STUDIES ON THE RELATIONSHIPS BETWEEN OLIGONUCLEOTIDE PROBE PROPERTIES AND HYBRIDIZATION SIGNAL INTENSITIES

by

Raad Zuhair Gharaibeh

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Information Technology

Charlotte

2009

Approved by:

____
Dr. Cynthia G. Gibas

____
Dr. Anthony A. Fodor

____
Dr. Jennifer W. Weller

____
Dr. Ann E. Loraine

____
Dr. Brian T. Cooper

ABSTRACT

RAAD ZUHAIR GHARAIBEH. Studies on the relationships between oligonucleotide probe properties and hybridization signal intensities. (Under the direction of DR. CYNTHIA G. GIBAS)

Microarray technology is a commonly used tool in biomedical research for assessing global gene expression, surveying DNA sequence variations, and studying alternative gene splicing. Given the wide range of applications of this technology, comprehensive understanding of its underlying mechanisms is of importance. The focus of this work is on contributions from microarray probe properties (probe secondary structure: ΔG_{ss}, probe-target binding energy: ΔG, probe-target mismatch) to the signal intensity. The benefits of incorporating or ignoring these properties to the process of microarray probe design and selection, as well as to microarray data preprocessing and analysis, are reported. Four related studies are described in this thesis. In the first, probe secondary structure was found to account for up to 3% of all variation on Affymetrix microarrays. In the second, a dinucleotide affinity model was developed and found to enhance the detection of differentially expressed genes when implemented as a background correction procedure in GeneChip preprocessing algorithms. This model is consistent with physical models of binding affinity of the probe target pair, which depends on the nearest-neighbor stacking interactions in addition to base-pairing. In the remaining studies, the importance of incorporating biophysical factors in both the design and the analysis of microarrays has been investigated. First, the impact of incorporation of a complete model of the hybridization equilibrium was tested. The results suggest that the use of probe ‘percent bound’, predicted by equilibrium models of hybridization, is a useful factor in predicting and assessing the behavior of long oligonucleotide probes. However, a universal probe-property-independent three-parameter Langmuir model has also been tested, and this simple model has been shown to be as, or more, effective as complex, computationally expensive models developed for microarray target concentration estimation. The simple, platform-independent model can equal or even outperform models that explicitly incorporate probe properties, such as the model incorporating probe percent bound developed in Chapter Three. This suggests that with a “spiked-in” concentration series targeting as few as 5-10 genes, reliable estimation of target concentration can be achieved for the entire microarray.

DEDICATION

This work is dedicated to:

Mahmood K. Al-Waked and *Mohammad A. Gharaibeh*

You both will always be remembered.

ACKNOWLEDGMENTS

Now that I realized it is the journey not the destination, I would like to express my sincere gratitude to: Zuhair, Sabha, Reham, Batool, Rkan, Sakher, Muneera, Yousef and Ahmad, who without their help, support and encouragement, my journey would not have been possible.

I would like to thank The Graduate Assistant Support Plan (GASP) of UNC-Charlotte for awarding me the financial support for my journey. I am very grateful to the Department of Bioinformatics and Genomics' faculty, staff and students for providing me with a home away from home. I would like to extend my appreciation and gratitude to Dr. Lawrence Mays, Ms. Patricia Artis and Ms. Elise Marshall for their great help and support.

This dissertation would not have been completed without the dedicated efforts of my committee members: Dr. Anthony Fodor, Dr. Jennifer Weller, Dr. Ann Loraine and Dr. Brian Cooper. I'm thankful for your time, efforts, comments, critiques and discussions.

I owe a huge debt of gratitude to Dr. Anthony Fodor who was a pivotal factor in the process of completing this dissertation. His contributions to my academic, professional and personal life out balance a simple thank you.

My sincere acknowledgment to my advisor Dr. Cynthia Gibas, whom diligent dedication ensured my academic, professional, financial and moral well-being during my journey. Cynthia, thank you for being there for me in my dark and bright moments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Parallel to the growing pace of high-throughput genome sequencing, there are similar efforts for transforming genomic sequence information into functional biological knowledge. DNA Microarray technology has enabled the investigation of gene expression at the transcript level on a genome-wide scale, and is considered to be one of the primary research tools in the post-genomic era. DNA microarray technology [1] allows thousands of genes to be assayed at once, offering global views of biological processes by providing a systematic way to survey gene expression and subtle DNA sequence variations [2]. These applications have attracted great interest and investment from scientists and companies all over the world.

A microarray is a surface that provides support to tethered nucleic acids called probes [3]. These probes are designed to interact with a mixture of labeled nucleic acids called targets [3]. The synthetic oligonucleotide probes used in microarray experiments vary from short (20-25mer) probes to long (50-70mer) probes [4]. The target mixture for gene expression arrays is derived from total mRNA extracted from the cell under specific conditions. Common experimental designs include comparison of transcript levels at steady state under different conditions, or comparison of samples taken at several time points in a process (e.g.

cell division) [2]. When a labeled target binds to its complementary probe, a stable interaction results and a signal is detected. Signal intensity is interpreted as reflecting transcript abundance, presence and/or absence of that target [5-6].

In order to ensure that the signal intensity detected in the experiment truly represents what targets are present in the mixture, careful selection of appropriate probes based on a solid understanding of the contributions of probe properties to the hybridization signal intensity is important. This is an essential step in order to add quantitative power to microarray technology. This problem was approached from different angles, addressing both specific properties which might later serve as design criteria, and integrated models of probe hybridization behavior, as used in signal interpretation. First, probe secondary structure was examined through thermodynamic modeling combined with statistical analysis to understand the consequences of intramolecular structure formation on the behavior of microarray probes. By employing physical models of binding affinity, which depend on the nearest-neighbor stacking interactions in addition to base-pairing, a similar approach to the one above was used to study probe affinity. Using a multi-state equilibrium model of solution hybridization provided insights into the usefulness of an N-state solution simulation of probe-target interactions in predicting the behavior of long microarray probes. Collapsing the collected knowledge from the previous parts resulted in accurate estimation of target concentration on microarrays.

