Genome-wide expression analysis of environmental stress in the cyanobacterium Synechocystis PCC 6803

Aguirre von Wobeser, E.

Publication date
2010

Chapter 1

General Introduction

THE GENOMIC REVOLUTION

Deoxyribonucleic acid (DNA) carries the information of living cells. The defined order in which the monomers C, G, A and T are lined up in a sequence contains the unique information that a living cell needs for its functions and multiplication. In prokaryotes, the information encoded by DNA is organized in discrete stretches, called open reading frames (hereafter referred to as genes). The genes' DNA is transcribed and renders a product called messenger RNA (mRNA). Each gene determines the sequence of a unique mRNA that is subsequently translated at the ribosomes into distinct polypeptides (proteins). Polypeptides are then folded into proteins, which perform the structural and enzymatic functions required by a living cell. The relative abundance of mRNA transcripts is a measure for the frequency at which the corresponding genes are transcribed, and can be used as a witness for the physiological status of the cell.

To understand the processes that an organism performs, biochemistry and molecular biology have traditionally focused on the functions of individual genes and the proteins they encode. These efforts have lead to a vast number of proteins being studied, and their genes being sequenced. Ultimately, researchers desired a complete view of the genomic information of model organisms, and endeavored into sequencing their whole genome sequences. The first article announcing a complete genome sequence of a free-living organism (the bacterium *Haemophilus influenzae* Rd), stated in its first sentence (Fleischmann *et al.*, 1995):

"A prerequisite to understanding the complete biology of an organism is the determination of its entire genome sequence."

Indeed, the genome sequence of an organism gives a nearly complete overview of what an organism is capable of, in biochemical terms. By searching for homology between the predicted genes of a new genome and all known protein sequences (Altschul *et al.*, 1990), one can search for ortholog genes; which are genes that probably have the same origin, and also the same function (e.g., Kaneko *et al.*, 1996). This process is known as genome annotation, and leaves usually many open reading frames without functional assignment for their protein product because of a lack of similar proteins in the databases. 'Genes' left without annotation are usually involved in functions specific to the form of life of the newly sequenced organism. Many of these genes are found to be conserved in taxonomically related organisms (Martin *et al.*, 2003).

The availability of genome sequences has transformed the way molecular biologists work. The genome information delimits the working space within a species, and is used as a reference point for any further molecular biology studies. Researchers aim to relate their findings to genes identified in the genome sequence (e.g., Chatterjee *et al.*, 2004). Moreover, genome sequences allow studies to be conducted at a global (i.e., whole-genome) level, focusing on how a phenomenon affects all the genes of an organism simultaneously. These applications define the genomic era, an era of the study of the cell as a whole. At present, about 600 species have been fully sequenced, and the number is rapidly increasing (www.genomesonline.org).

THE GENOMIC ERA AND MICROARRAY ANALYSIS

The first genome sequences generated great expectations on the knowledge and applications they would provide. However, the information encoded by the genomes proved to be complex, and its in-depth utilization is requiring further technological, mathematical and conceptual developments.

One field of research that has amply benefited from genome sequences is phylogenetics, the study of the evolutionary relations between organisms. Whole genomes allow sequence comparisons to be made using all the available information, as an extension of traditional molecular phylogenetics, where the evolutionary distance between certain ubiquitous sequences is compared (Eisen, 2000). Also, genomic sequences provide additional phylogenetic markers, like the presence or absence of genes or biochemical pathways, the arrangements of genes in the genomes, etc. (Bansal and Meyer, 2002; Boore, 2006).

Genome sequences are also utilized for high throughput analysis of physiological responses and capabilities of organisms, usually in conjunction with other technological developments. Studies analyzing the way cells utilize their genomic information under different environmental conditions or in specific mutants focus on the accumulation of mRNA (transcriptomics) or proteins (proteomics). Still other studies measure quantitative changes in the concentrations of metabolic products (metabolomics). While proteomic studies are more directly related to physiology, they can only identify a subset of the proteins in a cell (Fulda *et al.,* 2000; Simon *et al.*, 2002; Herranen *et al.*, 2004). Transcriptomics studies, targeted at the earlier mRNA stage of the gene expression process, provide information about the primary responses of the cells to the growth conditions under investigation, that may or may not result in changes in protein accumulation. However, gene expression studies correlate fairly well with proteomics data (Conway and Schoolnik, 2003; Suzuki *et al.*, 2006).

Whole-genome gene expression is analyzed using microarrays, first introduced by Schena *et al.* (1995). Microarray analysis uses Watson and Crick base pairing to detect the concentration of mRNA for each of the different genes individually in the RNA pool extracted from cells. One or more probes complementary to the mRNA of each gene are attached in multiple copies to a rigid surface (typically made of glass), in a matrix of well-localized spots (Figure 1). The labile mRNA, or rather a reverse transcribed stable DNA copy of it (cDNA) is labeled with a fluorescent dye or radioactively, and is exposed to the microarray for complementary hybridization. The fluorescence or radioactivity signal observed in each spot on the microarray is an indication of the concentration of the mRNA specific for the corresponding probe.

Several variations in microarray technology exist, that differ in the length of the probes and in the way they are constructed (Lockhart and Winzeler, 2002; Heller *et al.*, 2002). One type of microarrays, known as spotted microarrays, use PCR products that are deposited by a robot in spots on the slides. Those microarrays typically contain long probes, covering most or all of the sequence of the target genes. More sophisticated methods involve the synthesis of the probes directly on the array by photolithography (Lipshutz *et al.*, 1999) or by ink-jet technology (Hughes *et al.*, 2001). These methods have the advantage of using short in silico designed oligonucleotide probes that contain information unique to their

