BIOINFORMATICS

Vol. 16 no. 11 2000
Pages 1014–1022

Processing and quality control of DNA array hybridization data

T. Beißbarth[1,2], K. Fellenberg[1], B. Brors[1,3], R. Arribas-Prat[2], J. M. Boer[4], N. C. Hauser[3], M. Scheideler[3], J. D. Hoheisel[3], G. Schütz[2], A. Poustka[4] and M. Vingron[1,*]

[1]Abt. Theoretische Bioinformatik, [2]Abt. Molekularbiologie der Zelle I, [3]Abt. Funktionelle Genomanalyse and [4]Abt. Molekulare Genomanalyse, Deutsches Krebsforschungszentrum, INF 280, D–69 120 Heidelberg, Germany

Received on October 11, 1999; revised and accepted on June 1, 2000

Abstract

Motivation: *The technology of hybridization to DNA arrays is used to obtain the expression levels of many different genes simultaneously. It enables searching for genes that are expressed specifically under certain conditions. However, the technology produces large amounts of data demanding computational methods for their analysis. It is necessary to find ways to compare data from different experiments and to consider the quality and reproducibility of the data.*

Results: *Data analyzed in this paper have been generated by hybridization of radioactively labeled targets to DNA arrays spotted on nylon membranes. We introduce methods to compare the intensity values of several hybridization experiments. This is essential to find differentially expressed genes or to do pattern analysis. We also discuss possibilities for quality control of the acquired data.*

Availability: *http://www.dkfz.de/tbi*

Contact: *M.Vingron@dkfz-heidelberg.de*

Introduction

This special issue of *Bioinformatics* is one of the many signs of the increasing importance of DNA arrays and chips for the study of gene expression (Piétu *et al.*, 1996; DeRisi *et al.*, 1996; Spellman *et al.*, 1998; Khan *et al.*, 1998; Roth *et al.*, 1998). While the experimental technology has been developed very rapidly, it appears that the computational processing of the resulting data is lagging behind. In this paper, we report on our experience with the processing of data generated with arrays on nylon membrane using radioactive hybridization (Lennon and Lehrach, 1991).

The techniques for simultaneous determination of expression levels of a large number of genes can be roughly divided into two categories. In the first approach, a sample of the molecules in a library is characterized by determining certain tags. For example, in EST (expressed sequence tag) sequencing the sequence read constitutes such a tag (Adams *et al.*, 1991). In SAGE (serial analysis of gene expression) very short tags are concatenated and then sequenced (Velculescu *et al.*, 1995). After establishing which tags correspond to the same gene, such techniques give an estimate of the proportion at which a gene occurred.

The other approach works by immobilizing those genes whose expression level will be investigated and then hybridizing the sample under study to the immobilized genes. The detection of the hybridization signal may rely on radioactive (Friemert *et al.*, 1989) or fluorescent labeling (Shalon *et al.*, 1996). In the case of radioactive labeling, the amount of radioactivity as detected after exposure is the indicator of the amount of RNA present. Fluorescent labeling allows the comparison of different samples by labeling them differently. The resulting competitive hybridization makes the ratio between the two signals an indicator of the ratio at which particular genes are expressed in the two samples (DeRisi *et al.*, 1996; Chen *et al.*, 1997).

Nylon filters that are used for hybridization with radioactively labeled samples are at the moment easier to produce and do not require specialized hardware for the read-out of the signal. At our center, several groups have been working with this set-up, and this paper describes the lessons we learned from analyzing data produced with those filters. Arrays for yeast have already been described by Hauser *et al.* (1998). With similar technology, arrays containing clones from *Arabidopsis* were produced. In the context of the German Human Genome Project, clones selected from Unigene clusters have been spotted to obtain a comprehensive filter of human genes. Additionally, commercially available filters for mouse have been used (see **Systems and methods**).

*To whom correspondence should be addressed.

Commercial software is available for detecting spots and quantifying their intensity. Typically, such software will generate a table of intensities assigned to the individual spots. Many problems may arise at that point. Most prominently, spots may be missed due to incorrect grid assignment. Such software, however, is not the focus of our attention here and we assume that these problems have been solved. Nevertheless, a range of other problems will follow. Questions arise like: which genes are actually 'turned on', i.e. which genes are indeed expressed although perhaps at a low level? How reproducible are experiments? How do possible variations in the efficacy of experimental procedures influence the outcome?

Comparison of the results from different hybridizations requires standardization. Because of different background intensities, different labeling efficiencies or differing exposure times, two (or more) hybridization experiments are not readily comparable without prior standardization. Here, we will provide methods to deal with questions of additive and multiplicative distortion automatically. This is based on a physical model and has been successfully applied to several hundred hybridizations. We will show that a subpopulation of hybridization intensities across an array can be modeled by lognormal distribution. This distribution can be used to determine a threshold of reliability for these intensities. Because of the poor reproducibility of measured values one has to apply filtering in order to exclude highly variant spots in an array from further analysis. We will provide methods for quantifying the quality of spots.

