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

Aguirre von Wobeser, E.

Publication date
2010

Chapter 2

Effects of design parameters on the hybridization intensity of a *Synechocystis* PCC 6803 microarray

Eneas Aguirre von Wobeser[1], Bas W. Ibelings[2], Jef Huisman[1] & Hans C.P. Matthijs[1,3]

[1]*Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands*

[2]*Netherlands Institute of Ecology (NIOO-KNAW), Centre for Limnology, Rijksstraatweg 6, 3631 AC Nieuwersluis, The Netherlands.*

[3]*Corresponding author: Hans Matthijs, phone: +31 20 5257070, fax: +31 20 5257064, Email: J.C.P.Matthijs@uva.nl*

ABSTRACT

The hybridization efficiency of probes and their targets on DNA microarrays may differ for each probe-target pair. Here, we have studied how criteria set for the thermodynamic properties in probe selection may affect the intensity of the signals in a whole genome DNA microarray with thousands of probe-target pairs. Probes for the cyanobacterium *Synechocystis* PCC 6803 were designed by a systematic search strategy aiming at high-quality probes with comparable hybridization efficiencies. Effects of probe design parameters on hybridization intensities were tested, with control cells and cells subjected to salt stress as experimental system. We show that, even after normalization of the data, all design parameters had highly significant effects on hybridization intensity. Melting temperature and homology with other genes increased the hybridization signal, whereas probe length, hairpin and dimer structures, sequence length of genes, and the position of probes within genes significantly decreased the signal. The log ratio between treatment and control cells was also significantly affected by thermodynamic properties of the probes, although effects on the log ratio were relatively small. We conclude that a careful probe design strategy should aim at probes with similar thermodynamic properties across the entire microarray.

KEY WORDS. Microarray design, hybridization, thermodynamics, melting temperature, hairpin, dimer, homology, probe length, *Synechocystis* PCC 6803, salt stress.

ABBREVIATIONS. Tm - melting temperature; ΔG - Gibbs free energy, NCBI - National Center for Biotechnology Information, BLAST - basic local alignment search tool.

INTRODUCTION

Living cells modulate the expression of their genes to cope with challenging environmental conditions. Knowledge of the mechanisms of gene expression regulation in model organisms is of great interest for the understanding of these acclimation responses. Once gene expression patterns can be related to certain stress conditions, they could be utilized for the detection of those stress conditions in environmental samples. Thus, the study of gene expression responses to varying growth conditions has great potential for use in basic and applied research. Traditionally, gene expression studies have focused on measuring mRNA concentrations of single genes, suspected to be involved in the response to a given stress, based on laborious previous research. Recently, microarray technology has been developed to measure mRNA concentrations of thousands of genes simultaneously, allowing the detection of genome-wide expression and therewith identification of genes that were previously unknown to be responsive to the studied environmental conditions. Whole-genome DNA microarrays require that the sequence of all genes is known. For each gene, one or more carefully selected probes are obtained by PCR amplification or by chemical synthesis of oligonucleotides. A probe is a copy of all or part of the sequence of the gene it represents. A full genome requires thousands of unique probes for coverage of all genes. The single stranded probes are arranged in an ordered grid of spots on a glass slide. A probe on the slide hybridizes to the target DNA, which requires that the probe DNA is complementary to the sequence of its target.

A typical microarray experiment comprises extraction of RNA from an organism, and synthesis of cDNA by reverse transcription, using the RNA pool as template. Colour-tagged nucleotides are incorporated in the mix of nucleotides available for reverse transcription. The use of red and green colour-tags allows colour-coded labeling of cDNA obtained from control and treatment conditions, respectively. The labelled cDNA targets are exposed for hybridization to the probes on the microarray. The microarray slide is then scanned to detect the colour of the tag used. The more cDNA binds to each probe, the stronger the colour of the corresponding spot. Therefore, the colour signal detected by the scanner reflects the cDNA concentration, which in turn depends on the mRNA concentration in the original sample. However, the relation between the mRNA concentration of each gene and the signal obtained is not the same for each probe, since the hybridization reaction is influenced by many known and unknown factors (Kerr and Churchill, 2001), some of which are probe-specific.

While spotted microarrays typically use quite large gene sequences for specificity, oligonucleotide microarrays have the advantage of *in silico* selection of shorter sequences. Oligonucleotide probes are designed to favour hybridization of the probes to their targets (Nielsen *et al.*, 2003) and to avoid interference from other genes. If an oligonucleotide probe for one gene is too similar to the sequence of another gene, the measured signal could be influenced by the expression levels of both genes. To avoid this cross-hybridization, unique sequences of each gene are considered for microarray design. Oligonucleotide microarray design focuses therefore on sequence uniqueness of probes towards their target gene. Another important aspect is to design probes with similar thermodynamic properties (Matveeva *et al.*, 2003). However, the choice of probe sequences is limited by the sequence of each gene, and the intended thermodynamic properties are accordingly met at a varying degree for the different genes. Therefore, microarray designs generally consist of probes with varying degrees of quality.

Some probe-design programs use a Basic Local Alignment Search Tool (BLAST; Altschul *et al.*, 1997) to exclude parts of a gene with significant similarity to other parts of the genome. While BLAST results are useful for microarray design, better predictors of cross hybridization include the number of shared bases between the probe and the cross-hybridizing gene and the length of the longest contiguous perfect match (Hughes *et al.*, 2001; Talla *et al.*, 2003).

Hybridization of nucleic acid pairs in solution is affected by thermodynamic properties dependent on their nucleotide sequence, and can be modeled with high precision (SantaLucia and Hicks, 2004). However, similar systematic studies of the thermodynamics of hybridization of labelled nucleic acid targets to surface-fixed probes are lacking. Therefore, models derived for systems in solution (SantaLucia, 1998) are commonly used to predict the probe-target hybridization strength on microarrays (Premier Biosoft International, 2002). Earlier studies have verified that these models successfully predict hybridization of known concentrations of nucleic acids to surface-attached probes (Wu and Irizarry, 2004). The effect of the surface has been studied to some extent using Langmuir adsorption models from surface chemistry (Hekstra *et al.*, 2003) and position dependent weights on the contribution of each nucleotide to the free energy of the hybrid (Zhang *et al.*, 2006).