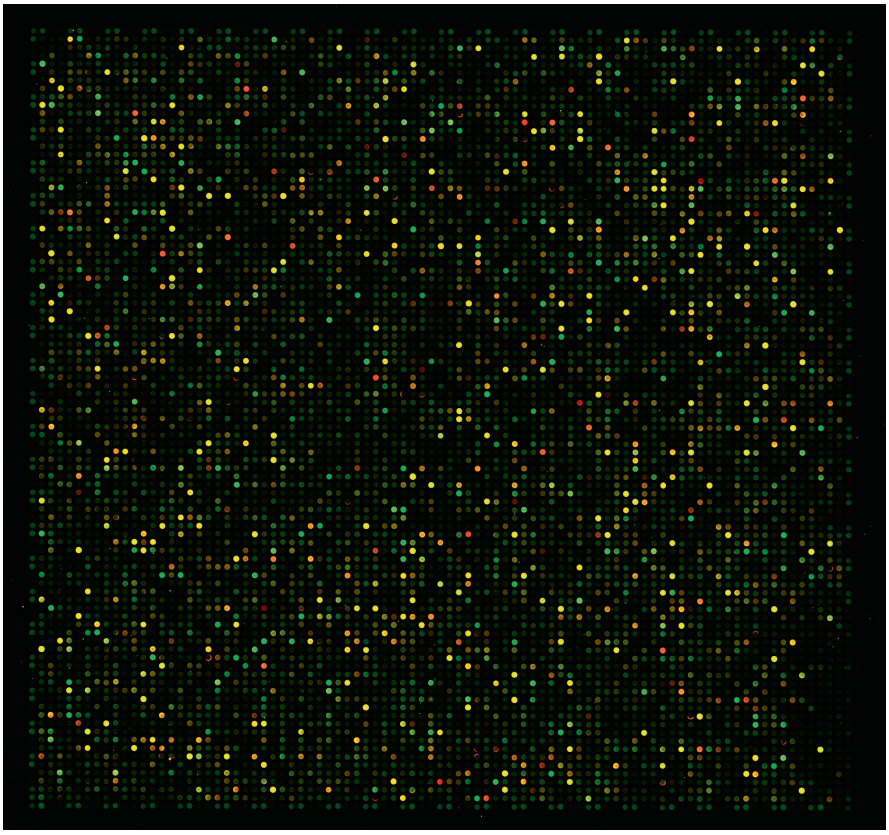

Figure 1. Example of a microarray slide actually used in this thesis project.

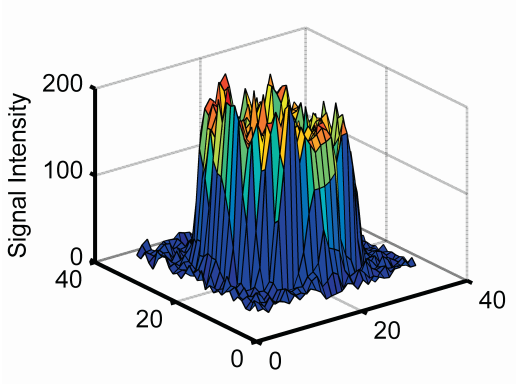

Figure 2. Signal intensity measured at a typical spot on a microarray slide.

target genes, thus minimizing hybridization of mRNA from genes different than their intended targets (Hughes *et al.*, 2001). Ink-jet synthesized microarrays produce very clear images (Figure 1), with well-defined spots that are clearly differentiated from the background noise (Figure 2). Each spot contains several hundreds of specific oligonucleotide probes, which warrants that both green and red targets find ample probes for binding.

WHOLE-GENOME GENE EXPRESSION ANALYSIS BY DNA MICROARRAY

DNA microarray analysis involves several steps, including the choice of microarray platform and microarray design, cell culturing and harvesting, RNA extraction and purification, synthesis of labeled-cDNA, hybridization of cDNA to the DNA microarray, scanning, feature extraction, data normalization and statistical testing. There are many variations for each of these steps, and the characteristics of the results will depend on the decisions taken. Since a major part of this thesis is focused on the bioinformatics part of microarray projects, namely microarray design, feature extraction and data analysis, emphasis on that part will be given in this section.

Microarray design

Microarray design consists of selection of adequate sequences from each single gene for use as probes. One important consideration is whether mRNA will be labeled directly or whether a DNA copy (cDNA) of the mRNA will be synthesized. The most common labeling method for microarray analysis is the synthesis of cDNA with one modified nucleotide containing a fluorescent chromophore for incorporation into the cDNA during reverse transcription. This is the method of choice for the following discussion.

Microarray probes should be chosen to be as specific as possible, in order to avoid confounding signals due to the expression of genes different from the target gene (Li and Stormo, 2001; Talla *et al.*, 2003). Two main approaches exist to assess the potential for hybridization of probes to mRNA from genes different than their targets. One approach consists of determining or controlling the fraction of shared bases between probes and stretches of other genes (Talla *et al.*, 2003), for example, using a Basic Local Alignment Searching Tool (Altschul *et al.*, 1990). A finer approach is based on calculation of the free energy of the hybrid between the probe and non-target genes with similar sequence (Chou *et al.*, 2004). However, while the effects of the number of shared bases between the probes and similar oligonucleotides have been tested experimentally, no such analysis has been reported at the hybridization free energy level. In terms of number of shared bases, sequences with more than 70 to 75 % homology can cause cross-hybridization in oligonucleotide probes of length 50-60 bases under typical hybridization conditions (Kane *et al.*, 2000; Hughes *et al.,* 2001). For longer probes the percentage of homology sufficient for cross-hybridization is probably lower (Talla *et al.*, 2003). The potential for cross-hybridization is greater when the shared bases between the probe and the target are contiguous (Kane *et al.*, 2000), in addition the position of the shared bases along the probe is also important (Hughes *et al.,* 2001).

Another characteristic that is important in microarray probe design is the sensitivity of the probes for their targets. The sensitivity has to be chosen as homogeneous as possible (Relógoio *et al.*, 2002; Nielsen *et al*., 2003), so that the measured signal reflects the mRNA concentration in the samples as close as possible. The stability of the probe/target hybrid determines the sensitivity, and it is thought to depend on their sequence at given hybridization conditions, as is the case with DNA in solution (SantaLucia and Hicks, 2004). The G+C content of the sequence can be used as a rough estimate of hybridization strength between the probe and the target, especially when long probes are used. A finer estimate is attained using nearest neighbor thermodynamic models (SantaLucia and Hicks, 2004; Yakovchuk *et al.*, 2006) to predict the free energy of the hybridization reaction or the melting temperature of the hybrid (the temperature at which half of the DNA is in the hybrid state and half is in the melted state).