This paper does not deal with the detection of correlations among genes in large numbers of experiments. Rather, we view the methods discussed here as a prerequisite for subsequent analysis of the data.

Systems and methods

The data analyzed in this paper were generated using complementary DNA arrays produced by PCR from 6116 ORFs (open reading frames) of *Saccharomyces cerevisiae* (Hauser *et al.*, 1998), representative *Homo sapiens* ESTs from the Unigene Collection Build 17, August 1997, representative *Arabidopsis thaliana* ESTs, and commercially available arrays from *Mus musculus* (Gene Discovery Array, Genome Systems, Inc., MO, USA). The cDNA samples were spotted onto nylon membranes. The radioactively labeled cDNA representation of the mRNA pool of a biological sample was then hybridized to the array.

The amount of radioactivity on the membrane was measured by means of a phosphorimager and converted into corresponding gray levels of an image. The gray levels are supposed to be linearly correlated to the amount of radioactivity on the filter. Every spot of the array needs to be recognized and assigned to its position in the array, i.e. to the corresponding clone number. For each of the spots in the array an intensity value needs to be assigned. Due to the large number of spots only automatic or semiautomatic procedures are suitable for this task. The data presented here have been obtained using the commercially available image analysis software Array Vision (Imaging Research, Ontario, Canada). Data were obtained as a list of intensity values and array positions for all the spots in the array.

Statistical analysis routines have been realized in MATLAB 5.3 (MathWorks Inc., MA, USA) and are available through a web-based interface (http://www.dkfz.de/tbi).

Definitions. We refer to the spots on the membrane as 'probe' and to the sample hybridized to the membrane-bound array as 'target', according to *Nature Genet.* **21** (Suppl.), 1999, p. 1. The target cDNA is derived from the total RNA or from poly-A^+ RNA prepared from a biological sample, referred to as 'mRNA pool'. Probe cDNA species are spotted in duplicate on the membrane. These spots are denoted 'primary' and 'secondary' spot, solely to indicate that there are two spots for each cDNA species. These cDNA species are either amplified fragments of an ORF (for *S.cerevisiae*) or PCR products derived from ESTs.

By the phrase 'hybridization' we refer only to the technical procedure of hybridizing and not to the complete experiment. Thus, 'repeated hybridization' will mean that the hybridization protocol has been repeated. There is no sharp definition of the term 'experiment'. This could either mean the process of transcribing and labeling a target mRNA sample followed by hybridization, or it could include additionally culturing of organisms or cells and subsequent RNA preparation. The term 'genome-wide array' refers to arrays that contain all available gene representatives of an organism. Except for yeast, which has been completely sequenced, this comprises only a portion of all genes of a genome.

Results

Data analysis

Methods to analyze data from a single filter. Typically, the first problem one encounters when dealing with data from radioactive hybridizations is the presence of a background signal. Phosphorimager screens are particularly sensitive and will make non-zero background intensity visible. The background may be inhomogeneously distributed, in which case we refer to it as 'local background'. In order to obtain a value for background intensity the following controls may be present on the filter and can be used for correction. At some positions in the array there could be array positions without DNA. These *empty spots* can be used to obtain a value for

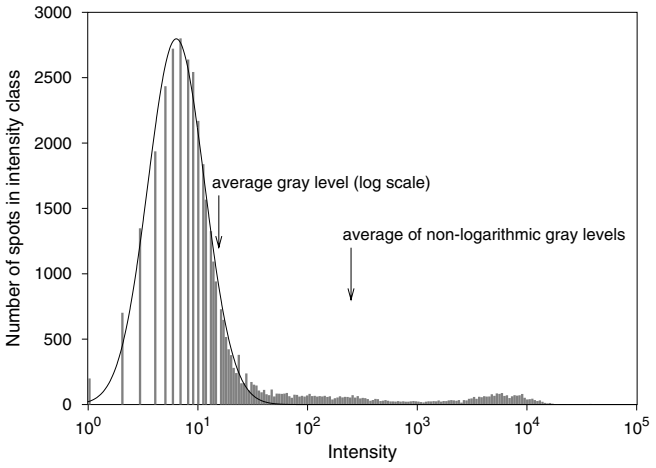

Fig. 1. Histogram of gray level distribution across an array of 17 280 EST clones (Human Unigene Collection) hybridized to a complex target. A fitted normal probability distribution is shown overlaid.

background intensity (and hence be subtracted from all other intensities). Intensities of spots representing supposedly non-expressed genes in the target mRNA pool provide another estimate of the signal intensity due to cross hybridization. Below we describe how to model these values. *Heterologous* DNA may be spotted on the filter to probe for unspecific hybridization. The value obtained for these spots can be used to define a threshold of reliability to mark which values we want to trust in the later analysis.

