Dynamic changes in gene expression of the cyanobacterium Synechocystis sp. PCC 6803 in response to nitrogen starvation

Krasikov, V.

Publication date
2012

Chapter 3

Method development for DNA-microarray data analysis of cyanobacteria

Method development for DNA-microarray data analysis of cyanobacteria

Vladimir Krasikov, Eneas Aguirre von Wobeser, Jef Huisman, and Hans C. P. Matthijs

Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam,
P.O. Box 94248, 1090 GE Amsterdam, The Netherlands

Summary

To investigate the response of cyanobacteria to changing environmental conditions, we designed an oligonucleotide DNA microarray of the unicellular freshwater cyanobacterium *Synechocystis* sp. strain PCC 6803. Our goal was the quantification of changes in gene expression across the entire genome of *Synechocystis*, by comparing gene expression patterns under different environmental conditions against a standard control. For this purpose, we developed a pipeline scheme for the analysis of experimental micro-array data. This scheme includes data import, correction for background noise, data normalization, fitting linear models to the data, and corrections for multiple hypothesis testing. Our data analysis makes use of the R language, a free open-source language for statistical computing. The output consists of lists of differentially expressed genes and an assignment of the statistical significance of the observed changes in gene expression. In addition, visualisation tools for facilitated data quality control and investigation of expression patterns by clustering algorithms are provided.

Introduction

Microarray technology is among the most promising tools available today to researchers in the life sciences (e.g., Schena *et al.*, 1995; Brown and Botstein, 1999; Hughes *et al.*, 2001). Microarray studies enable interrogation of the relative transcription of thousands of genes at the same time. Microarrays are solid substrates hosting large numbers of single-stranded DNA probes with a specific sequence, each precisely localized on a grid. In total, these probes may cover the entire genome of an organism. The probes on the grid are designed to hybridize with single-stranded cDNA molecules named targets. This target cDNA originates from the reverse transcription of mRNA obtained from the organism under investigation. Ideally, the probes on the grid have been selected such that hybridisation with target cDNA falls into a narrow range of melting temperatures to optimise selectivity via washes. As a result, washing at a selected temperature leaves only the hybridised pairs of probe and target attached to the grid. The target cDNA has been labelled with fluorescent dye during reverse transcription. Hence, the intensity of the fluorescence signals from the grid indicates the amount of mRNA isolated from the organism under investigation.

The elegance of the method is that changes in gene expression patterns of the organism can be quantified by using two different dyes, a red fluorescent and a green fluorescent dye (Churchill, 2002). For instance, cDNA obtained from organisms exposed to a specific experimental treatment (e.g., nitrogen limitation) may be stained red while the cDNA obtained from control experiments may be stained green. Subsequently, mixtures of cDNA from both the experimental treatment and the control are hybridized with the probes on the grid. The relative abundances of green and red fluorescent cDNA than give rise to fluorescent spots on the grid that range from pure green, via different shades of yellow (different ratios of green to red), to pure red. The colour of the fluorescent spots on the microarray thus reflects the relative expression of mRNA of the experimental treatment versus the control. In the above example, red spots indicate up-regulated genes and green spots indicate down-regulated genes in response to the experimental treatment. The microarray is scanned and the resulting image is analyzed such that the signal intensity from each spot on the grid can be quantified. The signal intensities are translated into long lists detailing the expression rates of all probes in the experimental treatment versus the control. In principle, these large data sets permit an answer to the biological key question: which genes are specifically regulated under which environmental conditions?

Raw data from microarray studies contain numerous sources of variability, however, which obscure straightforward determination of differentially expressed genes (Kerr and Churchill, 2002; Quackenbush, 2002). The sources of variability can be divided into several categories. Part of the variability is related to sample preparation, include RNA isolation, reverse transcription, and incorporation of the fluorescent dyes. Variability associated with the microarray slides is due to inhomogeneities in the spots and slides arising during microarray fabrication. Hybridization-related variability is determined by the hybridization characteristics of the probes and targets, non-specific hybridization and non-specific

background normalization, and the potential inclusion of artefacts like dust particles during the hybridization steps. The above-mentioned sources of variability give rise to noise in the data, and the only weapon against this is suitable replication of experiments and the application of appropriate statistical techniques afterwards (Draghici, 2003).