In addition to the stability of the probe/target hybrid, another factor that can affect hybridization sensitivity is the potential formation of competing secondary structures in the probes and targets (SantaLucia and Hicks, 2004). Secondary structures arising from self-complementary sub-sequences in a probe or target can compete with the hybridization reaction. While the secondary structures of long targets are difficult to predict, the potential hybridization of a bent probe with itself (hairpin hybridization) or of two contiguous probes in a spot (dimer hybridization) are typically considered to assess whether certain oligonucleotides should be excluded from a microarray design (Relógio *et al.,* 2002; Chou *et al.*, 2004).

A commercial software package for microarray design that takes into account similarities between genes and predicted thermodynamic properties of the hybridization between probes and targets is Array Designer (Premier Biosoft International). Alternative software packages for array design include ProbeSelect (Li and Stormo, 2001), OligoWiz (Nielsen *et al.*, 2003) and Picky (Chou *et al.*, 2004). In some microarray platforms, the complete gene sequence is cloned and spotted into the microarray (for an example, see Postier *et al.*, 2003). In those cases, microarray design does not apply, and the differences in probe sensitivities and specificities will depend only on the sequences of the genes.

The design of microarray oligonucleotide probes is conducted specifically for the microarray platform being used, probe length is a matter of heated debates. Manufacturer Affymetrix advocates use of several relatively short oligonucleotides (15 to 20 bp) for every gene and quantifying the average hybridization efficiency, whereas supplier Agilent advocates usage of a more limited number of longer oligonucleotides (60 basepairs) with the idea that the melting temperature and specificity can be more narrowly defined. Attachment of probes to the slides is done in different ways. Presynthesized probes may be spotted by a robot. Direct synthesis may also proceed in situ with synthesis taking place in defined spots on a solid surface. Two widely used technologies for synthesis directly on the surface are photolithography (Lipshutz *et al.*, 1999) and ink-jet technology (Hughes *et al.*, 2001). Photolithography, introduced by Affymetrix (USA), uses a mask to direct light to spots where nucleotides should not be attached, and leaves the spots that should receive the next nucleotide in turn activated (Lipshutz *et al.*, 1999). Ink-jet technology directs the nucleotides to the spots where they should be incorporated using a specialized printer (Agilent, USA). For work described in this thesis, the Agilent technology has been chosen.

Design of microarray experiments

Decisions regarding the experimental design depend strongly on the research question. It is always important to keep in mind that a sufficient number of biological replicates is essential for attaining meaningful data in DNA microarray analysis. A strict minimum of three independent cultures per treatment ('biological replicates') can be used as a guideline. Two channel green/red screening is an often used approach. Here, cDNA with green labeled nucleotide is synthesized from the mRNA of the control samples, while red labeled nucleotide is used for the treatment sample.
Traditionally, experimental design for microarray analysis has been constrained by the perceived need to compare only samples hybridized in the same microarray (Yang and Speed, 2002; Churchill, 2002; Kerr, 2003). Thus, to compare three samples in a two-channel microarray platform, one had to hybridize each sample twice, with each of the other samples. This apparent need for direct comparison of samples originates from the large variation in the mean intensities of spots between different arrays. Under this paradigm, one compares the logarithmic ratios (log ratios) of the two signals of each array. These effects can be modeled when data are analyzed, which allows for more flexible designs ('t Hoen *et al.* , 2004).

RNA extraction, labeling, hybridization and scanning

There are many different methods for RNA extraction, and the optimal protocol will depend on the organism being studied. The extraction protocols usually involve an isolation step (for example, with hot acid phenol and chloroform), a precipitation step (for example, using lithium chloride or ethanol) and a clean-up step using chromatography. The quality and quantity of the obtained RNA can be tested measuring absorption spectra and a size-sorting method like chromatography. Samples with wide ribosomal RNA size-distribution peaks have probably been subject to RNA degradation. A convenient instrument for generating absorption spectra of microliter volumes is Nanodrop (ND-1000, Nanodrop Technologies, USA). Using a Bioanalyzer instrument (Agilent 2100 bioanalyzer, Agilent Technologies, USA) the size distribution of an RNA sample can be analyzed to approve or disapprove its integrity. There are many protocols for the synthesis of labeled cDNA from RNA. Our protocol was based on guidelines of the Agilent company.

The green and red labeled material is exposed to a microarray slide and placed in a holder that permits optimal mixing in a temperature controlled incubator. Hybridization times are typically around 15 h, and the hybridization temperature is usually around 60°C. The high temperature during hybridization increases the stringency of the hybridization reaction, thus reducing non-specific hybridization.

After hybridization, the microarray slides are typically washed and let air dry, a fixation may be used. The slides are read with specialized scanners that usually have one laser beam for each color of dye used (Yang *et al.*, 2002). The typical output of the scanning process is one 16-bit tagged image file format (TIFF) file for each channel (red, green) on the microarray (Yang *et al.*, 2002).

Extracting data from microarray images

Once the scanned images have been generated, the intensity of each spot for each of the two color channels (green and red) has to be listed as a representative value, such as the mean or the median. The process of obtaining representative numerical values for the signal at each spot is known as data extraction. Feature Extraction Software (Agilent Technologies, USA) is an automated option for extracting Agilent microarray images. The extracted data normally include the mean and median intensity at each spot, a mean background signal measurement surrounding the spot and several quality measurements like the signal and background standard deviation and the presence of outlying pixels (Yang *et al.*, 2002).

Data normalization

After the data are extracted, they have to be normalized to remove artifacts introduced at various stages of the process. If the overall signal intensities are weak and they are correlated with the background signal, a background subtraction is necessary (Kim *et al.*, 2002). If, on the other hand, the background is uniform and the signal is well above the background, background subtraction is unnecessary, and even contra-productive as it could introduce noise to the data (Wernisch *et al.*, 2003; Thygesen and Zwinderman, 2004).