1.1 Oligonucleotide probe properties and oligoarray design

Basically, there are two major types of DNA microarrays: cDNA arrays and oligonucleotide arrays. A cDNA array is created by depositing PCR products on the array surface. An oligonucleotide array is made up from synthetic oligonucleotides that are either spotted or synthesized *in situ* on the array surface. The latter is the focus of this dissertation.

Regardless of the platform used, quantitative detection of transcripts requires that the probe exhibit a sensitive and predictable response to a range of concentrations of specific targets. Such a response must occur in the presence of complex mixture of non-specific targets. In fact, the signal intensity obtained from the array is highly dependent on the probe design. A good probe should demonstrate several characteristics such as high sensitivity, high specificity, low noise and low bias [4, 7]. The process of probe design is dependent on several characteristics of the probe (probe properties). Basic criteria that are used in probe design are summarized in Table 1.1.

It is also likely that factors such as reaction volume, electrostatics, diffusion and surface effects, reaction thermodynamics and kinetics, competitive binding effects, duplex stability and transition states and target concentrations have crucial consequences for probe-target interactions [3, 8-9].

TABLE 1.1 Probe design and selection criteria.
Commonly used criteria for probe design and selection in freely available oligonucleotide probe design tools [4].

Design Tool	Sequence Similarity Search	Contiguous Identity	% Identity	Target-Probe Mismatch Position	Forward/Reverse Strand Match
ArrayOligoSelector [10]	BLAST	No	?	No	No
GoArrays [11]	BLAST, w = 7	Yes	Yes	No	?
OligoArray [12-13]	BLAST, w = 7	?	No	No	No
OliCheck [14]	BLAST	Yes	No	Yes	Yes
Oligodb [15]	BLAST	No	No	No	No
OligoDesign [16]	BLAST, w = 9	No	Yes	No	No
OligoPicker [17]	BLAST, w = 8	Yes	Yes	No	No
OligoWiz [18]	BLAST	Yes	Yes	No	?
Oliz [19]	BLAST	Yes	Yes	No	No
Osprey [20]	BLAST	Yes	?	No	Yes (?)
Picky [21]	Suffix array	Yes	Yes	No	Yes
PRIMEGENS [22]	BLAST	No	?	?	?
PROBESEL [23]	Suffix tree	No	Yes	No	?
ProbeSelect [24]	Suffix array	Yes	No	No	Yes
Promide [25]	Suffix array	No	No	No	?
ROSO [26]	BLAST, w = 7	No	Yes	No	?
YODA [27]	SeqMatch, w = 4	Yes	Yes	No	?

Table 1.1 (continued)

Design Tool	%GC	ΔG	T_m	T_m Range	NSH	Probe Secondary Structure	Dimer	Hairpin
ArrayOligoSelector [10]	Yes	?	NN (?)	No	Yes	SW	Yes	?
GoArrays [11]	No	?	NN; SL98	Yes	No	MFOLD	?	?
OligoArray [12-13]	Yes	Yes	NN; SL98	Yes	Yes	MFOLD	Yes	Yes
OliCheck [14]	?	?	Yes (?)	?	No	?	?	?
Oligodb [15]	Yes	No	NN; melting	?	No	MFOLD	Yes	Yes
OligoDesign [16]	No	No	NN; SL98	No	No	Nussinov	Yes	?
OligoPicker [17]	No	No	GC; Schildkraut	Yes	No	BLAST	Yes	Yes
OligoWiz [18]	No	Yes	NN (?)	Yes	No	Yes (unknown)	Yes	?
Oliz [19]	Yes	No	Yes (?)	Yes	No	No	No	No
Osprey [20]	No	Yes	NN; SL98	Yes	Yes	MFOLD	Yes	Yes
Picky [21]	Yes	?	NN; SL96	Yes	Yes	Yes (unknown)	Yes	Yes
PRIMEGENS [22]	Primer3	?	Breslauer	No	?	Primer3	Yes	?
PROBESEL [23]	No	No	NN; SL98	No	No	No	No	No
ProbeSelect [24]	Yes	Yes	NN; SL98	No	No	No	No	?
Promide [25]	No	No	NN; SL98	Yes	No	Yes (unknown)	Yes	No
ROSO [26]	Yes	Yes	NN; SL98	Yes	No	Yes (unknown)	Yes	Yes
YODA [27]	Yes	?	NN; SL98	Yes	No	Yes (unknown)	Yes	?

Sequence Similarity Search: Algorithm used to determine sequence similarity between a probe and a non-target sequence; Contiguous Identity: Uses stretches of contiguous sequence identities between a probe and a non-target sequence for cross-hybridization check; % Identity: Uses the overall percentage of sequence identity between a probe and a non-target sequence for cross-hybridization check. (Legend continues on the next page)

Table 1.1 (continued)

Target-Probe Mismatch Position: Account for the impact of mismatch positions between the target and probes; Forward/Reverse Strand Match: Uses similarity searches against both the forward strand and the reverse-complement; %GC: Uses percent of G and C nucleotides in the probe for heuristics; ΔG: Calculates and uses the Gibbs free energy ΔG of the probe-target duplex as a measure of duplex stability; T_m: Calculates and uses the melting temperature of the probe-target to characterize and compare the thermodynamic behavior of probe candidates; T_m Range: Whether the melting temperature of probe candidates is thresholded, a wider range of T_ms provides a larger search space and gives more flexibility for finding specific probes; NSH: Non Specific Hybridization, whether the cross-hybridization potential of a candidate probe with all its non-targets is calculated; Probe Secondary Structure: Whether the tool tries to predict potential stable secondary structures that the probe may form; Dimer: Whether the tool makes any calculations to predict dimerization of the probe, this is a special case of probe secondary structure; Hairpin: Whether the tool makes any calculations to predict hairpins within the probe, this is a special case of probe secondary structure; ?: Not known from information published; w: Word size used for BLAST search; NN: Nearest Neighbour two-state model; SL98: [28]; SL96: [29]; Schildkraut: [30]; Breslauer: [31]; melting: [32]; Nussinov: [33]; Primer3: [34]; BLAST: [35]; MFOLD: [36]; SW: [37]