In recent years, the need for statistical tools to analyze microarray data has resulted in a wealth of new computational methods in the field of bioinformatics. Normalization of raw data is one of the crucial first steps in microarray data analysis (Kerr and Churchill, 2002; Quackenbush, 2002). The aim of normalization is to correct for systematic differences across data sets (e.g., differences in the overall intensity of spots) and to eliminate artefacts (e.g., nonlinear read-out of dye fluorescence). Several normalization approaches are available, including techniques based on analysis of variance (Kerr and Churchill, 2001; Wolfinger *et al.*, 2001) and q-spline normalization (Workman *et al.*, 2002). After proper normalization, the classical metaphor of the needle in the haystack easily becomes an accurate description when thousands of genes are investigated, especially if only a small proportion of these genes have responded to the experimental treatment. Classical statistical approaches cannot be applied to determine significant changes in gene expression, because the number of response variables (thousands of probes) in microarray experiments is much greater than the number of experiments (usually not more than ten). This would easily lead to the detection of numerous false positives (i.e., changes in gene expression that are declared significant, while in fact they are not). Therefore, the assignment of statistical significance in microarray experiments requires corrections for multiple hypothesis testing (e.g. Bonferroni correction, or control of the false discovery rate; Cui and Churchill, 2003; Ge *et al.*, 2003; Storey and Tibshirani, 2003).

This chapter serves as an illustration of the bioinformatics approach for microarray data analysis. We designed an oligonucleotide microarray of the unicellular freshwater cyanobacterium *Synechocystis* sp. strain PCC 6803. The array consists of 8091 probes, covering the full genome of 3264 Open Reading Frames (ORFs). To analyze the large data sets obtained from these microarrays, we present a pipeline scheme that includes data import, correction for background noise, data normalization, fitting linear models to the data, and corrections for multiple hypothesis testing. The pipeline scheme makes use of the Limma package (Smyth, 2004, 2005), which is a software package for microarray data based on the open-source language R for statistical computing (Dalgaard, 2002; Gentleman *et al.*, 2005). Our approach is illustrated with examples from a detailed study of changes in gene expression of cyanobacteria in response to nitrogen starvation.

Materials and Methods

DNA microarray platform. We designed a custom-made 60-mer oligonucleotide microarray, which has been *in situ* spotted on glass slides by Agilent Technologies (Palo Alto, USA) as described by Hughes *et al.* (2001). In total, 8091 probes were designed to cover the 3264 genes of *Synechocystis sp.* PCC 6803 (Kaneko *et al*., 1996). The full sequence and annotation of *Synechocystis* is freely available

at CyanoBase (www.kazusa.or.jp/cyano/). For each gene, 1 to 4 different oligonucleotides were used as probes in the array. Each probe was placed in an accurately known position in the matrix. Probes for the same gene were placed at randomized distances from each other to exclude local effects. Hybridization conditions were as described in the "Agilent *in situ* hybridization kit-plus" and in the "Agilent 60-mer oligo microarray processing protocol, SureHyb, SSPE Wash, version 2.0" protocols (Agilent Technologies, Palo Alto, USA). Slides were scanned at 10 micron resolution in an Agilent microarray scanner and image processing was performed by Feature Extraction Software version 7.5 (Agilent Technologies) providing one data file for each microarray experiment. A microarray experiment is defined as a comparison of a control and a treatment on one array. Our full experimental design involved 3 replications with biological controls.

Software. We used the open-source language R Version 2.8.0 for statistical computing (www.r-project.org), and the software package Limma to test for differential expression patterns (Smyth, 2005; www.bioconductor.org). Limma is an R package for the analysis of gene expression microarray data, using linear models for the assessment of differential gene expression. The data treatment included data import; pre-processing with background subtraction, within- and between-array normalisation, visualisation of the normalization procedure, application of linear models to the normalized data, extraction of differential expression patterns from the data with assignment of the statistical significance, data export, visualisations of the obtained results, and processing data in Excel. The complete R script that we developed for this work is presented in the Supplementary Materials (Appendix 3).