Several microarray normalization strategies are available, including some platform-dependent variations. Spotted microarrays may use different print-tips to produce different areas of the array simultaneously. In those cases, normalization for the mean effect of each print-tip is necessary (Dudoit *et al.*, 2002; Smyth and Speed, 2003). Some experiments include RNA-sets of defined sequences and concentrations for cDNA synthesis such that internal calibration of the hybridization mix and hybridization to the counterpart spots in the array can be used for normalization. This method is known as spike-in normalization (Badiee *et al.*, 2003). A related idea is the use of housekeeping genes as internal references for normalization (Khimani *et al.*, 2005). Housekeeping genes are genes that are supposed to have a constant expression level in all growth conditions comprising the experiment. The stability of the gene expression of a housekeeping gene set needs to be validated at the relevant experimental conditions with independent methods (Khimani *et al.*, 2005; Jain *et al.*, 2006), for which real-time quantitative polymerase chain reaction (RT-qPCR) is often used. Alternatively, this normalization technique can be useful if the research question involves the expression changes of the genome relative to the expression changes of a group of genes.

A very common practice, known as within-array normalization, is to normalize each microarray separately focusing on the differences between the two channels. This normalization scheme is directed towards removing the effects of using different dyes, i.e differences in dye incorporation efficiency and different quantum yields of fluorescence emission. This practice can be sufficient when used with a connected experimental design (see above), as has been proposed (Yang and Speed, 2002; Churchill, 2002; Kerr, 2003). A simple method for within-array normalization is to subtract the difference of the mean log ratios of the two channel's signals from one of them. A more refined method that removes intensity-dependent dye effects is to fit a locally weighted regression line (lowess) to the log ratios as a function of intensity and subtract that line from the log ratios (Dudoit *et al.*, 2002; Quackenbush, 2002).

Between-array normalization is aimed at normalizing different arrays of one experiment. Quantile-based normalization methods equalize the distributions of all the data-series of one experiment (Workman *et al.*, 2002). Quantile-based normalization methods assume that the experimental treatments do not affect the distribution of gene expression levels significantly. In other words, these methods assume that the amount of under-expression and over-expression is comparable. Analysis of Variance (ANOVA) based normalization methods explicitly model several experimental artifacts as linear effects on the measured intensities on microarrays (Kerr and Churchill, 2001; Wolfinger *et al.*, 2001). Typical effects included for normalization in ANOVA models are the main effect of each of the arrays in the experiment, the main effect of the different dyes and the array-specific effect of the dyes (Kerr and Churchill, 2001; Wolfinger *et al.*, 2001). The effect of different print-tips can be included when it applies. Additionally to the normalization terms, ANOVA-based microarray analysis methods can include the effects of genes, treatments and their interactions in the model (Kerr and Churchill, 2001). In that case, the statistical testing and the normalization are done simultaneously. As an alternative, the procedure can be split into a normalization model and a gene-by-gene model for statistical testing (Wolfinger *et al.*, 2001). As opposed to quantile-based methods, ANOVA-based normalization methods do not make assumptions about the distribution of gene expression in the samples. However, it does depend on the assumption that the measuring error is normally distributed.

Variance stabilization is another normalization method worth considering (Huber *et al.*, 2002). This method substitutes the use of logarithms for an inverse hyperbolic sine transformation tuned for the variance to become independent of the intensity. The procedure results in transformed data, in which differences in expression (equivalent to the log ratio) are independent of the intensity.

Statistical testing and the multiple testing problem

To evaluate the effects of the applied treatments on gene expression, one needs to perform statistical testing on the normalized data (Dudoit *et al.*, 2002). For this purpose, a microarray experiment can be viewed as a set of many hypotheses, one hypothesis for each gene represented in the microarray. Statistical tests are applied to assign a probability (p-value) to the outcome of each of those hypotheses. The p-values are then used to assess whether the genes are significantly differentially expressed (see below).

The correct statistical analysis method depends on the experimental design and on the microarray platform. Commonly used statistics for microarray experiments are the t-statistic (Dudoit *et al.*, 2002) and the F-statistic obtained by ANOVA (Kerr, 2003). The choice of parametric or non-parametric statistics depends on how reasonable the assumptions of homoscedasticity and normality are. The number of replicates in microarray experiments is usually too low to test for the validity of those assumptions. One alternative to avoid distributional assumptions with parametric statistics is to determine the p-values by permutation (Dudoit *et al.*, 2002). The basic idea behind determining p-values by permutation is to count the number of times a hypothesis is more significant than a random set with the same distribution. The random set is generated by randomly shuffling the labels (treatment *vs.* control) of the data series.

Statistical testing of microarray data generates one p-value per gene, yielding thousands of p-values per experiment. These p-values are estimates of the probability that a given hypothesis is truly null (i.e., that the gene is not differentially expressed). Small p-values indicate a low probability of the hypothesis being truly null, and therefore a high probability that the gene is truly differentially expressed. However, since a large number of hypotheses are tested, the number of genes with a p-value lower than an arbitrary threshold can be very large. This problem is referred to as multiple testing (Dudoit *et al*., 2002; Storey and Tibshirany, 2003). There are many methods for addressing the multiple testing problem, including the family-wise error rate (FWER) and the false discovery rate (FDR). The FWER is the probability of declaring at least one truly null hypothesis significant and the FDR is the expected proportion of truly null hypotheses declared significant (Dudoit *et al.,* 2003). The FDR is more adequate for microarray experiments where one identifies a large number of differentially expressed genes, and a small proportion of false positives are acceptable. Most methods for controlling for multiple testing are based on creating a data set with the same distribution as the original data, but where all the hypothesis are truly null (Dudoit *et al.*, 2003). Storey and Tibshirany (2003) introduced a new FDR estimation method that is now widely used. It does not depend on permutation, but compares the observed p-value distribution with the expected uniform p-value distribution under the complete null hypothesis.

