Transcriptional Profiling using DNA Array Data

Deutsches Krebsforschungszentrum Heidelberg

Nicole Hauser[1], Kurt Fellenberg[2], and Tim Beißbarth[2,3]

[1] Functional Genome Analysis (H0800)
[2] Theoretical Bioinformatics (H0300)
[3] Molecular Biology of the Cell I (A0200)

Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280
D - 69120 Heidelberg
Germany

Abstract

Subsequent to sequencing projects new strategies are required for efficient functional characterization of the identified genes. DNA array technology was proven to be a powerful tool by monitoring the transcriptional levels of all or a significant portion of genes of an organism simultaneously. Here, we present data obtained by spotting PCR products derived from yeast ORFs as high density arrays onto nylon or polypropylene membranes. Total RNA of yeast cells was then prepared, radioactively labeled and hybridized to such arrays. The signal intensity detected for each spot in the array is at least a semiquantitative measure of the expression level of the corresponding gene.

Transcriptional profiles can be obtained comparing multiple conditions, like varying growth media or temperatures and different mutant strains. In order to compare these conditions, the hybridization signal intensities must be brought to roughly the same level ('normalization'). Systematic differences are corrected for by linear regression. Changes in the expression levels of certain genes are detected by direct comparison to a control condition.

Coexpressed genes, i.e. genes that show similar profiles across all conditions are found by clustering techniques. It is often assumed that these genes are functionally related. Correspondence analysis reveals under which of several conditions the transcriptional profile of an organism will change significantly, as well as the identity of the involved genes.

Spot Organization

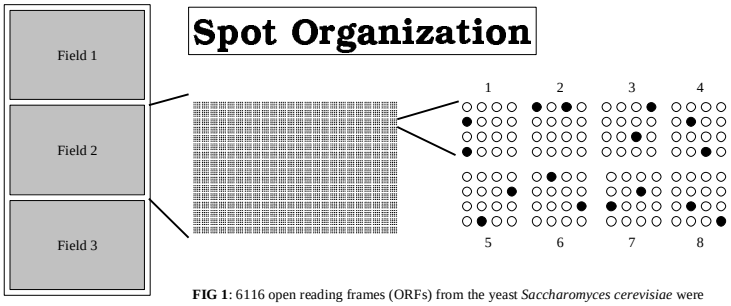

FIG 1: 6116 open reading frames (ORFs) from the yeast *Saccharomyces cerevisiae* were PCR-amplified on genomic DNA. PCR products were spotted in duplicate in a defined pattern using robotic devices.

Normalization

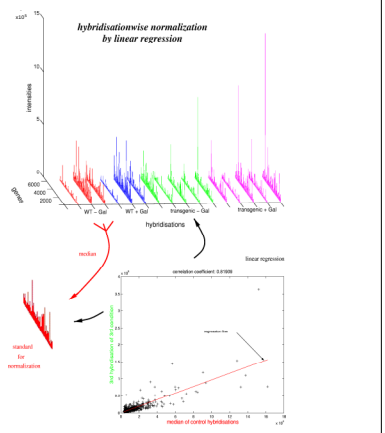

FIG 3: In a multiconditional experiment, hybridizations may differ both by additive background or multiplicative factors [1]. While an additive offset will destroy the proportions within the data, a difference in e.g. incubation time influences the ratios multiplicatively.

In order to prepare a proper comparison we correct for these distortions by means of linear regression. Each of our time courses or other multiconditional studies (e. g. the experiment shown above and explained in Fig. 5) contains a control condition which serves as an anchor for comparison (also with other experiments). We use the median for each gene of the control condition hybridizations as a standard for normalization x. This vector consists of intensity values x_i for every gene i spotted on the membrane. Assuming that the vast majority of the genes does not change their transcription levels under the applied conditions, we normalize a hybridization y by fitting the regression line $y_i = mx_i + b$ into the scatter plot of y_i against x_i before calculating the fitted hybridization f as $f_i = (y_i - b) / m$. By applying this procedure to every hybridization of the multiconditional experiment, all hybridizations are adjusted to the same level and are thus made comparable.