Experimental layout. Nitrogen starvation was induced by culturing *Synechocystis* sp. PCC 6803 in BG11 medium (Rippka *et al.*, 1979) without a nitrogen source. Cells were collected after 0, 6, 12, 24, 96 h of nitrogen starvation. After 96 hours, nitrogen was added to default BG11 medium concentration of 18 mM $NaNO_3$, and cells were subsequently collected after 6 and 12 h to monitor the recovery from nitrogen starvation. The experiments were run in triplicate. The reference time point at t = 0 h was based on 6 replicates, which were pooled and used as reference in all experiments. RNA was isolated by hot phenol chloroform extraction, LiCl precipitation, and clean-up with RNeasy mini kit (Qiagen, Germany). Target cDNA of the reference samples at t = 0 h was labelled with a green fluorescent dye (Cy3) and the target cDNA of all treatment samples was labelled with a red fluorescent dye (Cy5) in a direct reverse transcription reaction. Further details on the experimental procedures are presented in Chapters 4 and Chapter 6.

Results and Discussion

Data import and initialization files. Each of 3264 ORFs *Synechocystis* PCC 6803 was represented by two to four different probes, yielding a total of 8091 probes. The probes were described in a targets description file. This file is basic to the array design and to all our subsequent microarray experiments. In R the name of each experiment and the names of targets in the red and green channels are assigned. Figure 1 shows a raw image of one of the microarrays obtained from a nitrogen starvation experiment. Green spots indicate down-regulated genes, red spots indicate up-regulated genes, while yellow spots are intermediate. Raw intensity data files are interpreted by Agilent image processing software Feature Extraction 7.5 that creates separate tab-delimited text file for each microarray. Raw data are

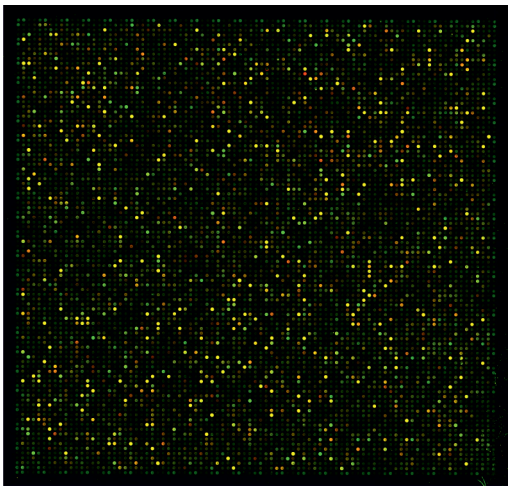

Figure 1. Raw microarray image. Image produced by data extraction software shows raw data for one of the microarrays in the nitrogen starvation experiment.

loaded inside the R environment with help of *read.maimages*() function producing the R(ed) G(reen)-object (Fig. 2A). This RG-object contains information about the intensity of each spot in both the red and green channels, and includes the spot's annotations as well.

Background correction and normalization procedures. Background correction was performed according to the "minimal method" of the Limma package, using the *backgroundCorrect* function. This method is used to avoid negative intensity values after background substraction. With this method any intensity which is zero or negative after background subtraction is set equal to the half-minimum of the positively corrected intensities for that array. Background correction proceeded with "within-array normalization" for each microarray separately using global LOWESS normalization (LOcally WEighted polynomial regression; Cleveland and Devlin, 1983) with the *normalizeWithinArrays* function, which is used to make the average fluorescence emission of the R and G channels on the single array comparable. At this step, data are log-transformed using log base 2. For each spot, we calculated the mean log intensity A and the log ratio M, defined as $A=(1/2)(^{2}\log R+^{2}\log G)$ and $M=^{2}\log(R/G)$, where R and G are the intensities in the red and green channels, respectively (Yang et al., 2002; Smyth and Speed, 2003). Next, "between-array normalization" is performed with the *normalizeBetweenArrays* function. Here, we applied A-quantile normalization, which ensures that the A-values have the same empirical distribution across arrays leaving the M-values unchanged (Yang and Thorne, 2003). Figures 2A and 2C show the RG-plot and MA-plot of the raw data, while Figures 2B and 2D show the RG-plot and MA-plot after normalization The results of the normalization procedure for all microarrays in the nitrogen starvation experiment are visualized by plotting the density distributions of the intensity in each channel before normalization (Fig. 2E) and after normalization (Fig. 2F). An