CYANOBACTERIAL GENOMICS AND STRESS PHYSIOLOGY

Cyanobacteria are oxygenic photosynthetic prokaryotes that occur in many ecosystems throughout the world. They are important contributors to the global primary production of our planet, notably by their widespread dominance in oceanic ecosystems (Partensky *et al.*, 1999). Moreover, due to the ability of some cyanobacteria to fix nitrogen, they may exert control over the nitrogen to phosphorus ratio of oceanic ecosystems (Tyrrell, 1999), provided that nitrogen fixation is not limited by iron (Falkowski, 2000). The ability of cyanobacteria to change their environment is widely recognized, since they are thought to be responsible for the accumulation of oxygen in the Earth's atmosphere, in the Paleoproterozic era. Besides their ecological relevance, they are believed to have important biotechnological potential, even as hydrogen producers for energy production (Tamagnini *et al.*, 2002; Burja *et al.*, 2003).

Cyanobacteria have attracted a considerable genome sequencing effort. One of the first whole-genome sequences completed was the one of *Synechocystis* PCC 6803 (Kaneko *et al*., 1996). At present, about 55 cyanobacterial genome-sequence projects are finished or close to being finished (Bryant and Frigaard, 2006).

The availability of whole-genome sequences of cyanobacteria has led to the use of microarrays for the study of cyanobacterial global gene expression. Microarray cyanobacterial studies have been conducted using spotted microarrays (Hihara *et al.*, 2001; Postier *et al*., 2003; Stowe-Evans *et al.*, 2004; Kucho *et al.*, 2004) and short oligonucleotide microarrays (Steglich *et al.*, 2006). Cyanobacterial microarray studies have been targeted at the study of gene expression under different environmental and/or mutational perturbations. Genome wide gene expression of *Synechocystis* PCC 6803 has been studied under high light stress (Allakhverdiev *et al.*, 2002; Hihara *et al.*, 2001), cold stress (Allakhverdiev *et*

al., 2002; Suzuki *et al.*, 2001; Mikami *et al.*, 2002; Inaba *et al.*, 2003), heat shock (Suzuki *et al.*, 2005; Suzuki *et al*., 2006), salt and osmotic stress (Allakhverdiev *et al.*, 2002; Mikami *et al.*, 2002; Kanesaki *et al.*, 2002; Marin *et al.*, 2003; Postier *et al*., 2003; Marin *et al.*, 2004, Asadulghani *et al.*, 2004; Paithoonrangsarid *et al.*, 2004; Shoumskaya *et al.*, 2005), metal starvation (Yamaguchi *et al.*, 2002; Singh *et al.*, 2003), darkness (Osanai *et al*., 2005), heterotrophic growth (Kahlon *et al.*, 2006) and short-term nitrogen starvation (Osanai *et al.*, 2006). Furthermore, gene expression of a variety of mutants of *Synechocystis* PCC 6803 have been studied by microarray analysis, particularly knock-out mutants of regulatory proteins, including two-component systems (Yamaguchi *et al.*, 2002; Mikami *et al.*, 2002; Inaba *et al.*, 2003; Marin *et al.*, 2003; Paithoonrangsarid *et al.*, 2004; Shoumskaya *et al.*, 2005; Suzuki *et al.*, 2005; Kahlon *et al.*, 2006; Panichkin *et al.*, 2006), transcriptional regulators (Fujimori *et al.*, 2005) and RNA polymerase sigma factors (Osanai *et al*., 2005). The effects of other knock-out mutants on whole-genome gene expression have been analyzed, including mutations of the oxygen-evolving complex of photosystem II (Schriek *et al.*, 2008), mutations of heat shock proteins (Asadulghani, 2004), mutations that affect the rigidity of the membranes (Inaba *et al.*, 2003) and the response to oxidative stress (Kobayashi *et al*., 2004; Li *et al.*, 2004) and the iron-stress inducible protein isiA (Singh *et al.*, 2005). Other studies used inhibitors of the photosynthetic electron transport chain to analyze the effect of different membrane redox potentials (Hihara *et al.*, 2003) and oxidative stress (Kobayashi *et al.*, 2004) on global gene expression of *Synechocystis* PCC 6803. Oxidative stress has also been induced with water peroxide for microarray analysis (Li *et al*., 2004; Singh *et al.*, 2005).

OUTLINE OF THIS THESIS

This thesis work presents the design and implementation of oligonucleotide microarrays for global gene expression analysis of *Synechocystis* PCC 6803. The thesis includes the development of new statistical methods that improve the analysis of microarray data. The microarray designed in this study was used to analyze changes in gene expression in response to nitrogen limitation, light limitation and inorganic carbon limitation.

The implementation of oligonucleotide microarray technology to a sequenced species requires a microarray design. The design consists of choosing unique probes from the sequences of all genes in the genome. In this study, the focus was on selection of oligonucleotides with a length of 60 bases. We carefully selected suitable probes using the thermodynamic properties of probe-target hybridization as the selection criterium. In **chapter 2** our strategy for microarray design is described and the impact of the design parameters on the observed hybridization signals is explored.

Microarray signals are affected by many artifacts, including differences in dye incorporation and quantum yield, differences in the RNA quantity and quality of the arrayed samples, etc. Therefore, normalization is essential for the proper analysis and interpretation of microarray data. In **chapter 3** we introduce a new parametric normalization method. The method models the signal distributions of data sets with the generalized extreme value (GEV) distribution, and subsequently normalizes these data sets to the same GEV distribution using the estimated distributional parameters. A key advantage of this parametric approach is that it preserves the internal structure of the

original data, which makes it especially suitable for the detection of subtle changes in gene expression.

In **chapter 4**, the microarray designed and presented in chapter 2 was used to analyze the effect of 12 h nitrogen starvation in batch cultures of *Synechocystis* PCC 6803. The microarray data were based on a high-quality experiment with 6 independent biological replicates for the control and the treatment conditions. The experiment includes dye-swaps and a self-self hybridization to control for technical error.