The strength of hybridization of a probe with a target can be expressed as Gibbs free energy (ΔG) of the probe-target complex, or as melting temperature. Melting temperature is defined as the temperature at which half of the nucleic acid molecules in a solution are hybridized and half are in single stranded state (SantaLucia, 1998). Since hybridization is more difficult at higher temperatures, higher melting temperatures indicate stronger hybridization efficiency.

Another parameter that is usually considered in microarray design is the strength of possible secondary structures of probes (Premier Biosoft International, 2002). If one part of a probe is complementary to another part, the probe can fold to form a hairpin structure, which hampers hybridization of the target to that probe (SantaLucia and Hicks, 2004). Furthermore, the high density of probes in microarray spots could allow the formation of dimer structures between two probes with the same sequence if they are self-complementary. Therefore, stretches that are complementary with themselves or other parts of a probe should be avoided. The strength of both hairpin and dimer structures can be quantified with the same models as the probe-target affinities, and can be expressed as Gibbs free energy values (SantaLucia and Hicks, 2004).

A better understanding of the factors influencing the hybridization of target strands to microarray probes could lead to more reliable estimates of gene expression levels (Matveeva *et al.*, 2003; Wu and Irizarry, 2004). In this study, we use multiple linear regression to analyze the effects of several design parameters on the hybridization signals in a microarray experiment. We have designed an oligonucleotide microarray for the freshwater cyanobacterium *Synechocystis* PCC 6803 and present results of an experiment using RNA from salt-stressed and control cells. In particular, we show that important probe design parameters (e.g., melting temperature, hairpin and dimer structures, the size of genes, the length of probes, homology of probes with other parts of the genome, and positions of probes within genes) affect hybridization signal intensity and, thus, the measurement of gene expression patterns.

MATERIALS AND METHODS

DNA microarray design strategy

A DNA microarray for the freshwater cyanobacterium *Synechocystis* PCC 6803 was designed using the genomic information of this strain (Kaneko *et al.*, 1996; http://www.kazusa.or.jp/cyanobase). One to four oligonucleotide sequences of 45 to 60 bases long of each Open Reading Frame (hereafter referred to as genes) were chosen according to several design features, specified below. A total of 8091 probes representing the 3264 genes of *Synechocystis* PCC 6803 were designed.

Probe design was performed with assistance of the software package Array Designer 2.0 (Premier Biosoft International, 2002). The probe-design strategy consisted of finding all non-overlapping probes starting at very stringent conditions, followed by stepwise loosening of the criteria to obtain at least one probe for each gene. Array Designer uses the NCBI Basic Local Alignment Search Tool (BLAST; Altschul *et al*., 1997) to find and

exclude probe sequences with strong homologies with other parts of the genome, using a user-specified threshold for the expect value (E value). The E value of a probe indicates the number of times that a stronger homology would be expected between two random sequences with the same lengths as the probe and the genome. Thus, a unique probe for a gene has an E value close to zero with its target sequence in the genome. Probes that also have small E values with other parts of the genome are not very specific, however, and are preferably excluded from microarray design. NCBI BLAST can also exclude probes with low complexity, for instance probes with long repeats of one or two nucleotides. Probes with low complexity are not expected to be highly specific for their targets.

Another feature from the Array Designer program that we used for probe design is a rating parameter that is applied automatically to every probe found when performing a probe search (Premier Biosoft International, 2002). The rating parameter evaluates the quality of a probe according to its deviation from the target melting temperature (Tm), position along the gene (favouring probes closer to the 5' end of the sequence), hairpin ΔG, dimer ΔG, and long repetitions of the same one or two bases. According to the rating values, the probes are classified into 'good', 'poor' and 'bad' probes.

Array Designer performs probe searches with custom-defined melting temperatures and lengths on acceptable stretches of the genes as determined by BLAST search. We used four E-value thresholds between probes and non-target sequences within the genome, namely 10, 1, 0.1 and 0 (no BLAST search) to find probes that minimize cross-hybridization. At each E-value threshold, probes were searched with similar Tm values to have similar hybridization affinities across the whole array (SantaLucia, 1998). The target Tm was set at 88°C. We applied three Tm ranges, from 88±2 °C, via 88±5 °C to 88±10 °C. We searched for probe lengths, starting from 60 bases and decreasing the length stepwise down to 45 bases. When comparing probes within different Tm ranges, the rating parameter was recalculated manually assuming the same Tm window for all probes, using the algorithm in the manual of Array Designer (Premier Biosoft International, 2002). At each parameter combination, we searched for up to 10 non-overlapping probes per gene.

The probes found at the most stringent conditions (E value with other parts of the genome >10, Tm=88±2 °C and length=60) were exported to Microsoft EXCEL spreadsheet. Thereafter, probes from less stringent conditions were added to the spreadsheet in a stepwise fashion (loosening first the length requirement, then the Tm requirement, and at last the BLAST requirement), provided these probes did not overlap with probes added earlier to the spreadsheet. As an exception, shorter probes rated 'good' substituted overlapping longer probes rated 'poor' or 'bad'. Finally, when the same level of stringency yielded multiple probes, the probes were ranked according to their rating value.

For each gene, this design procedure yielded a list of probes in descending order of quality. The two probes that ranked highest were selected for the microarray.
In order to increase the coverage of some biologically interesting genes for our research group, one or two additional probes were added for a number of specific genes.

The designed probes were synthesized on microarray slides using ink-jet printing at the company Agilent (Hughes *et al.*, 2001).

Cell culture conditions

To investigate the performance of the microarray, *Synechocystis* PCC 6803 was grown in 6 batch cultures using BG-11 growth medium (Rippka *et al.*, 1979). The batch cultures were incubated in an Orbital Incubator (Gallenkamp) at 30°C. Once the batch cultures had reached an optical density (at 750 nm) of approximately $OD_{750} = 0.5\ cm^{-1}$, 0.5 M NaCl was added to 3 of the cultures to induce salt stress (Jeanjean *et al.*, 1993). Thus, we obtained 3 cultures exposed to salt stress, and 3 control cultures. After 8 h, the cultures were centrifuged at 4000 rpm for 20 min, and the supernatant was discarded. The cell pellets were resuspended in 1.3 mL of 0.3 M sucrose, and transferred to vials for storage. The vials were immersed till frozen in liquid nitrogen and stored at -80 ^{o}C afterwards.