The objective of probe design and selection is to provide a set of probes with uniform hybridization properties at the reaction conditions employed. The design and selection process must maximize probe specificity ("the ability of a probe to provide a signal that is influenced only by the presence of the target molecule" [38]) and sensitivity ("a measure of how little is lost of the signal reflecting specific hybridization between the probe and its target" [4]), and minimize probe bias ("the systematic deviation of the measurement from the true signal due to probe-specific or other technical effects" [4]) and noise ("random signal variation" [4]).

Probe sequence design is a complex process for many reasons, ranging from simple technical difficulties to sophisticated theoretical problems. To mention some of these difficulties, we can consider the fact that the prediction of actively transcribed regions of the genome is still far from exhaustive, so the mixture of target transcripts that need to be discriminated in a sample is not always fully known. In a case where the transcripts are known, then the prediction of the properties of a candidate probe under competitive hybridization with a known mixture of different transcripts is, on its own, a very complex thermodynamic modeling challenge. There are many biophysical factors (Table 1.1) that produce inter- and intra-molecular interactions that in turn affect probe-target binding behavior and affinity. These factors should be understood in the first place, and then taken into account when designing oligonucleotide probes and analyzing microarray data.

For probes of very short length, such as those used on Affymetrix arrays and with the Motorola CodeLink technology, some excellent studies of probe properties, target properties and the behavior of probe-target pairs have been published [39-42], but the findings of these studies cannot be assumed to apply to longer oligonucleotide arrays. This is because the coil to duplex transition is unlikely to be two-state [29] due to multiple nucleation sites on longer probes [43]. As a consequence, most oligonucleotide probe design approaches regress to crude approximations and heuristic approaches, with modeling replaced by sequence similarity searches and alignments, and *ad hoc* rules of thumb regarding probe sequence complexity and secondary structure. These simplified approaches are not sufficiently reliable on their own and usually require experimental validation of candidate probes in a final probe selection step. Such experimental validation is rarely performed for reasons of complexity and high cost.

There are two broad effectors that can contribute to both sensitivity and specificity: those related to inherent probe properties (which are the focus of this research) and those related to the kinetics and the thermodynamics of the hybridization reaction (probe concentration, target concentration, sample complexity, hybridization temperature and post hybridization washing, hybridization and post hybridization washing time, quality of target molecules and mixing of the hybridization solution) [9]. Specificity and sensitivity effectors inherent to the probe properties and target pairs include, but are not limited to: probe-target binding energy: ΔG, probe-target melting temperature: T_m, probe-target sequence similarity, probe-target matching position, probe-target contiguous match length and position, probe base composition, probe secondary structure: ΔG_{ss} and probe length. Effectors inherited to the target properties include but also not limited to: target secondary structure, labeling effects, target-target interactions, target diffusion and target length.

A microarray system has two major players: probes and targets. Probes are more controllable by the researcher, while targets, the subject of the assay, are commonly less controllable. Both contribute to the generated signal and this signal is dependent on the probes [7, 44]. Based on that the focus of this dissertation is on microarray probes only and target effects are assumed to be constant. In chapter two and chapter three of this dissertation the relationship between probe properties (probe secondary structure and probe nearest-neighbor base composition) and the hybridization background signal is examined. Effect of both on microarray data preprocessing and analysis is reported.

1.2 Probe-Target mismatches

Many applications in molecular biology require rapid and sensitive prediction of hybridization, or partial hybridization, between an oligonucleotide and potential targets in a genomic DNA or mRNA pool. Microarray technologies, PCR primer design, sequencing by hybridization, and gene diagnostic methods, are all technologies for which these predictions are very important [45-47]. Fundamentally, all of these applications rely on the specificity of hybridization, the process by which single stranded DNA oligonucleotides form stable duplexes or hybrids with a complement and/or partially complementary strand. The binding properties of mismatched oligonucleotides are less understood than those of perfectly matched oligonucleotides. Understanding the effects of mismatches at various positions in microarray probes on the extent of probe-target hybridization is critical for any microarray experiment, because microarray results are often interpreted as if they are quantitative, even when all of the variables that may affect hybridization are not well understood.

Hybridization interference that results from the presence of mismatches in the probe represents a disadvantage and an advantage in the same time. Such mismatches may degrade the sensitivity and the specificity of the probe by producing misleading results through cross hybridization with non-specific targets. On the other hand, the properties of mismatched oligonucleotides are important for various applications including the technology for detecting SNPs across the genome. To ensure there are no cross hybridization events between the probe and non-specific sequences in the hybridization pool, several approaches are taken by probe design algorithms (Table 1.1) to monitor for specificity and sensitivity of the generated probes. These approaches also suffer from the two limitations mentioned in the previous section.