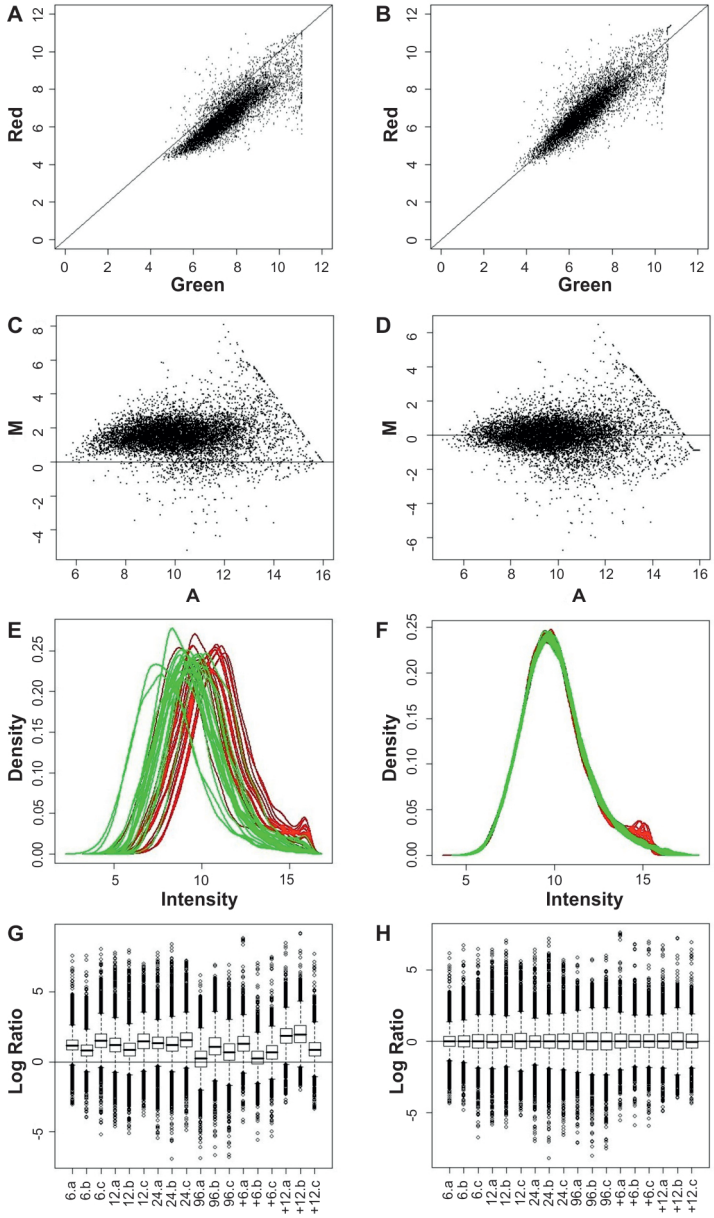

Figure 2. Normalization of the microarray data. Raw data from one of the microarrays plotted in a RG-plot (**A**) and MA-plot (**C**). Normalized data from the same microarray plotted in a RG-plot (**B**) and MA-plot (**D**). Density distribution of the intensities across the red and green channels for the raw data (**E**) and normalized data (**F**) of all microarrays in the experiment. Box-plots of the log ratios observed in the microarrays, using the raw data (**G**) and normalized data (**H**). The x-axis indicates the experimental time points (6, 12, 24 and 96 hours of starvation and +6 and +12 hours of recovery) and biological replicates (a, b, c) of each time point. The boxes span the range of log ratios that fall in the middle 50% of the distribution. The whiskers show extreme upper and lower observations on each microarray.