RNA extraction, cDNA synthesis and labeling

For RNA extraction from the harvested cultures, cells were disrupted and homogenized with a Fast-Prep-Instrument (Q-Biogene) and matrix B. RNA was extracted from the homogenates, using RNAPro-Solution (Q-Biogene) and purified with RNeasy columns (Qiagen, Germany). The purified RNA was eluted using nuclease-free water. The quality of the RNA was verified using a Bioanalyzer (Agilent Technologies).

Dye-swap design

The RNA extracted from each sample was reverse-transcribed in two separate reactions in the presence of fluorochrome-modified nucleotides (Cy3 and Cy5) using random primers. For this purpose, 10 µg of RNA was used in each reaction. The labelled cDNA was cleaned to remove nucleotides that were not incorporated. This yielded cDNA labeled with Cy3 (green) and Cy5 (red) dyes.
Samples were hybridized in pairs (one salt stressed and one control) in a replicated dye-swap design (Churchill, 2002). Each sample pair was hybridized in two slides; in one slide using the red-labelled control sample and green-labelled salt-stressed sample while using the opposite combination in the other slide. This design yielded 6 microarrays (3 sample pairs), and a total of 12 series of 8091 data each. The microarray slides were placed in a rotating hybridization chamber (Agilent Technologies) and were left to hybridize for 16 h overnight. Thereafter, the hybridized microarray slides were scanned with an Agilent Microarray Scanner to generate image files. The signal intensities were determined from the image files using Feature Extraction Software (Agilent Technologies) at its default settings.

Statistical analysis

Signal intensities were normalized to the same distribution using the qspline algorithm (Workman *et al.*, 2002). The intensities of the 12 data series had the same distribution after normalization. Background subtraction was not performed, since the signal intensities were clearly distinguishable from the background, and the background was fairly homogeneous between and among slides (Wernisch *et al.*, 2003; Thygesen and Zwinderman, 2004).

For each spot on each slide, we calculated the mean log intensity A:

$$A = \frac{1}{2}\left({}^{2}\log I_{salt} + {}^{2}\log I_{control}\right)$$

where I_{salt} is the signal intensity of the salt-stressed sample, and $I_{control}$ is the signal intensity of the control sample. The change in gene expression in response to salt stress was expressed as the log ratio M:

$$M = {}^{2}\log\left(\frac{I_{salt}}{I_{control}}\right)$$

The log ratio M was plotted against the mean log intensity A, in so-called MA plots, to assess intensity-dependent effects on the log ratio (Smyth and Speed, 2003).

After normalization by the qspline algorithm, we test for significant changes in gene expression in response to salt stress using a gene-by-gene ANOVA model (Wolfinger *et al.* 2001). In contrast to Wolfinger *et al.* (2001), we take into account that different probes for the same gene may have different effects on signal intensity. Hence, our gene model is defined as:

$$r_{gabj} = G_g + (GT)_{ga} + P_b + (GTP)_{gab} + (GA)_{gj} + \gamma_{gabj}$$

where r is the normalized intensity for gene g (g=1,…,3264), treatment a (a=1,2), probe b (b=1,…,8091), and array j (j=1,2), G is the main effect of each gene, GT is the interaction effect between gene and treatment, P is the main effect of each probe, GTP is the probe-specific treatment effect on each gene, and GA is the interaction effect between gene and array. The error term γ is assumed to be normally distributed with zero mean and variance σ_{γ}^{2}.

The main effect G of each gene reflects the mean expression level of the gene across all treatments, probes, and arrays. The effects P, GTP and GA serve the same function as GA in Wolfinger *et al.* (2001), namely to model the confounding effects of probes and arrays on the intensities observed at each spot. The effect of interest is GT, which is an estimate of the change in gene expression due to the treatment (Kerr and Churchill, 2001; Wolfinger *et al.* 2001).

Since we investigate thousands of genes, traditional statistical criteria may assign significance to many genes that are not differentially expressed. For instance, suppose that the null hypothesis of no differential gene expression is tested at a significance level of $p < 0.05$, then we may expect a priori that for about 163 of the 3264 genes of *Synechocystis* PCC 6803 the null hypothesis will be rejected just by chance, even if these genes are not significantly expressed. These are called false positive rejections. We therefore used the false discovery rate introduced by Storey and Tibshirani (2003) to limit the number of false positive rejections. The false discovery rate is quantified by the q value, which expresses the proportion of significant results expected to be false positives. The q value is used quite

similarly as the traditional p value. We set the false discovery rate at $q < 0.05$, meaning that only 5% of those genes declared significant are expected to be false positives.

Effects of design parameters on the mean log intensity and log ratio of all probes were explored using multiple linear regression analysis (Sokal and Rohlf, 1995) using the software package R (R Development Core Team, 2005). The design parameters included in the regression analysis were melting temperature, hairpin and dimer hybridization free energy, probe length, maximum homology, gene length, and position of the probe along the gene sequence. Maximum homology of each probe was calculated as the size (measured in base pairs) of the longest stretch homologous to any gene other than its target gene. The values of partial regression coefficients depend on the measurement scales of the design parameters. To eliminate differences in measurement scale, we calculated beta coefficients (also known as standard partial regression coefficients). Beta coefficients allow comparison of the magnitudes of the effects of different design parameters (Sokal and Rohlf, 1995).

RESULTS

Distribution of microarray design properties

A microarray consisting of 8091 synthetic oligonucleotide probes representing the 3264 genes of the *Synechocystis* PCC 6803 genome was assembled. Figure 1 plots the log ratio of the signal intensities of the salt stress versus control treatment (M) as function of the mean signal intensity (A), for all 8091 probes. This MA plot shows that normalization of the data was successful as most of the log ratio data clustered around the value of 0. The MA plot also indicates that the probes captured the variation in gene expression between control and salt treatment. Several genes responded to the salt treatment, as their log ratio differed markedly from 0, with up-regulated genes above and down-regulated genes below the lowess line (Figure 1). Genes that were significantly up-regulated or down-regulated in response to salt stress are listed in Appendix 2A and 2B, respectively. These results are in general agreement with earlier studies on gene expression patterns of *Synechocystis* PCC 6803 in response to salt stress (Kanesaki *et al.*, 2002; Marin *et al.*, 2004). In the remainder of this paper, we will not focus on the biology of salt stress in cyanobacteria, but instead we will use this data set to study the properties of microarray design parameters.