Recently, Oligonucleotide Modeling Platform, a new software tool for designing and modeling primers and probes, has been released [48]. The program depends mainly on the nearest-neighbor thermodynamic modeling, an approach developed originally by Tinoco in the 1970s [49]. The nearest-neighbor (NN) model represents the covalent structure of double-stranded DNA as a collection of unique four base-quartets, each consisting of a base pair and its neighboring base pair. Each quartet makes a specific energy contribution to the overall duplex structure, and energy contributions from a string of nearest-neighbors are combined with a helix initiation parameter to predict the free energy of formation of the structure [28, 50]. The nearest-neighbor model is parameterized with an experimentally validated collection of thermodynamic parameters that account for a wide variety of structural phenomena such as mismatches, dangling ends, and loops [51]. The model also accounts for the energy needed to unfold unimolecular structures formed by each sequence in the duplex, and can be expanded to include the competition between the perfect match duplex structure and possible mismatch(es) or alternate duplex structures that might form. In this form it is known as the N-state model for the important interactions, which are known to occur within and between molecules. The nearest neighbor N-state model is known to perform well for those sequences that fall in the range of 9-30-mer, and in a solution context [28-29, 50-59]. However, since all the parameter information the software uses is derived from measurements of short sequences in solution, while microarray probes are tethered to a surface, the applicability of the software to longer microarray probes needs further investigation.

In chapter four, experimental and theoretical assessments of these mismatches are provided and used to weight the effects of mismatch number and identity on the probe design, selection and analysis. Information gathered from this part also helped validating the accuracy of the current mismatch computational hybridization modeling software packages. It is also important to mention that all of the mismatch effects were tested under different target concentrations and in the presence and absence of other sequences to clarify the effects and contributions of competitive binding between targets to perfect match and mismatch probes on hybridization signal intensity.

1.3 How much does modeling probe properties contribute to array analysis results?

Given its importance as a one of the primary tools for global gene expression assessment, microarray design and analysis are well-investigated areas at the current time. Both received tremendous attention from the scientific community, and probably every detail in this technology has been investigated. From the design steps to the analysis steps, one can find plethora of resources that address general and specific details about this technology.

The availability of resources and studies about this technology have inspired development of many models that relate microarray probe properties to the obtained hybridization signal intensity [60-61]. And it has been shown that the signal intensities obtained from microarray probes hybridized to series of target concentrations are well described by the Langmuir isotherm [62-64]. The Langmuir isotherm is a general model, which describes adsorption of a solute or a gas to a surface [65]; in a microarray reaction that conforms to this model, there is an increase in signal intensity in response to increasing target concentration until the signal reaches a saturation concentration. Consequently, many groups have attempted to employ the isotherm in the process of microarray analysis [62, 66-67]. To date, many variants of the Langmuir isotherm have been adapted to array analysis with varying degree of success [62, 66-67]. Some variants depend on probe sequence composition [62] and show acceptable results on one datasets and an unreliable output on another. Other variants depend on the free energy of the probe-target hybridization (ΔG) and the rate equation for duplex formation and melting [66] and involve extensive modeling and a large number of free parameters.

Previous approaches from other groups and from our lab have used sequence characteristics of the probe or target sequence to improve the performance of microarray analysis statistics [68-70]. Each approach introduces a set of assumptions that may be valid for certain array platforms, but not for others. For example, the signal intensity that represents transcript abundance is a composite signal collected from more than one probe, in Affymetrix experiments, while it can be a signal from a single probe in other types of microarray experiments. Significant difference of this sort between platforms can limit the applicability of models developed specifically for one type of data.

Many researchers have argued for the importance of incorporating probe specific effects into the process of microarray analysis [62, 69-70]. Effects such as probe secondary structure, probe GC content and probe binding energy are all dependent on probe sequence. Hence, probe sequence has been used in some models as the primary factor in analyzing microarray results. For example, Hekstra *et al.*, [62] used a combination of the Langmuir isotherm and probe sequence composition to estimate absolute target concentration on Affymetrix GeneChips. Held *et al*., [66] developed another analysis model based on the Langmuir isotherm and probe-target hybridization thermodynamic properties. Abdueva *et al*., [67] employed a probe-specific Langmuir isotherm model to the analysis of GeneChips, combining a non-linear minimization approach with *log* saturation intensity and *log* non-specific intensity of the probe. Those two values are predicted from probe thermodynamic properties, based also on sequence content. More recently, Li *et al*., [71] introduced a competitive hybridization model that utilizes probe-specific effects to predict probe signal intensity and to quantify absolute target concentrations.

Many models described in the literature are rather complex, and use a large number of parameters, giving rise to the potential for overfitting. Others require computationally intensive preprocessing steps. Still others can be applied to only a fraction of the available probes on the array and/or are completely platform-specific. In the last chapter of this dissertation we describe the surprising result that a simple probe-properties-independent model based on the Langmuir isotherm, with only three free parameters, performs as well as or better than all of the available models to recover solution concentrations of target based on microarray signal.

1.4 The analysis of microarray hybridization

Analysis of microarray hybridization is a multi-faceted problem with different components that add levels of complexity to it. The biophysical aspects of microarray hybridization are complicated by many interlacing factors [72]. The chemical aspects are many and not well-understood or implemented in microarray analysis procedures [73]. The statistical aspects are usually sophisticated and involve assumptions that are not always met by microarray output [74].

A microarray system is composed of two major players: probes attached to the chip surface and targets present in the hybridization mixture. Here, the focus is on the probes and target effects are assumed to be constant. Our interest in microarray probes stems from the fact that in a typical microarray experiment, the researcher has more control over the probes but little or no control of the targets, since they are the subject of the assay. Given that probes are dynamic entities, they likely to exist in an ensemble of structures with different properties that affect their behavior and the way they respond to their targets. The biophysical characteristics of microarray probe are of interest to my research group. We were interested in the behavior of secondary-structure-prone probes on microarray chips. It is likely that modeling microarray probe secondary structure using algorithms developed for in-solution nucleic acid strands [75] does not always yield accurate output [72]. There are steric interactions and electrostatic and surface forces that act on microarray probes, but not on in-solution nucleic acid strands [76]. Nonetheless, we attempted to relate the minimum folding energy (ΔG_{ss}) of microarray probes calculated using common solution folding algorithms to their signal intensity. The results of this approach didn’t provide direct evidence, or a simple rule of thumb, for predicting the contributions of a probe’s intramolecular structure to its signal intensity.