In **chapter 5**, a time series of physiological and gene expression responses of *Synechocystis* PCC 6803 was studied in continuous culture. The experiment monitored four replicate chemostat cultures while they were shifted from a nitrogen-limited steady state to a light-limited steady state, and then back to a nitrogen-limited steady state again. Marked changes in photosynthesis, pigmentation and growth were observed during these experiments. Changes in gene expression patterns clearly related to the experimental treatment were observed for hundreds of genes.

Growth of cyanobacteria can become limited by the amount of available inorganic carbon in aquatic ecosystems. In **chapter 6**, we investigate concerted changes in physiological traits and gene expression that result from inorganic carbon limitation (Eisenhut *et al.*, 2007). Microarray analyses indicated stable up-regulation of genes for inducible uptake systems of carbon dioxide and bicarbonate. Furthermore, we found up-regulation of several photosystem I genes as well as a higher photosystem I content and activity. Surprisingly, down-regulation was observed for almost all carboxysomal proteins.

Chapter 7 presents a general discussion of the results obtained in the previous chapters. The relation between microarray design and microarray results is considered. The advantages and disadvantages of different normalization methods are discussed, with particular emphasis on the new normalization method introduced in this thesis. Also, a comparison between microarray results obtained from continuous culture versus batch culture is presented. Finally, the effects of different strategies for the experimental controls are discussed.

REFERENCES

Allakhverdiev S.I., Nishiyama Y., Miyairi S., Yamamoto H., Inagaki N., Kanesaki Y., Murata N. 2002. Salt stress inhibits the repair of photodamaged photosystem II by suppressing the transcription and translation of psbA genes in *Synechocystis*. Plant Physiology 130:1443-1453.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

Asadulghani S.Y., Nitta K., Kaneko Y., Kojima K., Fukuzawa H., Kosaka H., Nakamoto H. 2004. Comparative analysis of the hspA mutant and wild-type *Synechocystis* sp. strain PCC 6803 under salt stress: evaluation of the role of hspA in salt-stress management. Archives of Microbiology 182:487-497.

Badiee A., Eiken H.G., Steen V.M., Løvlie R. 2003. Evaluation of five different cDNA labeling methods for microarrays using spike controls. BMC Biotechnology 3:23.

Bansal A.K., Mayer T.E. 2002. Evolutionary analysis by whole-genome comparisons. Journal of Bacteriology 184:2260-2272.

Boore J.F. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology and Evolution 21:439-446.

Bryant D.A., Frigaard N.U. 2006. Prokaryotic photosynthesis and phototrophy illuminated. Trends in Microbiology 14:488-496.

Chatterjee A., Majee M., Ghosh S., Majumder A.L. 2004. Sll1722, an unassigned open reading frame of *Synechocystis* PCC 6803, codes for L-myo-inositol 1-phosphate synthase. Planta 218:989-998.

Chou H.H, Hsia A.P, Mooney D.L, Schnable P.S. 2004. Picky: oligo microarray design for large genomes. Bioinformatics 20:2893-2902.

Churchill G.A. 2002. Fundamentals of experimental design for cDNA microarrays. Nature Genetics 32:490-495.

Conway T., Schoolnik G.K. 2003. Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Molecular Microbiology 47:879-889.

Dudoit S., Yang Y.H., Callow M.J, Speed T.P. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12:111-139.

Dudoit S., Shaffer J.P, Boldrick J.C. 2003. Multiple hypothesis testing in microarray experiments. Statistical Science 18:71-103.

Eisen J.A. 2000. Assessing evolutionary relationships among microbes from whole-genome analysis. Current Opinion in Microbiology 3:475-480.

Eisenhut M, Aguirre von Wobeser E, Jonas L, Schubert HJ, Ibelings BW, Bauwe H, Matthijs HCP, Hagemann M. 2007. Long-term response toward inorganic carbon limitation in wild type and glycolate turnover mutants of the cyanobacterium *Synechocystis* sp. strain PCC 6803. Plant Physiology 144:1946-1959.

Falkowski P.G. 2000. Rationalizing elemental ratios in unicellular algae. Journal of Phycology 36:3-6.

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Doughetry B.A., Merrick J.M., McKenney K., Sutton G., FitzHugh W., Fields C., Gocayne J.D., Scott J., Shirley R., Liu L.I., Glodek A., Kelley J.M., Weidman J.F., Phillips C.A. Spriggs T., Hedblom E., Cotton M.D., Utterback T.R., Hanna M.C., Nguyen D.T., Saudek D.M., Brandon R.C., Fine L.D., Fritchman J.L., Fuhrmann J.L., Geoghangen N.S.M., Gnehm C.L., McDonnald L.A., Small K.V., Fraser C.M., Smith H.O., Venter J.C. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-498 + 507-512.

Fujimori T., Higuchi M., Sato H., Aiba H., Muramatsu M., Hihara Y., Sonoike K. 2005. The mutant of sll1961, which encodes a putative transcriptional regulator, has a defect in regulation of photosystem stoichiometry in the cyanobacterium *Synechocystis* sp. PCC 6803. Plant Physiology 139:408-416.

Fulda S., Huang F., Nilsson F., Hagemann M., Norling B. 2000. Proteomics of *Synechocystis* sp. strain PCC 6803: identification of periplasmic proteins in cells grown at low and high salt. European Journal of Biochemistry 267:5900-5907.

Herranen M., Battchikova N., Zhang P., Graf A., Sirpio S. , Paakkarinen V.,
Aro E.M. 2004.Towards functional proteomics of membrane protein complexes in *Synechocystis* sp. PCC 6803. Plant Physiology 134:470–481.

Hihara Y., Kamei A., Kanehisa M., Kaplan A., Ikeuchi M. 2001. DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. Plant Cell 13:793-806.

Hihara Y., Sonoike K., Kanehisa M., Ikeuchi M. 2003. DNA microarray analysis of redox-responsive genes in the genome of the cyanobacterium *Synechocystis* sp. strain PCC 6803. Journal of Bacteriology 185:1719-1725.