Figure 2 displays the *in silico* derived properties of the probes for various applied design parameters. The melting temperatures (Tm) of probes varied between 83 and 90°C, with most of them lying close to or at the targeted Tm value of 88°C (Figure 2A). The distribution of ΔG values of the probe-target hybrid provides another measure of hybridization affinity (Figure 2B). The probes with lowest predicted affinity for their targets had ΔG values around -306 kJ/mol, while the probes with highest predicted affinity had ΔG values around -440 kJ/mol. Probes with high values of hairpin and dimer hybridization (i.e. those with a very negative ΔG) were not selected. The remaining ΔG of hairpin hybridization of the chosen probes varied between -33.5 and 0 kJ/mol (Figure 2C). Similarly, the ΔG of dimer hybridization of pairs of adjacent molecules of the same probe varied between -46 and 0 kJ/mol (Figure 2D). Note that the ΔG of hairpin and dimer structures was about one order of magnitude lower than the ΔG of the probe-target hybrids

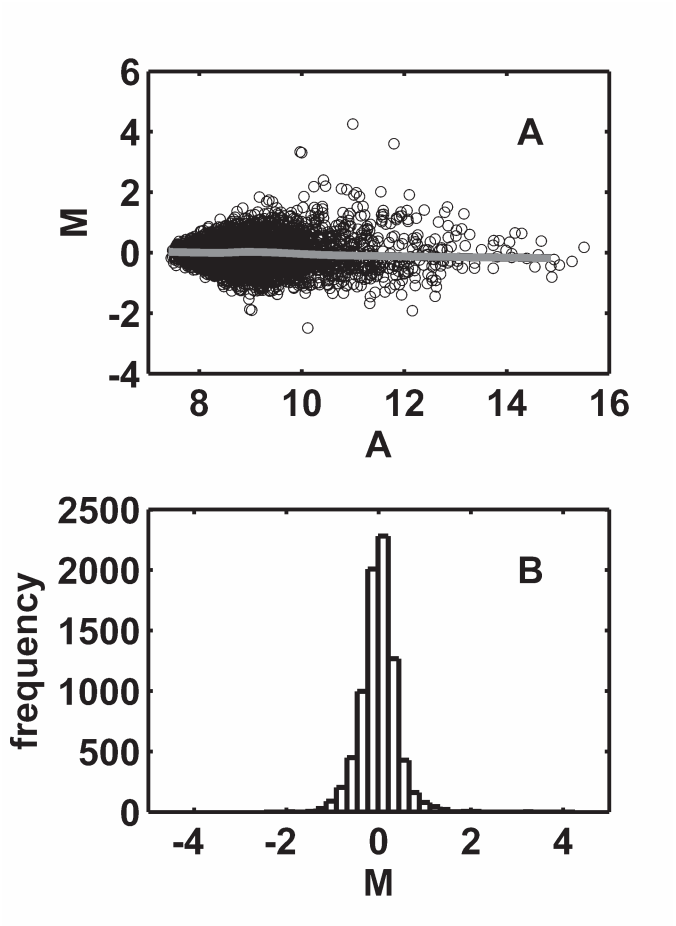

Figure 1. A) MA plot of the response in gene expression to salt stress. The log ratio M is plotted as a function of the mean log intensity A. Each datapoint corresponds to one probe. The gray line represents a lowess curve calculated from the data. B) Frequency histogram showing the distribution of the M values.

(compare Figure 2C,D with Figure 2B). The probes included in the microarray design were chosen to be 60 bases long, unless a significant gain in rating value or specificity was achieved by making them shorter. Therefore, about 5000 probes were 60 bases long, and the remaining 3091 probes were 45 to 59 bases long (Figure 2E). For detailed analysis of the effect of cross-homology on the signal intensities, the actual number of possible base pairings between a probe and other similar regions of the genome was counted. For each probe, the highest count was used as a predictor of cross-hybridization. The maximum number of shared bases with other genes was relatively large, with a value of around 30 bases for most probes (Figure 2F). The length of the gene sequences is a property of the genome of *Synechocystis* PCC 6803 and not a probe design feature. However, since the length of the gene sequence may also affect hybridization on microarrays, its distribution is also shown (Figure 2G). Gene lengths varied between 90 and more than 12000 bases, but most genes were shorter than 2000 bases.The position of the probes within gene sequences was selected near the 5' end (Figure 1H). Regions of the genome with low complexity were avoided from the design with a low complexity filter available at the NCBI BLAST tool. Probes with long repetitions of one or two bases motifs were also avoided. No probes had repeated regions longer than 5 bases.