However, extending a previously published model of hybridization background signal [77] to include ΔG_{ss} resulted in 3% gain in the explanatory power of the model [78].

The above result suggested the approach of incorporating a more fundamental description of the probe's biophysical behavior into the same model [77]. When we incorporated sequence nearest neighbor (NN) information as a parameter, the model was able to explain 10% of all the variance in control and experimental data gathered using Affymetrix GeneChip arrays [70]. The enhancement gained from these approaches and the available studies on microarray hybridization analysis (see [4, 79-80] and references therein) suggested that models would benefit from incorporating probe biophysical characteristics, and that the quality of the model predictions would likely improve as the biophysical models improved. At the same time, many of the available microarray datasets lack basic information needed for proper modeling (for example: probe concentration, probe density and probe length consistency). Consequently, systematic studies of probe hybridization profiles and efficiency of the current nucleic acid modeling algorithms using those datasets were not feasible. A well-defined dataset was needed to conduct these studies. We designed and carried out a microarray experiment that could be used to study the effects of single, double and triple central mismatches on the behavior of 50-mer probes under different target concentrations, which simultaneously provided comprehensive concentration response data for all probes on the array. The results of this investigation indicated that the behavior of 50-mer probes is predictable using the current solution hybridization modeling algorithms. Probe 'percent bound' calculated by equilibrium models of solution hybridization was found to be of special importance, since it provided an easy mean to predict and assess the behavior of microarray probes.

We view microarray as a qualitative technology that has the potential to be made quantitative. Given our interest in using probe properties to provide accurate quantification of targets, our goal is to employ the results of these studies in enhancing microarray data preprocessing and analysis approaches. Many of the available algorithms for microarray analysis report "expression measure" for each target [69, 81-82]. This "measure" is just a substitution for target concentration without quantitative power attached to it. That is, target X is present/absent from the hybridization mixture or target X is more/less abundant than target Y. In other words, "expression measure" can be described as "an unknown increasing function of target concentrations, modulo statistical noise" [74]. Utilizing the Langmuir adsorption model [65] of microarray hybridization and probe percent bound to estimate absolute target concentration on microarrays showed promising results on the 50-mer dataset. Applying the same approach to estimate target concentration on different array platform (Affymetrix) was successful. We found that a combination of both the Langmuir adsorption model of microarray hybridization and probe percent bound outperformed previously published models [62, 67, 71] for microarray target concentration estimation. However, working against a null model that totally ignores probe properties resulted in the surprising finding that ignoring probe properties yielded better estimates of target concentration on all the tested microarray platforms.

The latter approach requires a training set of 5-10 genes to yield reasonable results. Genes included in the training set must follow a Langmuir-like response to their targets; otherwise the performance of the model will be degraded. Our next step will be to use a larger dataset of 50-mer probes to study the factors that prevent a probe from showing a Langmuir-like response. The same dataset will also be used to compile a kit of 5-10 genes that show a consistent Langmuir-like response under different target concentrations. This kit would be used as a training set for the developed model to enhance the quantitative power of microarray technology.

CHAPTER 2: USING PROBE SECONDARY STRUCTURE INFORMATION TO ENHANCE AFFYMETRIX GENECHIP BACKGROUND ESTIMATE

2.1 Abstract

High-density short oligonucleotide microarrays are a primary research tool for assessing global gene expression. Background noise on microarrays comprises a significant portion of the measured raw data. A number of statistical techniques have been developed to correct for this background noise. Here, we demonstrate that probe minimum folding energy and structure can be used to enhance a previously existing model for background noise correction. We estimate that probe secondary structure accounts for up to 3% of all variation on Affymetrix microarrays.

2.2 Introduction

Microarray technology holds the promise of capturing global gene expression by providing global molecular snapshots of the cell's transcriptional machinery products [83]. The ultimate goal of gene expression microarrays is to measure the abundance of each known transcript in the sample under investigation. The abundance is inferred from the signal generated by each probe as a result of a hybridization reaction with a labeled target (transcript). However, this signal

This chapter is adapted from Gharaibeh *et al*. [78]

includes background noise that not only measures the target abundance, but also non-specific binding and autofluorescence of the chip surface.

In the Affymetrix GeneChip system, each transcript's abundance is measured by a set of 11-20 probe pairs. Each pair is composed of a perfect match probe (PM), which exactly complements a region on the transcript, and a mismatch probe (MM), which is identical to the PM probe except at the 13th base, where the reverse compliment nucleotide is introduced. MM probes were originally introduced by Affymetrix to measure background noise. However, it has been shown by many groups that MM contain significant amount of PM signal and are therefore unreliable as estimators of background noise [77, 84-86]. A true estimate of background noise would improve the quality of Affymetrix GeneChip data.

Inconsistency of the signal generated from each probe is a common phenomenon in GeneChip microarray experiments [81, 87]. The differences in the signal produced can be attributed to many sources: optical noise, cross-hybridization, dye-related contributions and probe sequence composition. Many algorithms have been developed to attempt to correct for these inconsistencies [69, 82, 88]. In particular, it has been found that probe sequence composition can significantly affect the intensity of the signal generated from that probe, independent of the concentration of its target. A number of groups have suggested models where the background intensity of probes could be estimated based on their sequence composition [69, 77].

The process of nucleic acid hybridization in solution has been well studied and models such as the nearest-neighbor model provide a robust description of hybridization thermodynamics [48]. Probe-target hybridization on the microarray surface, however, does not follow the solution analogue, and the nearest-neighbor parameters that describe solution hybridization appear to be different than those for microarrays [69]. On-chip DNA hybridization is likely to be complicated by the geometric constraints of having one strand (i.e. probe) attached to the surface of the chip [89]. In addition, many other factors like probe and target secondary structure, effective reaction volume, electrostatics, diffusion and surface effects, reaction thermodynamics and kinetics, competitive binding effects, hybridization buffer composition and probe-probe interactions are believed to affect microarray DNA hybridization [76, 90].