‘t Hoen P.A.C., Turk R., Boer J.M., Sterrenburg E., de Menezes R.X., van Ommen G.J.B., Dunnen J.T. 2004. Intensity-based analysis of two-colour microarrays enables efficient and flexible hybridization designs. Nucleic Acids Research 32(4):e41.

Huber W., von Heydebreck A., Sültmann H., Poustka A., Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18:S96-S104.

Hughes T.R., Mao M., Jones A.R., Burchard J., Marton M.J., Shannon K.W., Lefkowitz S.M., Ziman M., Schelter J.M., Meyer M.R., Kobayashi S., Davis C., Dai H., He Y.D., Stephaniants S.B., Cavet G., Walker W.L., West A., Coffey E., Shoemaker D.D., Stoughton R., Blanchard A.P., Friend S.H., Linsley P.S. 2001. Expression profiling using microarrays fabricated by ink-jet oligonucleotide synthesizer. Nature Biotechnology 19:342-347.

Inaba M., Suzuki I., Szalontai B., Kanesaki Y., Los D.A., Hayashi H., Murata N. 2003. Gene-engineered rigidification of membrane lipids enhances the cold inducibility of gene expression in *Synechocystis*. Journal of Biological Chemistry 278:12191-12198.

Kahlon S., Beeri K., Ohkawa H., Hihara Y., Murik O., Suzuki I., Ogawa T., Kaplan A. 2006. A putative sensor kinase, Hik31, is involved in the response of *Synechocystis* sp. strain PCC 6803 to the presence of glucose. Microbiology 152:647-655.

Kane M.D., Jatkoe T.A., Stumpf C.R., Thomas J.D., Madore S.J. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50-mer) microarrays. Nucleic Acids Research 28:4552-4557.

Kaneko T., Sato S., Kotani H., Tanaka A., Asamizu E., Nakamura Y., Miyajima N., Hirosawa M., Suqiura 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 Research 3:109-136.

Kanesaki Y., Suzuki I., Allakhverdiev S.I., Mikami K., Murata N. 2002. Salt stress and hyperosmotic stress regulate the expression of different sets of genes in *Synechocystis* sp. PCC 6803. Biochemical and Biophysical Research Communications 290:339-348.

Kerr K. 2003. Design considerations for efficient and effective microarray studies. Biometrics 59:822-828.

Khimani A.H., Mhashilkar A.M., Mikulskis A., O'Malley M., Liao J., Golenko E.E., Mayer P., Chada S., Killian J.B., Lott S.T. 2005. Housekeeping genes in cancer: normalization of array data. BioTechniques 38:739-745.

Kim J.H., Shin D.M., Lee Y.S. 2002. Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles. Experimental and Molecular Medicine 34:224-232.

Kobayashi M, Ishizuka T, Katayama M, Kanehisa M, Bhattacharyya-Pakrasi M, Pakrasi HB, Ikeuchi M. 2004. Response to oxidative stress involves a novel peroxiredoxin gene in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803. Plant and Cell Physiology 45:290-299.

Kucho K., Tsuchiya Y., Okumoto Y., Harada M., Yamada M., Ishiura M. 2004. Construction of unmodified oligonucleotide-based microarrays in the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1: screening of the candidates for circadianly expressed genes. Genes Genetic Systems 79:319-329.

Li F., Stormo G.D. 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics 17:1067-1076.

Li H., Singh A.K., McIntyre L.M., Sherman L.A. 2004. Differential gene expression in response to hydrogen peroxide and the putative PerR regulon of *Synechocystis* sp. strain PCC 6803. Journal of Bacteriology 186:3331-3345.

Lipshutz R.J., Fodor S.P.A., Gingeras T.R., Lockhart D.J. 1999. High density synthetic oligonucleotide arrays. Nature Genetics Supplement 21:20-24.

Marin K., Suzuki I., Yamaguchi K., Ribbeck K., Yamamoto H., Kanesaki Y., Hagemann M., Murata N. 2003. Identification of histidine kinases that act as sensors in the perception of salt stress in *Synechocystis* sp. PCC 6803. Proceedings of the National Academy of Sciences USA 100:9061-9066.

Marin K., Kanesaki Y., Los D.A., Murata N., Suzuki I., Hagemann M. 2004. Gene expression profiling reflects physiological processes in salt acclimation of *Synechocystis* sp. strain PCC 6803. Plant Physiology 136:3290-3300.

Martin K.A., Siefert J.L., Yarrapragada S., Lu Y., McNeill T.Z., Moreno P.A., Weinstock G.M., Widger W.R., Fox G.E. 2003. Cyanobacterial signature genes. Photosynthesis Research 75:211-221.

Mikami K., Kanesaki Y., Suzuki I., Murata N. 2002. The histidine kinase Hik33 perceives osmotic stress and cold stress in *Synechocystis* sp PCC 6803. Molecular Microbiology 46:905-915.

Nielsen H.B., Wernersson R., Knudsen S. 2003. Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. Nucleic Acids Research 31:3491–3496.

Osanai T., Kanesaki Y., Nakano T., Takahashi H., Asayama M., Shirai M., Kanehisa M., Suzuki I., Murata N., Tanaka K. 2005. Positive regulation of sugar catabolic pathways in the cyanobacterium *Synechocystis* sp. PCC 6803 by the group 2 sigma factor sigE. Journal of Biological Chemistry 280:30653-30659.

Osanai T., Imamura S., Asayama M., Shirai M., Suzuki I., Murata N., Tanaka K. 2006. Nitrogen induction of sugar catabolic gene expression in *Synechocystis* sp. PCC 6803. DNA Research 13:185-195.

Paithoonrangsarid K., Shoumskaya M.A., Kanesaki Y., Satoh S., Tabata S., Los D.A., Zinchenko V.V., Hayashi H., Tanticharoen M., Suzuki I., Murata N. 2004. Five histidine kinases perceive osmotic stress and regulate distinct sets of genes in *Synechocystis*. Journal of Biological Chemistry 279:53078-53086.