Generally, in a genome-wide array, for most spots there will be no complementary mRNA species in the complex target simply because only a comparatively small number of genes is expressed in the biological sample under investigation. Yet, such spots may display a marked signal intensity in the hybridization. Looking at the histogram of gray level distributions for all spots on an array (Figure 1), we frequently observe two populations of spots. On a logarithmic scale, the gray levels in the low-intensity region follow a normal distribution comprising for most arrays more than 80% of all genes. This means that the majority of signals roughly follows a log-normal distribution typical for data that are centered at a minimum near zero and cannot extend below zero (Sachs, 1984). The logarithms of the remaining intensities extend to the right of the Gaussian curve; this gives the histogram a much heavier tail than expected from random data alone. We interpret the population that displays a normal distribution of intensities as the set of those spots that do not have a complement in the target or where the number of transcripts is below the detection limit. What we observe would then be the distribution of the signals for this population, mainly due to unspecific interaction with target DNA fragments, overlaid on the distribution of expression levels of all genes. We might not be able to distinguish lowly expressed genes from genes which are not expressed. From SAGE analysis it is expected that—for most tissues—only 10–20% of all genes are expressed under a given condition (Zhang *et al.*, 1997; Velculescu *et al.*, 1999). Therefore, it is not suprising to see that the signals of most genes fall into the background region. It should be noticed that genome-wide arrays based on ESTs might still have a bias towards highly expressed genes.

To characterize the dispersion of intensities, a normal distribution function can be fitted to the values of the low-intensity class. Since this is only a subset of all values, a method is needed for estimating which values belong to this class. We use an algorithm which starts by fitting a Gauss normal distribution function to the histogram by means of non-linear regression. To distinguish between 'noise' that needs to be modeled and signal that does not obey the normal distribution, the data set used for fitting is then iteratively reduced by truncating above a threshold. This threshold is calculated by adding one standard deviation to the mean of the calculated distribution. This is followed by a new fitting based on the reduced data set. The iteration stops when the mean of the distribution stays constant. We wish to emphasize that the mean of this distribution is not identical with the average of the logarithms of the entire data set, and it is clearly separated from the mean of the non-logarithmic data transformed to the same scale. The latter value is so far away from the center of the distribution of intensities that we do not ascribe meaning to it (see Figure 1).

When fitting logarithmic intensities by a normal distribution is successful this also provides a rational approach to the question of which genes are actually expressed. Obviously, even for higher intensities there is still a positive probability that such a signal might be due to chance. This probability is related to the area under the normal distribution above a certain threshold. In Figure 1, e.g. values above 100 are almost certainly 'real' signals.

In our experience this approach usually works well for most experiments done on genome-wide arrays. However, there are experimental setups where it does not hold true that most genes are not expressed, especially when using arrays made from a small selection of genes. In this case this method cannot be applied.

Comparison of two data sets. When we compare two data sets which originate from different hybridization experiments, we notice certain systematic differences in the measured intensities. We model two different types of systematic differences: one type is the background as has been described in the previous section. We assume that the influence of the background is additive with respect to the measured intensities. Further, we observe a constant

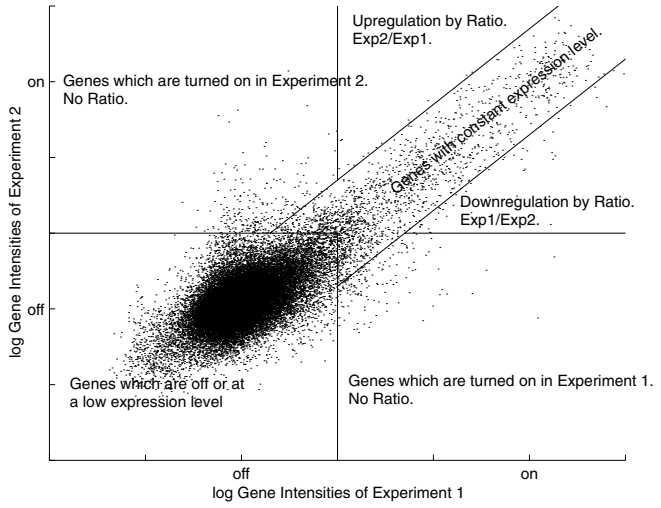

Fig. 2. Scheme of a scatter plot using logarithmic scale. As an example two hybridization experiments on an array containing 18 432 mouse ESTs are compared. Intensities are plotted from hybridization to target derived from mouse thymus with (ordinate) or without (abscissa) stimulation by dexamethason. A schematic representation of the different regions is given, as well as their interpretations.

multiplicative factor between the intensities of genes of two hybridizations, probably due to different labeling rates of the complex probe used for hybridization or to unequal exposure times of the filters.