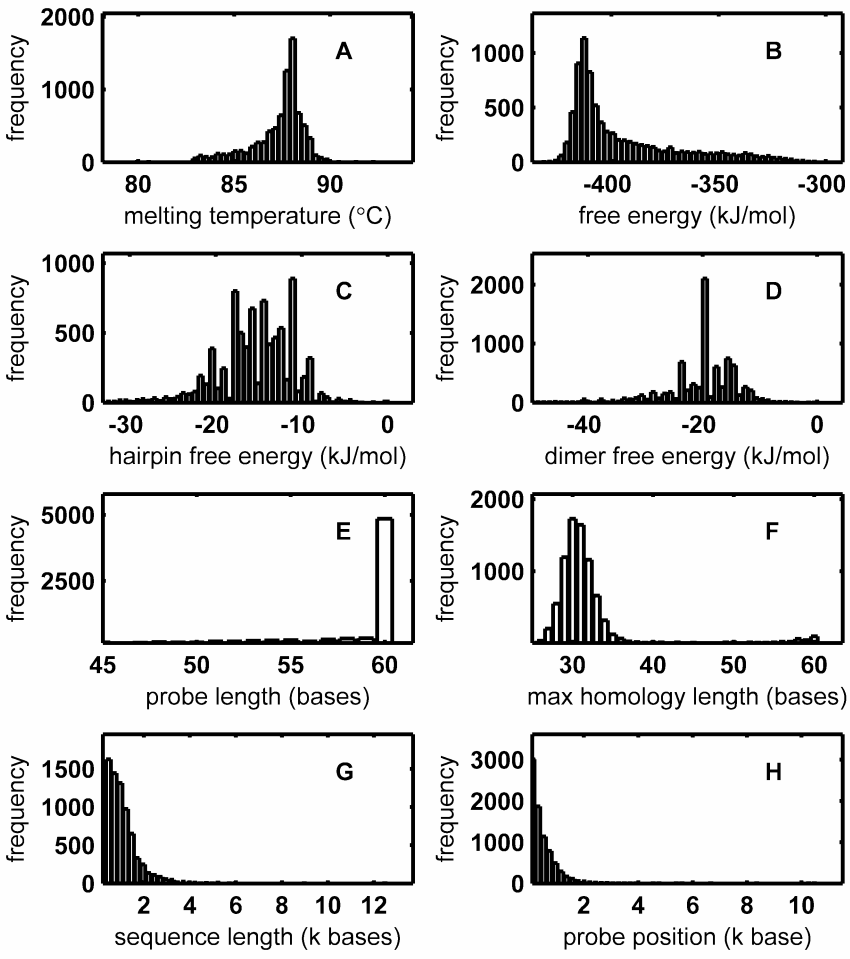

Figure 2. Frequency distribution of various properties of the probes in the microarray design.

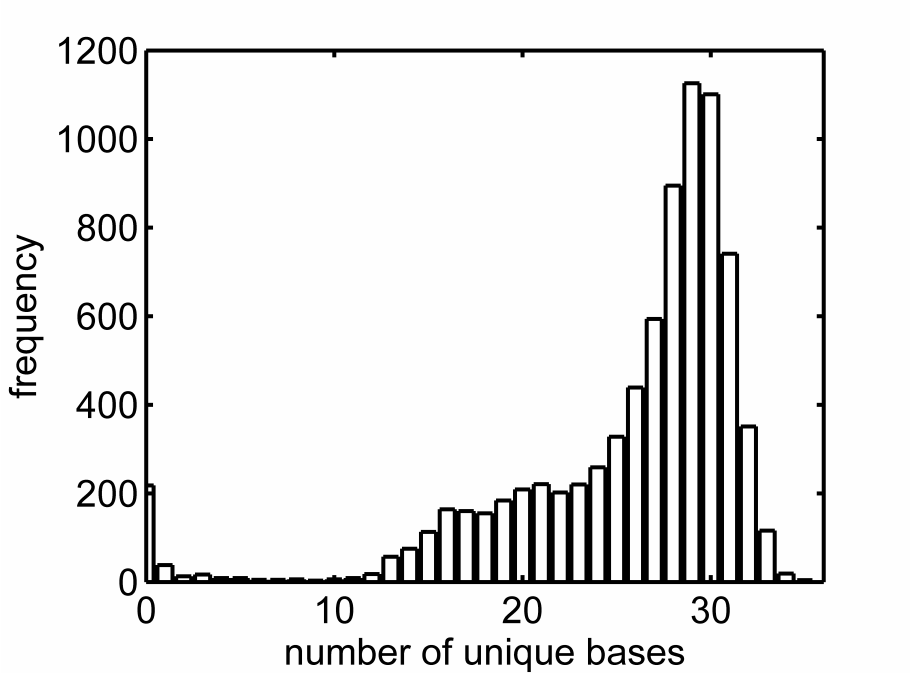

Figure 3. Frequency distribution of the number of unique bases (i.e., bases not shared with the most homologous stretch in other parts of the genome).

Unique bases and cross-hybridisation

The distribution of the number of unique bases (i.e., bases not shared between the probes and their most homologous stretch elsewhere in the genome) is shown in Figure 3. Homology of probes with other parts of the genome, outside the target area, was avoided as much as possible. Nevertheless, the occurrence of small shared stretches between the probes and other open reading frames in the genome appeared unavoidable. In the great majority of cases (89%) the number of unique bases was equal to or larger than 18 bases, which is sufficient for selective preference and low occurrence of cross hybridisation in the 60 mer design that was used by Agilent for the DNA microarrays in this study (see discussion).

Effect of design parameters on hybridization intensity

We tested to what extent probe design parameters affected the observed hybridization intensities (Table 1). Multiple linear regression analysis revealed that all design parameters had highly significant effects on the mean log intensity of the array signals. The design parameters accounted for almost 10% of the total variation, as determined by the coefficient of multiple determination of the model (Table 1). Melting temperature (Tm) had a positive effect on mean log intensity (Table 1; Figure 4A). To illustrate this pattern in more detail, Tm data were divided into bins of 0.1 ^{o}C. For each bin, the average value of mean log intensity was calculated. This shows that mean log intensity increased with melting temperature (Figure 4B). All else being equal, probe length had a negative effect on mean

log intensity (Table 1). That is, short probes had a higher hybridization signal than long probes. This indicates that short probes are better accessible for their targets than long probes. Less free energy gain in dimer and hairpin formation had positive effects on mean log intensity (Table 1), because hairpin and dimer hybridization compete with probe-target hybridization, and are therefore diminishing the wanted hybridization signal. Sequence length of the gene and the position of the probe along the gene in the microarray design had significant negative effects on mean log intensity (Table 1). The negative effect of gene sequence length likely results from secondary structure formation in the gene, which could compete with probe-target hybridization in a similar way as hairpin and dimer formation. The negative effect on signal intensity for probes selected further away from the 5' end of the gene is related to the occurrence of read through of the cDNA in random primed reverse transcription to the 5' prime end of each mRNA transcript, thus rendering more cDNA product of the 5' part of the gene sequence. The maximum homology length (i.e., the maximum number of bases shared between a probe and stretches of genes other than its intended target) had a significant positive effect on mean log intensity (Table 1; Figure 5A, B). However, the magnitude of this effect was small compared to the effects of the other design parameters (Table 1).