In this chapter, the effect of predicted probe secondary structure on background hybridization noise in Affymetrix microarrays is examined. Although microarray probes are attached to the surface of the chip, they are dynamic molecules that, depending on their sequence composition, can fold onto themselves into stable secondary structure. Such stable secondary structure has the potential to interfere with probe-target hybridization [90]. Consequently, the signal obtained from such probes may not reflect the actual transcript concentrations. It has been shown, for example, that a stable secondary structure motif in a 20-mer probe dramatically decreases the final signal obtained to a point where the probe is considered insensitive to its intended target [91].

Microarray probes are usually screened for the presence of stable secondary structure either by a simple base complementarity check or using more sophisticated and time consuming energy minimization algorithms [92]. The base complementarity check is more routinely used for its simplicity and speed. Discrepancies between methods do exist, and there are no guidelines that determine which method is preferable [93]. It is therefore likely that, despite these screening procedures, a significant amount of secondary structure is present in probes in microarray experiments.

In this chapter we propose that the hybridization background noise of each probe can be modeled as a function of its sequence composition and its minimum folding energy and secondary structure. By incorporating probe secondary structure information into a previously described model of background noise [77], we improved the fit of that model to microarray data by 1-3% with minimal addition of significant free parameters.

2.3 Methods

2.3.1 Datasets

Seven datasets were used in this study (Table 2.1): the human genome U133 Latin Square dataset [94], the Choe control dataset [95], a Leukemia dataset [96], a Malaria PM only dataset [97], an Etoposide response dataset [98], a BK potassium channel knockout dataset [99-100] and an alternative splicing PM only tiling microarray dataset [101].

2.3.2 System and software

All the computational work was done on a 73-node Apple cluster. Each node is a dual 2.7 GHz PowePC G5 with 2GB RAM running Mac OSX 10.4. Secondary structure prediction was done using the *hybrid-min-ss* program of the UNAFold-2.5 software package [92]. All probes were folded as single DNA strands at 45 °C and 1.0 M sodium concentration. All other options were set to the program defaults. Simple linear model fitting and *p*-value calculations were done using R linear model function (*lm*) [102]. The Naef and Magnasco model [77] and the position-dependent secondary-structure attenuated affinity model were implemented in Perl.

TABLE 2.1 PSAA performance on different datasets.
R^2 of Naef and Magnasco model [77] (NM) and the position-dependent secondary-structure attenuated affinity model (PPSA) for the seven data sets used in this study. Results presented as average R^2 ± SD.

Dataset	na[a]	np[b]		NM	PSAA[c]
Latin Square [94]	42	248152	PM	0.17±0.009	0.184±0.010
		248152	MM	0.40±0.009	0.416±0.009
Choe [95]	6	195994	PM	0.20±0.022	0.216±0.025
		195994	MM	0.46±0.017	0.49±0.017
Leukemia [96]	72	201800	PM	0.49±0.063	0.51±0.062
		201800	MM	0.60±0.036	0.61±0.035
Etoposide response [98]	60	496468	PM	0.05±0.040	0.06±0.040
		496468	MM	0.11±0.062	0.12±0.062
BK knockout [99-100]	20	496468	PM	0.09±0.035	0.10±0.036
		496468	MM	0.29±0.050	0.30±0.049
Splicing microarray [101]	75	505916	PM	0.30±0.062	0.31±0.063
Malaria [97]	17	173262	PM	0.36±0.043	0.38±0.043

[a]na: number of chips.
[b]np: number of probes.
[c] The differences in R^2 between NM and PSAA are all statistically significant (P<10^{-3}) using paired one-sided Wilcoxon test.

2.4 Results

2.4.1 Simple linear models

The signal intensity generated from each probe can be modeled as:

$$I_j = O_j + N_j + S_j \qquad \text{Eq. 2.1}$$

Where I is the raw intensity value of probe j, O is the optical noise, N is the background noise of non-specific binding, and S is the signal generated from specific binding between probe j and its intended target [82]. Here, we do not model the specific binding signal (S) and none of our models therefore contain terms for S.

Since the *S* term, which we are ignoring in our models, is significantly higher in the PM probes than the MM probes, each probe type was modeled separately.

Controlling the GC content of the probe is one of the basic principles of microarray probe design. A probe with high GC content tends to hybridize better and to form a stable duplex with both target and non-target sequences. A simple linear model that relates probe intensity to GC content can be written as follows:

$$I_j = B0 + B1\langle GC\rangle_j + \varepsilon_j \qquad \text{Eq. 2.2}$$

where *I* is the raw intensity of probe *j*, $\langle GC\rangle_j$ is the number of GC nucleotides in probe *j* (which is a number between 0 and 25), *B0* and *B1* are free parameters and ε_j is an error term. The model explains a modest amount of the overall intensity when applied to the Latin Square data set; $R^2 \approx 0.02$ for PM and 0.12 for MM (Fig. 2.1).