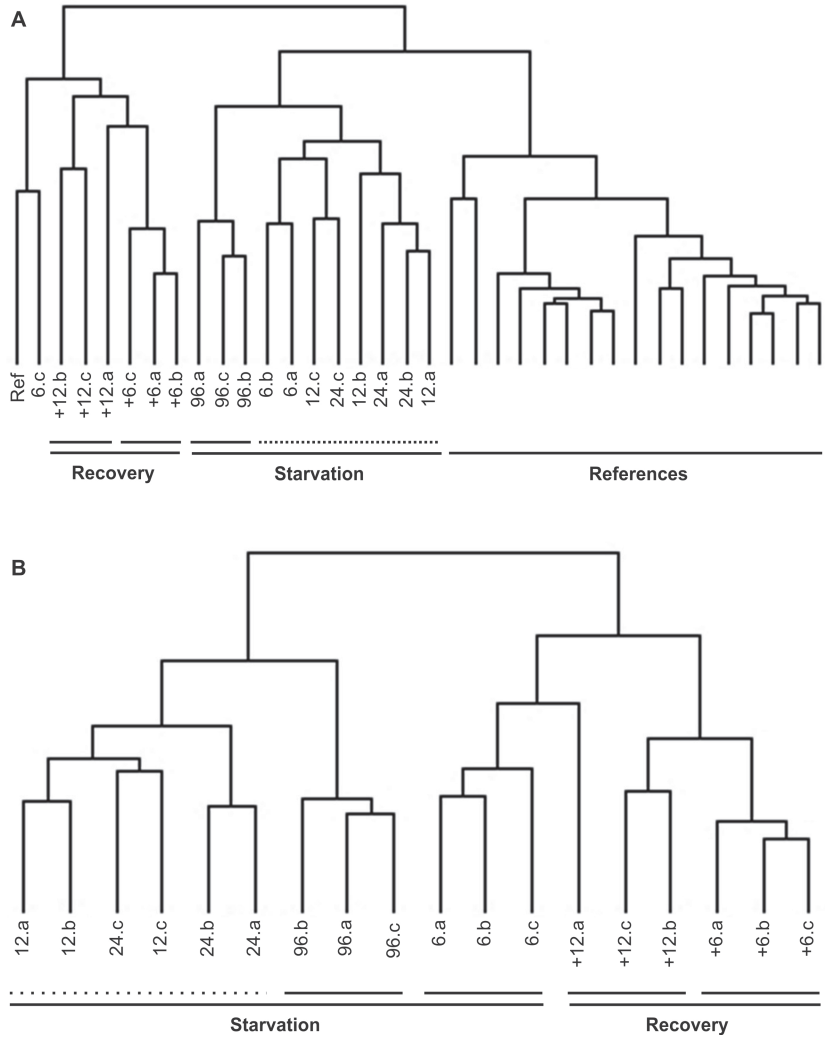

Figure 3. Clustering of all replicates in the experiment. Images represent values for the individual spots on the microarray. Hierarchical clustering is performed across replicates and time points of the experiment producing a similarity tree. Nitrogen starvation and recovery time points are underlined. (**A**) Clustering according to the intensities observed in individual channels form tree major branches. The reference experiments form a tight cluster in the right part of the tree. Clusters representing specific time points in the experiment (96 hours of nitrogen starvation; 6 and 12 hours of recovery) are underlined with solid lines. The mixed cluster of 6, 12 and 24 hours of nitrogen starvation is underlined with a dashed line. (**B**) Clustering according to the log ratio of the red to green channels forms two tight clusters. Clusters representing specific time points are underlined with solid lines; the mixed cluster of 12 and 24 hours of starvation is underlined with a dashed line.

alternative visualization is presented by so-called "box-plots", which highlights the smallest observation, lower quartile, median, upper quartile, and largest observation of the log ratio in each microarray (Fig. 2G, H).

To evaluate the overall performance of the microarray experiment and to test the effectiveness of the normalization procedure, we performed hierarchical clustering of all replicates (Johnson, 1967). Clustering across replicates visualizes similarities of the replicates in the experiment. Clustering according to the intensities of the individual red and green channels across all replicates and all time points of the nitrogen starvation experiment revealed that all reference experiments at t = 0 h, except one, cluster close together in branches on the right-hand side of the tree (Fig. 3A). Furthermore, the 6, 12, and 24 hours replicates form one mixed cluster in the center of the tree, highlighting the observation that expression at this time points was quite similar. The 96 hours starvation replicates also form one conspicuous branch of the tree, close to the earlier time points of the starvation experiment, and quite dissimilar from the replicates after 6 and 12 hours of recovery. Clustering according to the log ratios observed in individual microarrays revealed that expression profiles after 6 hours of starvation are close to the profiles in the recovery mode, the profiles after 12 and 24 hours starvation are close to each other, and the 96 hours replicates are quite distinct (Fig. 3B). The appearance of replicates in tight clusters is an indication of a correct normalisation procedure. We also tried other normalization procedures, including "variance stabilization and normalization" (from the vsn package, Huber *et al.*, 2002) and Median normalization (from the marray package, Dudoit *et al*., 2002). These alternative normalization procedures produced less structured trees (data not shown), and were therefore considered less suitable for our data set.