A good way to visualize the comparison of two hybridization experiments is a scatter plot, schematically shown in Figure 2. The intensity of every spot in Experiment 1 is plotted against its intensity in Experiment 2. It is appropriate to use logarithmic scale because one is interested in intensity ratios rather than absolute differences. The plot can be subdivided into regions with different interpretations: data points in the lower left corner represent genes which are inactive or are expressed only at a low level. In our experiments we noticed that these constitute the vast majority. In the upper left and lower right corner, points correspond to genes which are only expressed in one experiment and not (or not discernibly) in the other. The intensity ratios calculated for these regions are poorly reproducible. Genes in the channel around the diagonal are detected as expressed in both experiments. The farther a gene is away from the diagonal, the higher its intensity ratio between the measurements in the two experiments.

Frequently, transforming the data sets of either experiment to a standardnormal form is used as a method for standardization (Piétu *et al.*, 1996). This implies that the intensities follow a normal distribution as was discussed in the previous section. We have shown there that this holds true only for a subpopulation of spots and thus should not be used for standardization. The spots fall in either one of two classes, the first of which contains spots that display low hybridization signal intensities due to unspecific interaction with labeled target DNA fragments. The second class comprises all spots whose signal intensity is due to specific hybridization to a complementary target sequence. We consider it not appropriate to treat both classes as one.

Our preferred way to standardize the data for comparison is based on the linear model we sketched above. First we have to eliminate the effects of background by subtracting an additive constant, or offset. This offset has some influence on the adjusting factor that is calculated. The influence of background on data sets is outlined in Figure 3a,b. The scatter plot of two experiments with highly different background is distorted (Figure 3a) to yield an arc-shaped cloud of points around the identity line when plotted on a logarithmic scale. This distortion is corrected by subtraction of an offset (Figure 3b). If the image analysis software does reliably measure background intensity, we use these values for correction. Otherwise, we robustly and rapidly estimate the offset by taking the 5% quantile of intensity values in either data set and subtracting it from all corresponding intensity values.

Intensity values close to background raise additional problems since they usually display an unfavorable signal-to-noise ratio, leading to highly unreliable intensity values. Ratios formed with such values can get very high even when there is no significant difference in the expression level of the corresponding genes. It may be useful to define an intensity threshold in order to exclude these spots from being marked as 'differential' (see Figure 2).

We use the following procedure to determine a rough estimate of this threshold. If only high-intensity values are included in the computation, the linear correlation coefficient of intensities in Experiment 1 relative to intensities in Experiment 2 will increase when more and more intensity values are added to the analysis. Including lower intensity values, the point at which the linear correlation coefficient starts to decrease is chosen as the threshold.

Next, we try to find the systematic factor of change. Therefore, we need a set of genes which we believe should have an equal or similar expression level in the experiments we want to compare. For these genes, a median of the ratios is used as the adjusting factor, such that the ratio of intensities for these genes becomes 1. The effect of this calculation is demonstrated in Figure 3c,d.

If a set of housekeeping genes can be defined, these can be used to adjust the intensity values. These genes are believed to be expressed constitutively at a constant level, independent of the conditions of the experiment. However, aside from the difficulty of finding such genes, they may not behave uniformly under all conditions and sometimes display unexpected behavior. Another method relies on