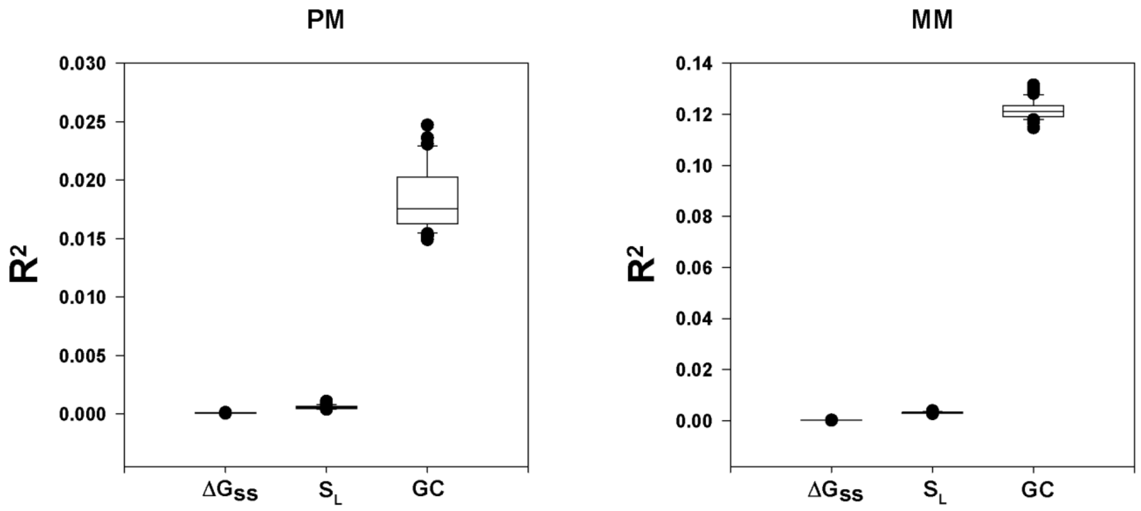

FIGURE 2.1 R^2 distribution for the simple linear models (Eqs 2.2, 2.3 and 2.4) for all the U133 Latin Square chips.
The null hypothesis that the *B1* parameter is equal to zero is rejected with high confidence ($P < 10^{-4}$) for all the models.

The model explains more of the MM probes intensity because most of the signal obtained from MM probes is background noise. MM intensity is therefore more independent of the concentration of the target gene.

We wondered, compared to the GC content, how much of the background noise probe secondary structure would explain when put into a simple linear model. The free energy of probe secondary structure formation (ΔG_{ss}) is an indicator of the stability of secondary structure in which the probe folds on itself. The more stable the secondary structure, the less a probe will be able to hybridize to its target or non-target sequences. As a result, one would expect to observe a low signal from such probe.

How much of all probe variance can be explained directly by secondary structure predictions? A simple linear model is:

$$I_j = B0 + B1\langle \Delta G_{ss} \rangle_j + \varepsilon_j \qquad \text{Eq. 2.3}$$

where $\langle \Delta G_{ss} \rangle_j$ is probe *j* minimum folding energy in Kcal/mol.

If we apply this simple linear model to the Latin Square data set, we find a very low r-squared values; $R^2 < 10^{-4}$ (Fig. 2.1). However, the *p*-value of the null hypothesis that the *B1* parameter is equal to zero is rejected with high confidence (Fig. 2.1). These data suggest that there is a statistically significant influence of ΔG_{ss} on the observed intensity, although this relationship does not explain very much of the overall intensity on the array.

One may argue that the low *r-squared* values in equation 2.3 are due to the fact that ΔG_{ss} value does not reflect the size of the secondary structure motif found in that probe and the number of free bases available for hybridization. The program hybrid-min-ss reports for the most stable secondary structure of a probe whether a given nucleotide is involved in secondary structure formation or not. We can define a value, S_L, which is the longest stretch of nucleotides that are not involved in secondary structure formation (for example, S_L =10 in Fig. 2.2).

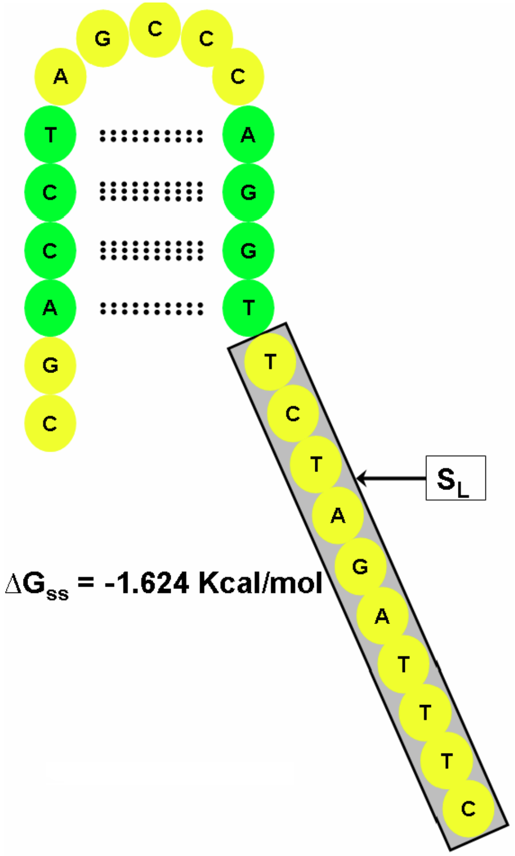

FIGURE 2.2 Folded probe showing its sequence, minimum folding energy (*ΔG*$_{ss}$) and minimum energy structure.
The longest free string of bases of the folded probe (S_L) is shown in gray box; bases involved in hydrogen bonding are shown in green ovals.

To investigate the relationship between the longest free string of bases of the folded probe (S_L) and the observed intensity, we can again apply a simple linear model:

$$I_j = B0 + B1\langle S_L \rangle_j + \varepsilon_j \qquad \text{Eq. 2.4}$$

where $\langle S_L \rangle_j$ is the longest free string of bases in probe *j* based on its minimum energy structure. This model also has a very low r-squared values when applied to the Latin Square data set; $R^2 < 10^{-3}$ (Fig. 2.1), only slightly higher than Eq. 2.3

suggesting little direct effects of probe S_L on the observed intensity. However, the *p*-value of the null hypothesis that the *B1* parameter is equal to zero is also rejected with high confidence (Fig. 2.1). We see that GC content can explain a modest amount of overall intensity. Models based on secondary structure explain much less of the intensity data, although they are still highly statistically significant.