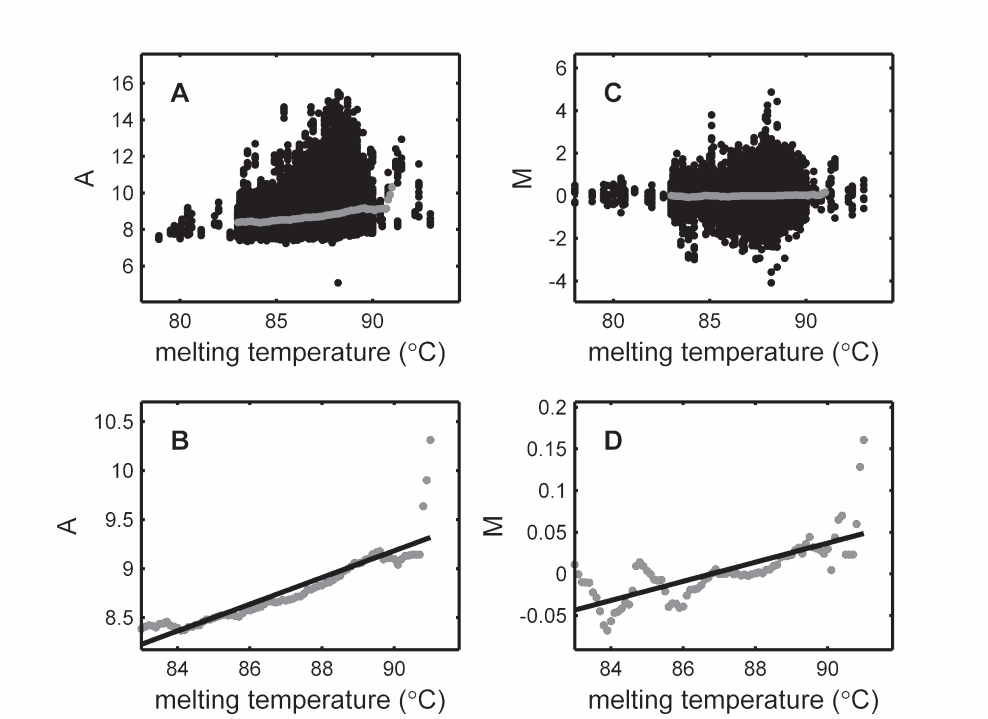

Figure 4. Effect of melting temperature (Tm) on the mean log intensity A (a measure of hybridization intensity) and the log ratio M (a measure of the response in gene expression). Light gray data points indicate mean values when the data set is subdivided into bins of 0.1 ^{o}C. Solid lines in (B) and (D) are trendlines based on linear regression.

Table 1. Partial regression coefficients obtained from multiple linear regression of the mean log intensities (A) against different design parameters.

Design parameter	Regression coefficient	St.Error	Beta coefficient	t	p
melting temperature (Tm)	0.1458	0.0028	0.2351	52.78	$< 10^{-15}$
probe length	-0.0294	0.0009	-0.1500	-34.58	$< 10^{-15}$
hairpin ΔG	0.0218	0.0009	0.1110	23.08	$< 10^{-15}$
dimer ΔG	0.0113	0.0007	0.0818	16.96	$< 10^{-15}$
log(probe position)	-0.0264	0.0021	-0.0650	-12.65	$< 10^{-15}$
log(sequence length)	-0.0558	0.0047	-0.0631	-11.77	$< 10^{-15}$
max. homology length	0.0050	0.0007	0.0032	7.26	$< 10^{-12}$

Note: The coefficient of multiple determination of the complete model was $R^2 = 0.096$ (N = 48,546; $p < 10^{-15}$). Units of measurement are: Tm, in ºC; hairpin and dimer ΔG, in kJ/mol; probe length, maximum homology length, sequence length and probe position, in number of bases

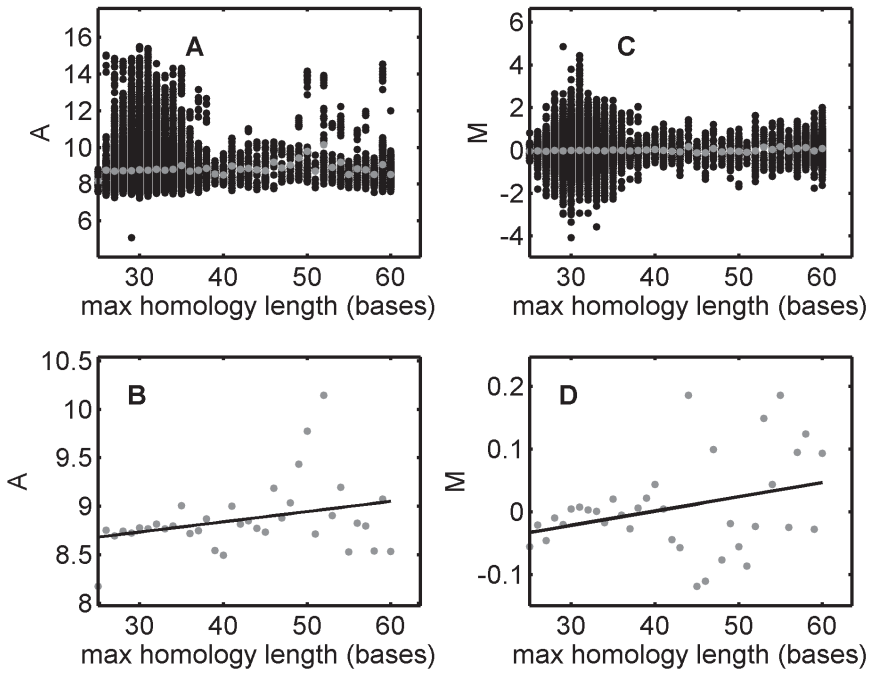

Figure 5. Effect of maximum homology length with other parts of the genome on the mean log intensity A (a measure of hybridization intensity) and the log ratio M (a measure of the response in gene expression). Light gray data points indicate mean values when the data set is subdivided into bins of 1 base. Solid lines in (B) and (D) are trendlines based on linear regression.

Effect of design parameters on control and treatment samples separately

Effects of design parameters on mean log intensity, shown in Table 1, were based on the combined data of treatment and control samples. The question remained whether these effects were confounded by the treatment. That is, perhaps the impact of the design parameters on hybridization intensity was affected by salt stress? To answer this question, the same multiple linear regression analysis was applied to the treatment samples and control samples separately. The coefficients obtained for the treatment samples (Table 2) and control samples (Table 3) were the same as for the combined data set (Table 1). The coefficients of multiple determination were also very similar. Therefore, we conclude that the effects of the design parameter on the hybridization intensity were not confounded by the experimental treatment.

Table 2. Partial regression coefficients obtained from multiple linear regression of the mean log intensities of the control samples against different design parameters.