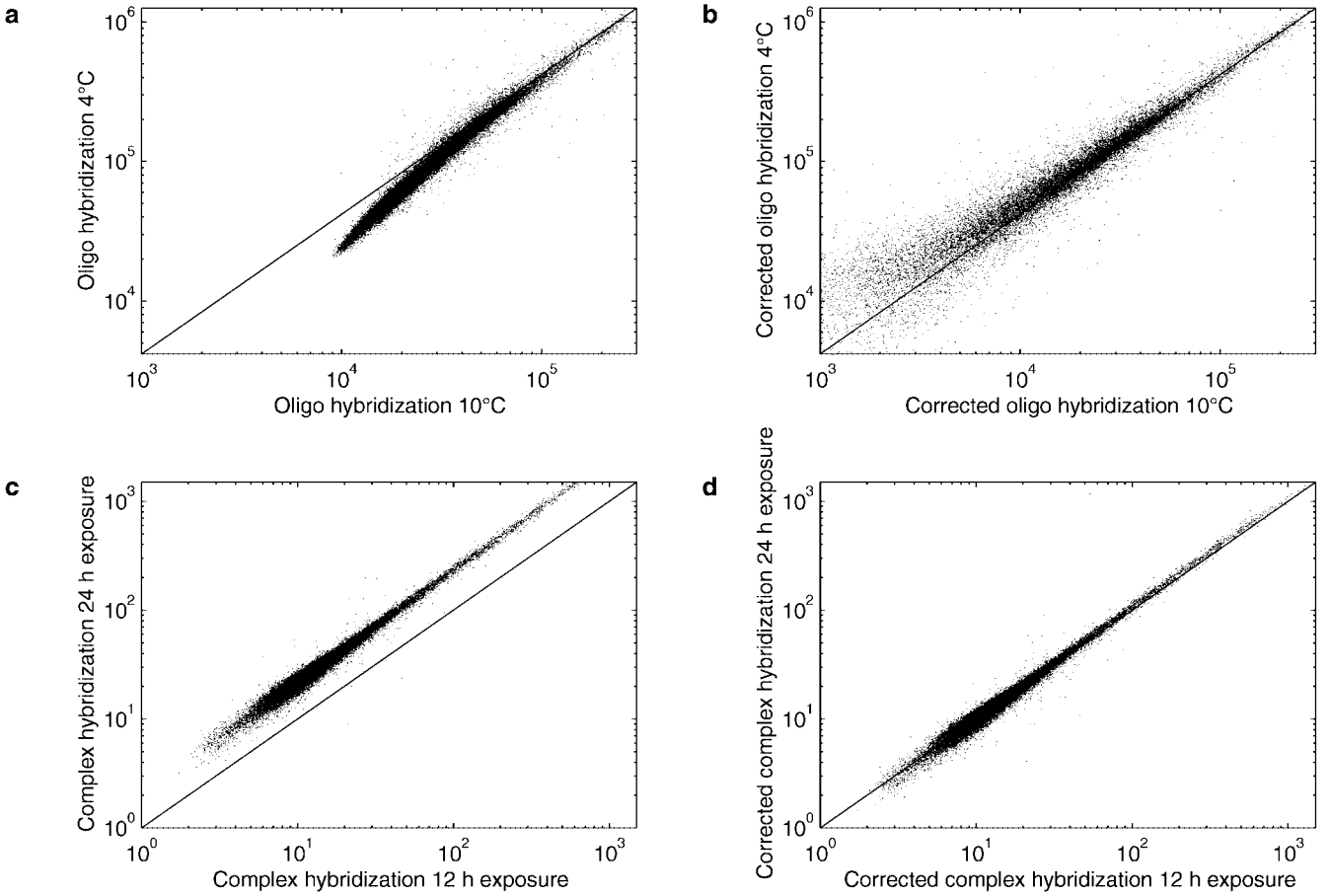

Fig. 3. Influence of correction for background and multiplicative factor. In (a) and (b) two hybridizations with an oligonucleotide recognizing all spots are compared, carried out on an array of 13 824 *Arabidopsis* ESTs at differing temperatures. The hybridizations show highly different background intensity. The raw data (a) display an arc shape which has been corrected by subtracting the 5% quantile (b). (c) and (d) display a factor difference in their intensities. The data are from one hybridization of a complex target to a mouse array, with different exposure times to the phosphorimager screen. Data before (c) and after (d) standardization are shown.

externally added controls, i.e. heterologous DNA spotted on the filter that hybridizes with a complementary sample added to the complex target.

Using genome-wide arrays, it is a good assumption that the expression level, and hence the signal intensity for most spots does not change when comparing closely related experiments. When comparing more distantly related experiments or using arrays which are biased towards a selection of genes where a lot of changes are expected this assumption does not hold true. To be able to compare these kinds of experiments one needs to be able to define a set of control genes, where no changes are expected. To estimate the factor of change between experiments we compute the arithmetic mean of the logarithmic differences. The intensities of, e.g. Experiment 2, can then be adjusted to be on the same scale as the other experiment by subtracting this mean from all intensities of Experiment 2:

$$\ln e_{2,k} - \frac{\sum_{i=1}^{n} (\ln e_{2,i} - \ln e_{1,i})}{n}$$

for each intensity $e_{2,k}$ ($k = 1, \ldots, n$). In this equation, $e_{2,\cdot}$ refers to the intensity data of Set 2, $e_{1,\cdot}$ to those of Set 1, and n is the number of spots on the filter. To make the results less sensitive to outliers, the arithmetic mean may be replaced by the median. Genes below an intensity threshold, which display a considerable variance in the intensity ratio, should not be included in the calculation of the mean or the median.