Determination of differentially expressed probes. The design of any microarray experiment can be represented in terms of linear model for each gene. The suitably normalized data were fitted with a probe-wise linear models (Yang and Speed, 2003; Smyth, 2004). Limma uses an empirical Bayes method to moderate the standard errors of the estimated log ratios (Efron and Tibshirani, 2002). This moderated t-statistic is the ratio of the M-value to its standard error and has the same interpretation as an ordinary t-statistic except that the standard errors have been moderated across probes, i.e., shrunk towards a common value. This has the effect of borrowing information from the ensemble of genes to aid with inference about each individual gene. This results in an assignment of the log ratio to each probe and of the p-value as a measure of the significance of the observation. These p-values were adjusted for multiple hypothesis testing by control of the false discovery rate (Hochberg and Benjamini, 1990; Reiner *et al.*, 2003). False discovery rate (FDR) was set to the level of 0.01 and probes with adjusted p-values less than FDR were judged as differentially regulated.

The resulting table of differentially regulated probes was exported into text format for downstream analysis in Excel, and may be used to graphically represent significant changes in the MA-plot (Fig. 4), to produce Venn diagrams (Fig. 5), or hierarchical clustering of similar expression behaviour for selected groups of genes (Fig. 6).

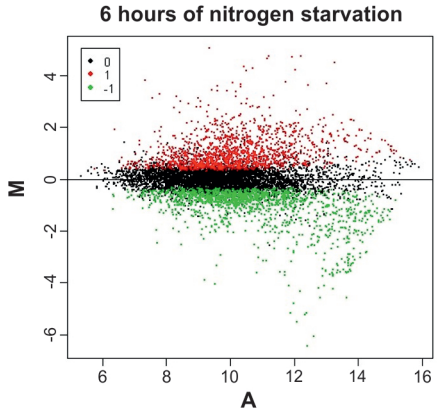

Figure 4. Graphical representation of differentially expressed genes in a MA plot. MA plot after 6 hours of nitrogen starvation; statistically significant probes are highlighted in red for up-regulated probes and in green for down-regulated probes.

Figure 5. Venn diagram of up- and down-regulated probes after 6, 12 and 24 hours of nitrogen starvation. The number of up- and down-regulated probes after 6, 12 and 24 hours of nitrogen starvation is shown. The total number of probes investigated is given in the lower right corner.