Panichkin V.B., Arakawa-Kobayashi S., Kanaseki T., Suzuki I., Los D.A., Shestakov S.V., Murata N. 2006. Serine/threonine protein kinase SpkA in *Synechocystis* sp. strain PCC 6803 is a regulator of expression of three putative pilA operons, formation of thick pili, and cell motility. Journal of Bacteriology 188:7696-7699.

Partensky F., Hess W.R., Vaulot D. 1999. *Prochlorococcus*: a marine photosynthetic prokaryote of global significance. Microbiology and Molecular Biology Reviews 63:106-127.

Postier B.L, Wang H.L., Singh A., Impson L., Andrews H.L, Klahn J., Li H., Risinger G., Pesta D., Deyholos M., Galbraith D.W., Sherman L.A., Burnap R.L. 2003. The construction and use of bacterial DNA microarrays based on an optimized two-stage PCR strategy. BMC Genomics 4:23.

Quackenbush J. 2002. Microarray data normalization and transformation. Nature Genetics 32:496-501.

Relógio A., Schwager C., Richter A., Ansorge W., Valcárcel J. 2002. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Research 30:e51.

SantaLucia J., Hicks D. 2004. The thermodynamics of DNA structural motifs. Annual Review of Biophysics and Biomolecular Structure 33:415-440.

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

Schriek S., Aguirre von Wobeser E., Nodop A., Becker A., Ibelings B.W., Bok J., Staiger D., Matthijs H.C.P., Pistorius E.K., Michel K.P. 2008. Transcript profiling indicates that the absence of PsbO affects the coordination of C and N metabolism in *Synechocystis* sp PCC 6803. Physiologia Plantarum 133:525-543.

Shoumskaya M.A., Paithoonrangsarid K., Kanesaki Y., Los D.A., Zinchenko V.V., Tanticharoen M., Suzuki I., Murata N. 2005. Identical Hik-Rre systems are involved in perception and transduction of salt signals and hyperosmotic signals but regulate the expression of individual genes to different extents in *Synechocystis*. Journal of Biological Chemistry 280:21531-21538.

Simon W.J., Hall J.J., Suzuki I., Murata N., Slabas A.R. 2002. Proteomic study of the soluble proteins from the unicellular cyanobacterium *Synechocystis sp.* PCC 6803 using automated matrix-assisted laser desorption/ionization-time of flight peptide mass fingerprinting. Proteomics 2:1735-1742.

Singh A.K., McIntyre L.M., Sherman L.A. 2003. Microarray analysis of the genome-wide response to iron deficiency and iron reconstitution in the cyanobacterium *Synechocystis* sp. PCC 6803. Plant Physiology 132:1825-1839.

Singh A.K., Li H., Bono L., Sherman L.A. 2005. Novel adaptive responses revealed by transcription profiling of a *Synechocystis* sp. PCC 6803 delta-isiA mutant in the presence and absence of hydrogen peroxide. Photosynthesis Research 84:65-70.

Smyth G.K., Speed T. 2003. Normalization of cDNA microarray data. Methods 31:265–273.

Steglich C., Futschik M., Rector T., Steen R., Chisholm S.W. 2006. Genome-wide analysis of light sensing in *Prochlorococcus*. Journal of Bacteriology 188:7796-7806.

Storey J.D., Tibshirani R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences USA 100:9440-9445.

Stowe-Evans E.L., Ford J., Kehoe D.M. 2004. Genomic DNA microarray analysis: identification of new genes regulated by light color in the cyanobacterium *Fremyella diplosiphon*. Journal of Bacteriology 186:4338-4349.

Suzuki I., Kanesaki Y., Mikami K., Kanehisa M., Murata N. 2001. Cold-regulated genes under control of the cold sensor Hik33 in *Synechocystis*. Molecular Microbiology 40:235-244.

Suzuki I., Kanesaki Y., Hayashi H., Hall J.J., Simon W.J., Slabas A.R., Murata N. 2005. The histidine kinase Hik34 is involved in thermotolerance by regulating the expression of heat shock genes in *Synechocystis*. Plant Physiology 138:1409-1421.

Suzuki I., Simon W.J., Slabas A.R. 2006. The heat shock response of *Synechocystis* sp. PCC 6803 analysed by transcriptomics and proteomics. Journal of Experimental Botany 57:1573-1578.

Talla E., Tekaia F., Brino L., Dujon B. 2003. A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization. BMC Genomics 4:38.

Tamagnini P., Axelsson R., Lindberg P., Oxelfelt F., Wünschiers R., Lindblad P. 2002. Hydrogenases and hydrogen metabolism of cyanobacteria. Microbiology and Molecular Biology Reviews 66:1-20.

Thygesen H.H., Zwinderman A.H. 2004. Comparing transformation methods for DNA microarray data. BMC Bioinformatics 5:77.

Tyrrell T. 1999. The relative influences of nitrogen and phosphorus on oceanic primary production. Nature 400:525-531.

Wernisch L., Kendall S.L., Soneji S., Wietzorrek A., Parish T., Hinds J., Butcher P.D., Stoker N. 2003. Analysis of whole-genome microarray replicates using mixed models. Bioinformatics 19:53-61.

Workman C., Jensen L.J., Jarmer H., Berka R., Gautier L., Nielser H.B., Saxild H.H., Nielsen C., Brunak S., Knudsen S. 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology 3(9): research 0048.

Yakovchuk P., Protozanova E., Frank-Kamenetskii D.F. 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Research 34:564-574.

Yamaguchi K., Suzuki I., Yamamoto H., Lyukevich A., Bodrova I., Los D.A., Piven I., Zinchenko V., Kanehisa M., Murata N. 2002. A two-component Mn^{2+}-sensing system negatively regulates expression of the mntCAB operon in *Synechocystis*. Plant Cell 14:2901-2913.

Yang Y.H., Speed T. 2002. Design issues for cDNA microarray experiments. Nature Reviews Genetics 3: 579-588.

Yang Y.H., Buckley M.J., Dudoit S., Speed T.P. 2002. Comparison of methods for image analysis on cDNA microarray data. Journal of Computational and Graphical Statistics 11:108-136.