We have noted that hybridization experiments to be compared are best performed with the same filter, or with filters from the same production batch, because we

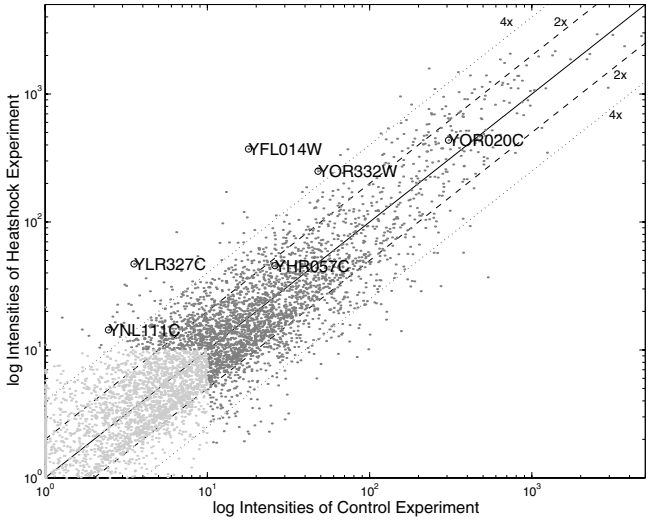

Fig. 4. Scatter plot of hybridization intensities obtained with an array containing PCR fragments from 6103 *S.cerevisiae* ORFs. Yeast has been cultured at 30 (abscissa) or 37 °C (ordinate). Several heat-induced genes have been marked. Explanation of ORF identifiers is given in Table 1.

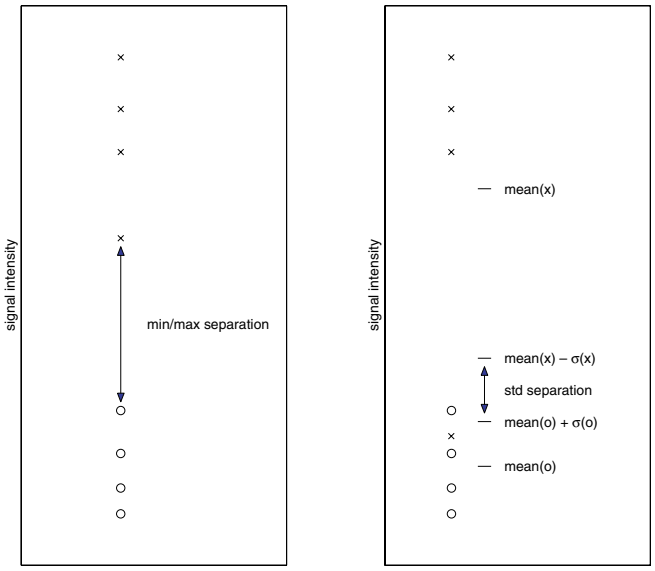

Fig. 5. Schematic illustration of the proposed quality measures. Signal intensities are logarithmized.

observed significant differences with filters from different batches, often rendering the experiments incomparable.

In Figure 4 we show the comparison of two hybridizations of *S.cerevisiae* that has been cultured at 30 and 37 °C, respectively. Standardization is essential to define the region of heat-induced genes. Some genes known to be induced by heat shock are marked. These include chaperones, cytochrome *b5* which is involved in membrane lipid remodeling, as well as a V-type ATPase subunit that, when knocked out, will render the null mutant heat sensitive. Explanation of the ORF identifiers is given in Table 1. The highest-induced gene (induced by a factor of 21) is *HSP12*, a small heat shock protein.

Quality control

A convenient way to handle highly variable data is to repeat an experiment. There are several levels of repetition one can envisage when dealing with expression array hybridization data.

The first level of repetition is to have all gene representatives spotted in duplicate on the membrane. In our experience, comparison of the intensities of the primary and secondary spot reveals reading errors due to problems with image analysis and spotting. Erroneous grid allocation or overlapping signals produced by very intense spots can be detected in a scatter plot where the intensities of the secondary spots are plotted against the intensities of the corresponding primary spots. All spots one wishes to rely on should be located in a narrow zone around the identity line. Outliers should be excluded from further analysis.

When comparing two hybridization experiments, four differences between the intensity values corresponding to one gene can be formed with the values for primary and secondary spot on either filter. It is conservative to expect that, in the case of differentially expressed genes, even the smallest of these differences should exceed the variance of the differences. We consider it less informative to calculate the mean of primary and secondary spot for each filter prior to comparison because variance information is lost.

The next level of repetition is to repeat parts of or even entire experiments. Repeated hybridization starting with the same RNA sample will reflect variations in labeling, hybridization measurement and intensity assignment. Repetitions comprising the whole protocol including culturing, sampling and RNA preparation will reflect variations in all of the performed steps giving a more complete image of the variance for each signal observed. We investigated data comprising up to four replicates, providing eight intensity values for each probe species in the array.

