is the only one stemming from array series (i.e. production batch) 59, whereas all other arrays were of series 61. From other experiments, we generally observed sufficient comparability among arrays of the same production series, whereas arrays of different batches could not be directly compared. To show the distorting effect of this artifact to an otherwise revealing dataset, we provide a correspondence analysis plot calculated from this dataset without the outlying measurements in the supplemental material. The other six extracted candidates proved to be unable to characterize the selected cluster as described also in the supplemental material.

Sometimes, especially with higher numbers of measurements, it is desirable to aggregate values for annotations of continuous range (see no. 16 and 17 in Table 2). 'Label incorporation rate' may thus be discretized into e.g. low, medium and high values. We provide methods enabling discretization of annotation ranges into a chosen number of bins due to their particular distribution or by expert knowledge.

DISCUSSION

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; http://www.ebi.ac.uk/microarray/MGED/Annotations-wg/index.html), 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 e.g. those proposed by EBI (MIAME) is independent from our storage schema. The experimenter defining the annotations decides about standard compliance and level of detail. The sets of hierarchically ordered annotations are listed on the associated web page (http://www.dkfz.de/tbi/services/mchips). The entire descriptions of all hybridizations stored in our databases can be analyzed statistically. We currently keep 1659 hybridizations in 12 databases. They belong to the above five fields of research and comprise both radioactive-label and multichannel experiments.

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

Table 2. Frequencies of characteristic annotation values

More than or exactly 2x over/underrepresented:

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

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

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

annotation 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)
```

completeness exceeding minimal standards. For analysis, it provides the capability to include 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.

Previously published microarray database concepts have focused on the ability to include intensity data from different platforms and to make these comparable (Aach *et al.*, 2000; Brazma *et al.*, 2000). Some projects have started to develop controlled vocabulary for experiment description (e.g. ArrayExpress, RAD and GEO). However, little effort has so far been made to categorize the descriptions down to minute detail and make them amenable to analysis. As our concept is not meant to be implemented in a large public gene expression database, we have not dealt with including additional platforms like oligonucleotide chips or SAGE, but have concentrated on mining the wealth of information contained in the experiment annotations. However, we 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 yeast, arabidopsis, *T. brucei*, *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.

Experiment annotation is web-based to ensure that any experiment can be annotated from remote by the experimenters themselves. Efforts for annotating experiments are minimized. Data analysis comprises 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 (Beißbarth *et al.*, 2000; Fellenberg *et al.*, 2001). High-level analysis techniques include hierarchical clustering (Eisen *et al.*, 1998) and correspondence analysis (Fellenberg *et al.*, 2001). 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 like 'cell cycle' in their gene annotation or to import multiple sets of gene tags from text lists.

Statistical analysis of experiment annotations can be applied for arbitrary sets of hybridizations 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 experimental 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 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 be directly applied, too, because, unlike with free text annotation, instances of occurrence are readily countable for all annotation values. We consider correspondence analysis particularly useful for the exploratory analysis of microarray data. Future plans comprise integrated visualization of both transcription intensities and experiment annotations by multiple or joint correspondence analysis, 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.

ACKNOWLEDGEMENTS

We are grateful to Sonja Bastuk and Melanie Bier for excellent technical help. Tim Beißbarth and Dieter Finkenzeller contributed with analysis algorithms. We would also like to thank Judith Boer, Marcel Scheideler, Frank Diehl, Susanne Grahlmann, Verena Aign, Helene Tournu, Arno Meijer, Luis Lombardia, Manuel Beccera, Andy Hayes, Nikolaus Schlaich, Roland Eils, Ugis Sarkans and Alvis Brazma for testing, reporting bugs, defining experiment annotations, for criticism or suggestions. This work was funded by the European Commission as part of the Eurofan-2 project under contract BIO4-CT97-2294.

REFERENCES

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

Ballard,C., Herreman,D., Schau,D., Bell,R., Kim,E. and Valencic,A. (1998) *Data Modeling Techniques for Data Warehousing*. IBM International Technical Support Organization, San Jose, CA, www.redbooks.ibm.com.

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

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

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

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

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

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

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

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

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

Lennon,G.G. and Lehrach,H. (1991) Hybridization analyses of arrayed cDNA libraries. *Trends Genet.*, **7**, 314–317.

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

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Schena,M. (1996) Genome analysis with gene expression microarrays. *BioEssays*, **18**, 427–431.

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

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

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

Stoeckert,C., Pizarro,A., Manduchi,E., Gibson,M., Brunk,B., Crabtree,J., Schug,J., Shen-Orr,S. and Overton,G.C. (2001) A relational schema for both array-based and sage gene expression experiments. *Bioinformatics*, **17**, 300–308.