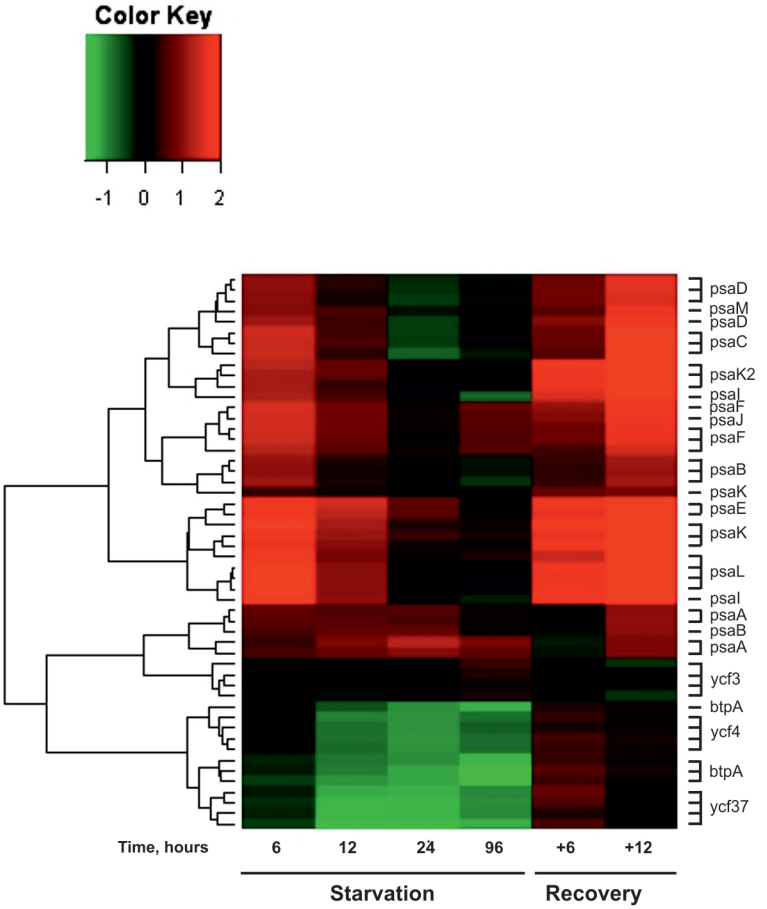

Figure 6. Example of cluster analysis for a functional group of genes. Hierarchical clustering of probes for the selected structural genes of Photosystem I during the nitrogen starvation experiment. Although the probes are located wide apart on the microarray grid, they show similar expression behaviour. Genes that are up-regulated (red) or down-regulated (green) differentially in time form separate clusters.

Selection of differentially expressed genes. As each gene on the microarray is represented by several different probes, a gene was judged to be regulated if all probes for that gene had p-values of less than 0.01. The p-value for that gene was set as the maximum of the p-values of the corresponding probes, and the log ratio of the gene was calculated as the mean log ratio of the corresponding probes. The resulting tables of differentially expressed genes were analysed to investigate dynamic changes in gene expression during the starvation experiment, as will be reported in subsequent chapters.

Conclusion and application

We presented a pipeline scheme for exploratory microarray data analysis. It includes statistical data treatment, and results in an inventory of differentially expressed genes. Differences in gene expression are the consequence of the treatment condition, and different environmental conditions typically give rise to specific patterns of gene expression. In this thesis, the emphasis has been on nitrogen starvation. The pipeline scheme is applied to study patterns of gene expression in a pilot experiment after 12 hours of nitrogen starvation (Chapter 4), and in a time-series nitrogen starvation experiment followed by recovery from the nitrogen stress (Chapter 6).

Acknowledgements

The research of V. K. and J. H. was supported by the Earth and Life Sciences Foundation (ALW), which is subsidized by the Netherlands Organization for Scientific Research (NWO). E. A. v. W. was financially supported by a scholarship from Consejo Nacional para la Ciencia y Tecnología (Mexico).

List of supplemental data

Apendix 3. R-script used to assess differential expression patterns in the nitrogen starvation experiments.

References

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**: 33-37

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32**: 490-495

Cleveland W, Devlin S (1983) Locally weighted regression: An approach to regression analysis by local fitting. *J Am Stat Assoc* **83**: 596-610

Cui XQ, Churchill GA.(2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* **4**: 210

Dalgaard P (2002) Introductory Statistics with R. Springer, New York

Draghici S, Kuklin A (2003) Data Analysis Tools for DNA Microarrays. CRC Press, New York

Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**: 111–139

Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* **23**: 70-86

Ge YC, Dudoit S, Speed TP (2003) Resampling-based multiple testing for microarray data analysis. *Test* **12**: 1-77

Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds.) (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York

Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. *Stat Med* **9**: 811-818

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stablization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics* **18**: 96–104

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley P (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**: 342-347

Johnson SC (1967) Hierarchical Clustering Schemes. *Psychometrika* **2**: 241-254

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**: 185-209

Kerr MK, Churchill GA (2001) Statistical design and the analysis of gene expression microarray data. *Genet Res* **77**: 123-128

Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* **32**: 496-501

Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**: 368–375

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol* **111**: 1-61

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467-470

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. *Methods* **31**: 265-273

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article 3

Smyth GK (2005) Limma: linear models for microarray data. *In*: Bioinformatics and Computational Biology Solutions using R and Bioconductor. Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (eds.), Springer, New York, pp. 397–420

Storey J. D., and Tibshirani R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**: 9440-9445

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**: 625-637.

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**: research0048

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**: e15

Yang YH, Speed TP (2003) Design and analysis of comparative microarray experiments. *In*: Statistical Analysis of Gene Expression Microarray Data, Speed TP (ed.), Chapman & Hall/CRC Press, London, pp. 35-91

Yang YH, Thorne NP (2003) Normalization for two-color cDNA microarray data. *In*: Science and Statistics, Goldstein DR (ed.). Institute of Mathematical Statistics Lecture Notes – Monograph Series, Volume 40, pp. 403-418