To check for reliability in comparison of experiments under two different conditions, each of which has been repeated several times, we calculate quality measures to see whether signals for a spot corresponding to a putative differentially expressed gene are well separated. The ‘min–max separation’ is calculated by taking the minimum of all distances between data points in the first data set and those in the second data set. Well separated data sets should not overlap and therefore display a positive min–max separation (Figure 5). This is

Table 1. Description of yeast ORFs found to be differentially expressed in heat shock experiment

ORF id	Description	Factor of induction
YFL014W	*HSP12*, heat shock protein	21.7
YLR327C	chaperone, involved in assembly of protein complexes	15.9
YNL111C	cytochrome *b*5, cofactor of fatty acid desaturases	7.0
YOR332W	*VMA4*, V-ATPase subunit, null mutant heat sensitive	5.1
YHR057C	*CYP2*, cyclophilin	1.8
YOR020C	*HSP10*, heat shock protein	1.5

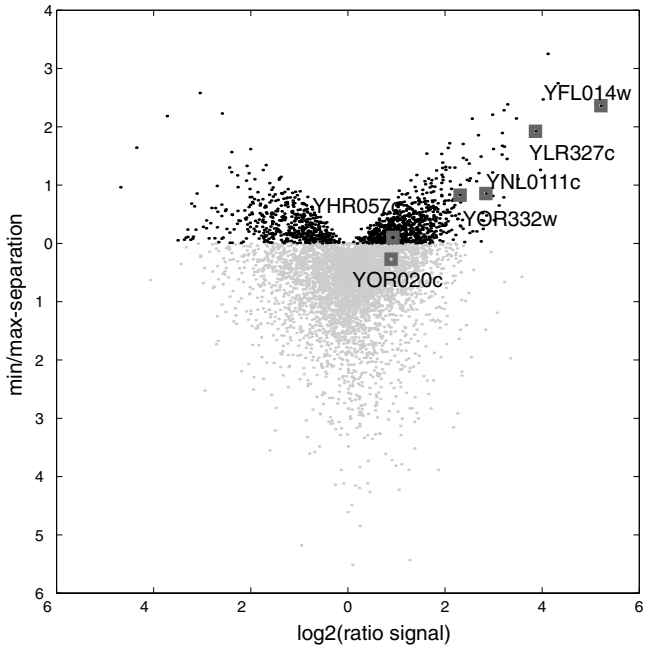

Fig. 6. Judging quality of several reproductions of hybridizations with an *S.cerevisiae* array and yeast cultured at 30 or 37 ° C (as shown in Figure 4). The $\log_2$ of the ratio of the median intensities (abscissa) is plotted vs the min–max separation (ordinate).

a very restrictive measure which is frequently disturbed by outliers. A less stringent criterion for separation would be a distance which we call 'std separation', defined as the difference of the means of the two data sets diminished by one standard deviation of either data set (Figure 5).

In the heat-induction experiment with *S.cerevisiae* described above, three replicates of the condition cultured at 30 ° C and two replicates cultured at 37 ° C have been generated. Each gene has two representative values in each replicate. To evaluate the reproducibility of the results the measured ratio of the median intensities is plotted versus the min–max separation (Figure 6). A min–max separation higher than zero indicates a reliable result. It can be noticed that the highest-induced gene *HSP12* also is among the genes with the highest min–max separation.

The most general level of repetition is to repeat entire comparisons of experiments. In this case, one can compare the lists of differentially expressed genes and see whether certain genes are repeatedly included.

In the comparison of several data sets, e.g. a time-course, a concentration series or a collection of mutants, standardization is equally required. Two questions are of particular importance: first, a standard has to be defined when a control condition has been repeatedly analyzed. In this case we use a virtual standard that is obtained by taking the gene-wise median of intensity values across all replicates of a hybridization under the control condition. Second, all values, including those of the control condition hybridizations, must then be standardized to this virtual standard. Otherwise, cross-comparison between conditions or clustering of gene or experiment profiles will not be possible. Taking the median from all replicates of a certain condition is only meaningful after standardizing. This applies also to cases where each condition is accompanied by its own control hybridization. Also here, all data sets must be standardized to the same virtual standard in order to allow inter-condition comparison.

Discussion

The technique of expression profiling by means of hybridization to high density DNA arrays offers a new tool to investigate the expression levels of thousands of genes at the same time. Differentially expressed genes should, in theory, be detectable by comparisons of hybridization data from pairs or series of experiments provided that the signal intensities are precisely related to the proportion of the complementary mRNA in the mRNA pool. But comparison is hampered by the fact that differences in signal intensities might not only be due to true expression changes but also to experimental variabilities, which are often in the same range as the differences one expects to occur by differential expression. Thus, careful correction for the various influences on the experiment, like incorporation of radioactive label or exposure time, is needed.