Design parameter	Regression coefficient	St.Error	Beta coefficient	t	p
melting temperature (Tm)	0.1401	0.0029	0.2181	48.74	$< 10^{-15}$
probe length	-0.0292	0.0009	-0.1437	-32.98	$< 10^{-15}$
hairpin ΔG	0.0222	0.0010	0.1094	22.65	$< 10^{-15}$
dimer ΔG	0.0118	0.0007	0.0827	17.07	$< 10^{-15}$
log(probe position)	-0.0290	0.0022	-0.0689	-13.36	$< 10^{-15}$
log(sequence length)	-0.0496	0.0049	-0.0541	-10.04	$< 10^{-15}$
max. homology length	0.0036	0.0007	0.0221	4.96	7.15×10^{-7}

Note: The coefficient of multiple determination of the complete model was $R^2 = 0.096$ (N = 48,546; $p < 10^{-15}$). The units of measurement are: Tm, in ºC; hairpin ΔG and dimer ΔG, in kJ/mol; probe length, maximum homology length, sequence length and probe position, in number of bases

Table 3. Partial regression coefficients obtained from multiple linear regression of the mean log intensities of the salt treated samples against different design parameters.

Design parameter	Regression coefficient	St.Error	Beta coefficient	t	p
melting temperature (Tm)	0.1514	0.0029	0.2366	53.03	$< 10^{-15}$
probe length	-0.0296	0.0009	-0.1463	-33.68	$< 10^{-15}$
hairpin ΔG	0.0213	0.0010	0.1052	21.85	$< 10^{-15}$
dimer ΔG	0.0107	0.0007	0.0754	15.61	$< 10^{-15}$
log(probe position)	-0.0237	0.0022	-0.0567	-11.03	$< 10^{-15}$
log(sequence length)	-0.0621	0.0050	-0.0680	-12.67	$< 10^{-15}$
max. homology length	0.0065	0.0007	0.0403	9.06	$< 10^{-15}$

Note: The coefficient of multiple determination of the complete model was $R^2 = 0.087$ (N = 48,546; $p < 10^{-15}$). The units of measurement are: Tm, in ºC; hairpin ΔG and dimer ΔG, in kJ/mol; probe length, maximum homology length, sequence length and probe position, in number of bases.

Effect of design parameters on log ratios

The effect of probe design on the hybridisation intensity A has been reported above for a transcript group (i.e. red or green signals). It was questioned whether choices in probe design would also affect the log ratio M between two transcript groups. Multiple linear regression analysis revealed that most design parameters had small but significant effects on the log ratios of the spots (Table 4). It implies that design parameters affect assessment of significant changes in gene expression. However, in prospect the contribution of the design parameters to the total variability in the log ratio was small; they explained only 0.3% of the variation according to the coefficient of multiple determination. The homology of probes with other parts of the genome had the largest effect on the log ratio (Table 4; Figure 4 C, D). Probes with a higher homology with other parts of the genome would have a higher probability to cause variability to extend to the log ratio. The melting temperature also had a relatively large effect on the log ratio (Table 4; Figure 3 C, D), whereas the effects of sequence length, probe position, and the free energy of dimer formation were smaller (Table 4). Hairpin formation and probe length had no significant effect on the log ratio.

Table 4. Partial regression coefficients obtained from multiple linear regression of the log ratios (M) against different design parameters.

Design parameter	Regression coefficient	St. Error	Beta coefficient	t	p
max. homology length	0.0029	0.0004	0.0354	7.60	$< 10^{-13}$
melting temperature (Tm)	0.0112	0.0015	0.0345	7.37	$< 10^{-12}$
log(sequence length)	-0.0126	0.0026	-0.0270	-4.79	$< 10^{-5}$
log(probe position)	0.0052	0.0012	0.0245	4.54	$< 10^{-5}$
dimer ΔG	-0.0011	0.0004	-0.0151	-2.97	0.00296
hairpin ΔG	-0.0009	0.0005	-0.0091	-1.81	0.07063
probe length	-0.0004	0.0005	-0.0039	-0.85	0.39416

Note: The coefficient of multiple determination of the complete model was $R^2 = 0.003$ (N = 48,546; $p < 10^{-15}$). The units of measurement are: Tm, in ºC; hairpin ΔG and dimer ΔG, in kJ/mol; probe length, maximum homology length, sequence length and probe position, in number of bases.

DISCUSSION

Thermodynamic properties in microarray design

A good DNA microarray design should combine low artificial variability of the data with full coverage of all ORF and genes present on the genome. This requires careful negotiation between probe quality features while still attempting to cover all genes. Microarray design is therefore bounded by the necessity to suppress cross-hybridization and the availability of probes with suitable thermodynamic properties. The extent of cross-hybridization is a much debated issue in array application (Talla *et al.,* 2003; Wu *et al.*, 2005; Reilly *et al.*, 2006;

Chen *et al.*, 2006), and limitations in microarray design have become evident in several application fields (e.g., microbial community studies; Wagner *et al.*, 2007).

In our study, we considered a very large search space, searching for probes across different levels of stringency. More specifically, we searched for probes with similar thermodynamic properties, to ensure that the hybridization signal depends on the RNA concentration as much as possible. The homology between a probe and other parts of the genome than its intended target was controlled at the design stage using NCBI BLAST searches (Altschul *et al*., 1997). For the 60 mer probe length array design used in this work it has been shown that a minimum of 18 specific nucleotides is required to obtain unique hybridization results (Hughes *et al.,* 2001). This minimum number of specific nucleotides is only valid for a hybridization protocol based on narrowly defined thermodynamic properties for probe-target interaction that permit stringent washing for best selectivity. Effects of thermodynamic properties of probes on their hybridization with target molecules has been studied in great detail with DNA in solution (SantaLucia, 1998). To estimate the free energy of probe-target hybridization, hairpin and dimer hybridization respectively, we used established estimates for hybridization free energies of DNA in solution (SantaLucia and Hicks, 2004). Applicability of properties of DNA hybridization in solution for prediction of hybridization of targets to probes on a rigid surface has been studied with a limited set of spike-in targets with known concentration (Wu and Irizarry, 2004; He *et al.*, 2005). Here we have extended this application to a whole-genome DNA microarray. Our microarray design strategy

Our analysis shows that the selection of probes for microarray studies should be based on careful choices of design parameters, like melting temperature, hairpin hybridization, dimer hybridization and cross-homology (Southern *et al.*, 1999; Matveeva *et al.*, 2003; Talla *et al.*, 2003). Another parameter that should be considered for microarray design is the length of the probes (Southern *et al.,* 1999), which demonstrated a significant effect on the hybridization signals. The length of the target sequence is a natural property of the genes, and cannot be used for microarray design, but it also had significant effects on the hybridization reaction. Our results differ from Talla *et al.* (2003), who could not detect effects of probe length and the position of the probe along the gene when plotting signal intensity as a function of these design parameters. In contrast, we found significant effects of probe length and position along the gene on signal intensity by carefully taking into account contributions of the other design parameters using multiple linear regression.