Hybridization and Spot Detection

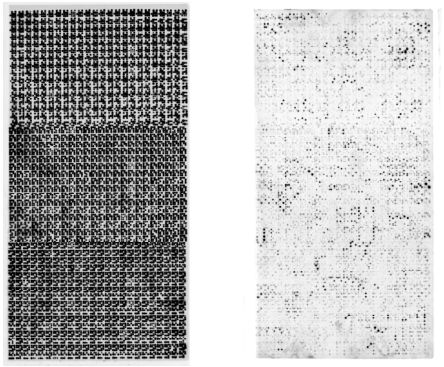

FIG 2: By hybridization with the universal 5'-sequence, introduced during PCR, each spot can be detected (left). Different transcript levels can be detected after hybridisation with a complex cDNA probe made from total RNA (right).

Differentially transcribed Genes

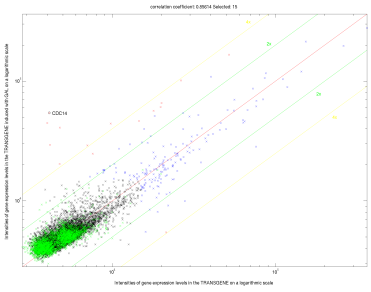

FIG 4: Two individual hybridization experiments done with yeast transfected with CDC14 under a GAL inducible promoter are compared, induced vs. not induced.

Measured expression levels of the genes under two different experimental conditions can be compared using a scatter plot. Here the intensities are plotted versus each other using logarithmic scales. Crosses close to the diagonal refer to genes that are equally expressed under both experimental conditions, points on the green line represent genes which are 2 fold up or down regulated and points on the yellow line represent genes which differ by a factor of 4. All genes that differ by more than a factor of 3 are marked in red. Black crosses represent genes that display an intensity close to the background level, they tend to produce erroneously high ratios. The green crosses mark empty positions on the membrane where no DNA is spotted.

Clustering by Correspondence Analysis

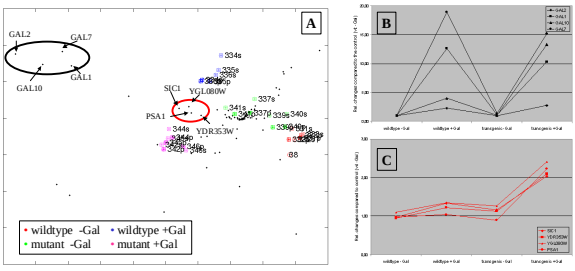

FIG 5: Two strains being wildtype and transgenic respectively for one gene were grown under two different conditions. In the transgenic strain the CDC14 gene was cloned to be inducible by galactose. Grown in standard medium containing glucose, both strains showed now significant phenotype, while the transgenic strain grown in galactose medium showed a dramatic increase of Cdc14p and an arrest of the cells in mitosis (data not shown).
RNA material derived from the 4 conditions (wt –Gal, transgenic –Gal, wt +Gal and transgenic +Gal) were used for DNA-array hybridization [2]. Raw intensity data of repetitive hybridizations were subjected to normalization (see Fig. 3), resulting in fitted data being comparable with each other. Several filters (e.g. intensity threshold, quality assessment) were applied to the fitted data. 154 genes with significant changes (> factor 2) of transcript level in either condition compared to the control (wt –Gal) could be identified. Clustering of genes and hybridizations was done by a multi dimensional scaling algorithm, were genes are described in the experimental-dimensional space and experiments in the gene-dimensional space and both reduced to two dimensions with a minimal loss of information. The two plains are combined into one plot by correspondence analysis (A). Genes are plotted as black dots while repetitive hybridizations are displayed in their respective color (see legend). Two gene clusters showing characteristic transcription profiles are shown exemplarily (B, C).

Conclusions

DNA arrays were proven to be a useful tool for detailed analysis of transcriptional activity in yeast cells. Intelligent analysis is crucial to reveal the information that is inherent in the enormous amount of data.
In order to make hybridizations within a series comparable, one has to correct for additive background and constant factors of distortion (Fig. 3) prior to further analysis.
Differentially transcribed genes can be detected by direct comparison of signal intensities under two conditions in a scatter plot (Fig. 4). When dealing with multiple conditions, clustering of genes and experiments may be more useful (Fig. 5). Coexpressed genes can be identified, and this may give a hint for functional classification of yet unknown genes.
We are currently developing a database for storage of experimental data that provides also the analysis tools presented here. The database will be available by use of a web-based interface. Clone names will be linked to gene annotations in public databases. However, an entire network of biochemical, physiological and genetic studies to be connected to transcriptional analyses is needed to assign functions to non-annotated genes and to better understand cell function.

References

[1] Beißbarth, Fellenberg *et al.*, *Bioinformatics*, submitted
[2] Hauser *et al.*, 1998, *Yeast 14*, 1209-21.