The most basic data processing techniques, background correction and linear transformation of the data, have been described in this paper. There is, however, discrepancy in the literature about the methods to find an adjusting factor for standardization. Chen *et al.* (1997) use a ratio distribution function derived for two normally distributed data sets with constant variance to standardize by an iterative procedure. Piétu *et al.* (1996) standardize by subtracting the mean of the logarithmic data and dividing by their standard deviation. Richmond *et al.* (1999) calculate the relative percentage of total signal as a means of standardization. We have built our standardization procedure on a linear model of systematic influences on microarray gene expression data. The assumptions underlying this model have been tested on several hundred hybridizations and found to be sound. In contrast, we have found that intensity values on an array do not in their entirety follow a lognormal distribution. Hence, this distribution may be used to define the background on one array but is not suited to standardize for comparison between hybridizations.

Correction for background or linear distortion, however, does not help with respect to the variability of the experimental data. To reduce this variability it is desirable to have several repetitions of the same experiment. We suggest including all data, rather than averaging since the individual differences give a rough estimate of the data quality. We would like to visually inspect these differences if a gene is believed to be differentially expressed. The reliability may be judged from the 'min–max separation' or the 'std separation'. Before calculating ratios of intensity values, we compute the median of the data because the median is more robust to outliers than the mean.

We often observed that the data in the low-intensity region of an array hybridization can be well fitted with a normal distribution function (Figure 1). The fitted function may be used to obtain an estimate of the proportion of spots in a given intensity class belonging to the normal distribution. This portion is believed to correspond mainly to non-expressed genes. However, we refrain from setting a fixed arbitrary threshold using this function as do Piétu *et al.* (1996). Rather, the predicted normal density in a particular interval provides a way to estimate the probability that an observed intensity is actually due to an expressed gene, rather than to experimental noise.

The methods introduced in this paper are only a prerequisite to a more thorough study of the data in order to reveal the inherent information. This is of particular interest when dealing with data from a series of experiments, e.g. with a time-course, a concentration series or different tumor stages, where changes in expression levels are more difficult to detect by pairwise comparison. We have had very encouraging experiences with various clustering methods (Carr *et al.*, 1997; Eisen *et al.*, 1998) and with methods of embedding high-dimensional data in a plane, or in three dimensional space (Spanakis and Brouty-Boyé, 1997; Khan *et al.*, 1998; Hilsenbeck *et al.*, 1999). Our developments in these methods will be discussed elsewhere.

Acknowledgements

We gladly acknowledge the following people for their help, discussions or suggestions: Rainer Spang, Tobias Müller, Richard Desper, Bernhard Korn, Hilmar Lapp (Novartis) and Maja Vujic. The presented work was carried out at the DKFZ, Department of Theoretical Bioinformatics. Data were generated at the Departments of Molecular Cell Biology I headed by Günther Schütz, Functional Genome Analysis headed by Jörg Hoheisel and Molecular Genome Analysis headed by Annemarie Poustka. Work was funded by the *Deutsche Forschungsgemeinschaft*, the *Bundesministerium für Bildung und Forschung*, and by the European Commission.

References

Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F., Kerlavage,A.R., McCombie,W.R. and Venter,J.C. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Carr,D.B., Somogyi,R. and Michaels,G. (1997) Templates for looking at gene expression clustering. *Statistical Computing and Statistical Graphics Newsletter*, **8**, 20–29.

Chen,Y., Dougherty,E.R. and Bittner,M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.*, **2**, 364–374.

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**, 14863–14868.

Friemert,C., Erfle,V. and Strauss,G. (1989) Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. *Methods Mol. Cell. Biol.*, **1**, 143–153.

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

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

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

Lennon,G.G. and Lehrach,H. (1991) Hybridization analyses of

arrayed cDNA libraries. *Trends Genet.*, **7**, 314–317.

Piétu,G., Alibert,O., Guichard,V., Lamy,B., Bois,F., Leroy,E., Mariage-Samson,R., Houlgatte,R., Soularue,P. and Auffray,C. (1996) Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.*, **6**, 492–503.

Richmond,C.S., Glasner,J.D., Mau,R., Jin,H. and Blattner,F.R. (1999) Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Res.*, **27**, 3821–3835.

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, **16**, 939–945.

Sachs,L. (1984) *Applied Statistics*. 2nd edn, Springer, Berlin, pp. 107–110.

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.

Spanakis,E. and Brouty-Boyé,D. (1997) Discrimination of fibroblast subtypes by multivariate analysis of gene expression. *Int. J. Cancer*, **71**, 402–409.

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

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

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,S.R., Vogelstein,B.E. and Kinzler,K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.