Principal observations on the impact of design parameters

One advantage of oligonucleotide microarrays is that they allow some control over the thermodynamic properties of probes. This enables selection of probes from the full genome with rather similar thermodynamic characteristics. Yet, even though we selected probes with similar thermodynamic properties, their thermodynamic properties still explained 10% of the variation in signal intensity (Table 1). These effects of probe design on signal intensity were independent of the experimental treatment (Tables 2 and 3). Effects of design parameters on the variation in signal intensity are expected to be larger in studies using microarrays with less stringently designed probes. We therefore recommend a consistent procedure to select probes with similar thermodynamic properties, as exemplified by this study. Criteria for the selection of probes should be based on those

design parameters with the largest influence on the signal intensities. According to our findings with the cyanobacterium *Synechocystis*, the magnitude of the effects of design parameters can be ranked as follows: homology to other parts of the genome > melting temperature > probe position > dimer ΔG.

Many thermodynamic properties of the probes also had significant effects on the log ratios of treatment versus control signal. As expected, however, effects on the log ratios were much smaller than effects on mean log intensity. More specifically, thermodynamic properties explained only 0.3% of the total variation in the log ratios (Table 4). This confirms the common usage of log ratios as a very effective means to remove undesired effects of probe design and to correct for differences in hybridization kinetics between samples (Quackenbush, 2002).

Understanding of the physiology of an organism would greatly benefit from quantitative comparison of gene expression levels of different genes across the genome. Whether use of absolute signal intensities of array data is allowed is a topic under debate (Kerr and Churchill, 2001; Hekstra *et al*. 2003; Draghici *et al.*, 2005). In our study, 10% of the variation in signal intensity was explained by thermodynamic properties of the probes. This leaves 90% of variation in signal intensity unexplained. Ideally, this remaining variation should largely reflect variation in gene expression levels, which is the signal of interest in microarray studies. However, part of this remaining variation may also reflect nonlinear effects of the probes' thermodynamic properties not captured by our multiple regression analysis. Or it may reflect other properties of probe design not investigated here, perhaps properties that are unique to each probe-target pair. Future studies should aim to resolve the remaining variation in signal intensity in further detail. This might ultimately permit interpretation of absolute signal intensities across the entire genome of an organism.

CONCLUSIONS

This study was focussed on the role of probe design properties on the performance of a DNA microarray. Our results support earlier findings that the thermodynamic properties of probes for DNA microarrays may markedly affect hybridization intensities (Matveeva *et al.*, 2003; He *et al.*, 2005). With an increasing amount of whole genome sequences available, the quality of DNA microarray designs will become of increasing interest. We recommend a careful probe design strategy aiming at probes with similar thermodynamic properties across the entire range spotted on the microarray.

ACKNOWLEDGEMENTS

We kindly thank Dr. Timo Breit for helpful comments that improved the manuscript. EAvW was financially supported by a scholarship from Consejo Nacional para la Ciencia y Tecnología (Mexico). The research of EAvW and JH was further supported by the Earth and Life Sciences Foundation (ALW), which is subsidized by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nature Genetics 21:33-37.

Chen YA, Chou CC, Lu X, Slate EH, Peck K, Xu W, Voit EO, Almeida JS. 2006. A multivariate prediction model for microarray cross-hybridization. BMC Bioinformatics 7:101.

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

Draghici S, Khatri P, Eklund AC, Szallasi Z. 2005. Reliability and reproducibility issues in DNA microarray measurements. Trends in Genetics 22:101-109.

He Z, Wu L, Li X, Fields MW, Zhou J. 2005. Empirical establishment of oligonucleotide probe design criteria. Applied and Environmental Microbiology 71: 3753-3760.

Hekstra D, Taussig AR, Magnasco M, Naef F. 2003. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. Nucleic Acids Research 31:1962-1968.

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. Nature Biotechnology 19:342-347.

Jeanjean R, Matthijs HCP, Onana B, Havaux M, Joset F. 1993. Exposure of the cyanobacterium *Synechocystis* PCC6803 to salt stress induces concerted changes in respiration and photosynthesis. Plant and Cell Physiology 34:1073-1079.

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 PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA research 3:109-136.

Kerr K, Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. Genetical Research 77:123-128.

Matveeva OV, Shabalina SA, Nemtsov VA, Tsodikov AD, Gesteland RF, Atkins JF. 2003. Thermodynamic calculations and statistical correlations for oligo-probes design. Nucleic Acids Research 31: 4211-4217.

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

Premier Biosoft International. 2002. Array Designer 2.0 Manual. www.PremierBiosoft.com.

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

R Development Core Team. 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

Reilly C, Raghavan A, Bohjanen P. 2006. Global assessment of cross-hybridization for oligonucleotide arrays. Journal of Biomolecular Techniques 17:163-172.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. Journal of General Microbiology 111:1-61.

SantaLucia J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences (USA) 95:1460-1465.

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

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

Sokal RR, Rohlf FJ. 1995. Biometry. 3rd Ed. Freeman, New York.

Southern E, Mir K, Shchepinov M. 1999. Molecular interactions on microarrays. Nature Genetics 21:5-9.

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

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.

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

Wagner M, Smidt H, Loy A, Zhou J. 2007. Unravelling microbial communities with DNA-microarrays: challenges and future directions. Microbial Ecology 53: 498-506.

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

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. Journal of Computational Biology 8:625-637.

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, 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.

Wu Z, Irizarry R. 2004. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. Proceedings of the 8th annual international conference on Research in Computational Molecular Biology, pp. 98-106. ACM Press, New York, USA.

Wu C, Carta R, Zhang L. 2005. Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Research 33:e84.

Zhang L, Wu C, Carta R, Zao H. 2006. Free energy of DNA duplex formation on short oligonucleotide microarrays. Nucleic Acids Research 35:e18.