2.4.2 Position-dependent secondary-structure attenuated affinity model (PSAA)

Since the three simple linear models (equations 2.2, 2.3 and 2.4) all hold significant relationships with the observed intensity (Fig. 2.1), we wanted to combine them into one model that takes into account the base composition, ΔG_{ss} and S_L of the probe. We found that a simple linear combination of GC, ΔG_{ss} and S_L did not significantly improve on the power of the individual models (data not shown). We reasoned that a model that is aware of each probe's base position and involvement in the overall secondary structure of the probe would outperform models that ignore this information.

The model of Naef and Magnasco [77] provides a starting point that meets our requirement for individual base information. In this model, probe background is modeled based on sequence composition:

$$\ln\left\langle B/M\right\rangle = \sum_{k=1}^{25} \sum_{l\in(A,T,C,G)} S_{lk}A_{lk} \qquad \text{Eq. 2.5}$$

where *B* is the raw probe intensity, *M* is the median intensity of the array, *l* is nucleotide index, *k* is the position of *l* along the probe, *S* a Boolean variable equals to

1 if the probe sequence has *l* at *k* and zero otherwise, and *A* is the per-site per letter affinity. To clarify, consider the probe shown in Fig. 2.2, then equation 2.5 will read:

$$\begin{aligned}
&\ln\langle B/M\rangle = \\
&(S_{1G}\times A_{1G})+(S_{1A}\times A_{1A})+(S_{1T}\times A_{1T})+(S_{1C}\times A_{1C})+ \\
&(S_{2G}\times A_{2G})+(S_{2A}\times A_{2A})+(S_{2T}\times A_{2T})+(S_{2C}\times A_{2C})+ \\
&(S_{3G}\times A_{3G})+(S_{3A}\times A_{3A})+(S_{3T}\times A_{3T})+(S_{3C}\times A_{3C})+ \\
&\cdots\cdots\cdots\cdots \\
&+(S_{25G}\times A_{25G})+(S_{25A}\times A_{25A})+(S_{25T}\times A_{25T})+(S_{25C}\times A_{25C})
\end{aligned}$$

$$\ln\langle B/M\rangle = A_{1C}+A_{2G}+A_{3A}+\cdots+A_{25C}$$

Equation 2.5 is a simple model that has four free parameters for each probe base (100 free parameters for a 25-base probe). The values of these 100 free parameters are generated by linear least squares fit [77]. Given the large number of probes on each chip (about half a million for the human genome U133 chip, for example) over-fitting is not a concern.

In our approach, we add the continuous variable θ to reflect the involvement of the probe nucleotides in secondary structure formation. The model now is written as:

$$\ln\langle B/M\rangle = \sum_{k=1}^{25}\sum_{l\in(A,T,C,G)} \theta\, S_{lk}A_{lk} \qquad \text{Eq. 2.6}$$

The θ term reflects the degree to which an individual probe base participates in secondary structure formation. In our model, it is represented by any value between 0 and 1. There are a large number of ways in which values for θ could be generated. We made the following simplifying assumptions. We begin by considering nucleotides that are not involved in secondary structure formation. In cases where a probe's ΔG_{ss} > 0 Kcal/mol we can set θ for all bases within that probe to 1. Likewise, when a base within a probe is not involved in secondary structure hydrogen bonding (yellow ovals in Fig. 2.2, for example), we can set θ to 1 for that base. To calculate θ for the remaining bases, we set a $\Delta G_{ss\text{-}cutoff}$ value below which θ will be constant (θ = *tb*). We assumed that the relationship between θ and ΔG_{ss} is linear in the region where ΔG_{ss} is between $\Delta G_{ss\text{-}cutoff}$ and 0 (Fig. 2.3).

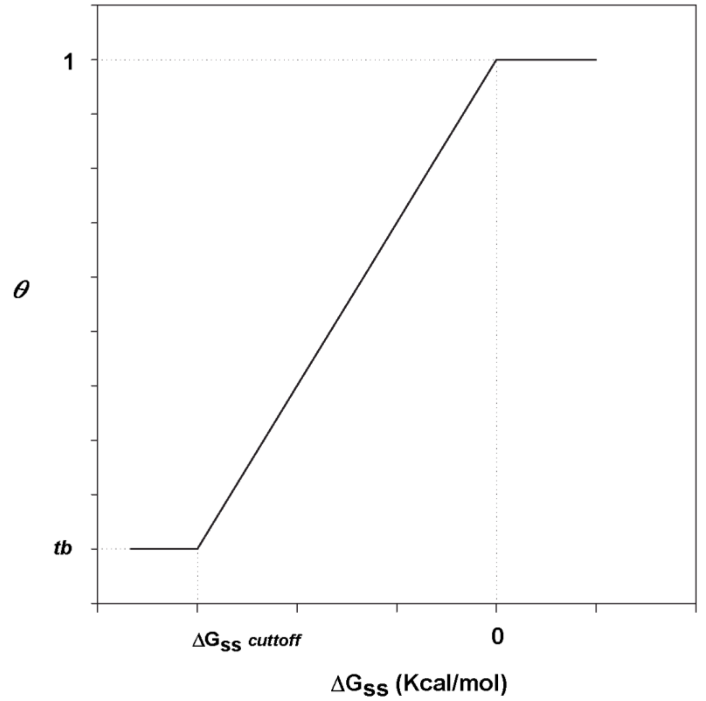

FIGURE 2.3 A model for the relationship between ΔG_{ss} and θ for the bases involved in secondary structure formation.

From the assumption of linearity, we can derive a slope and an intercept to yield:

$$\theta = \frac{tb - 1}{\Delta G_{ss-\mathrm{cutoff}}} \langle \Delta G_{ss} \rangle_j + 1 \qquad \text{Eq. 2.7}$$

This equation has two unknown parameters *ΔG_{ss-cutoff}* and *tb*. To find the best values for these parameters, we tested the effects of changing *ΔG_{ss-cutoff}* and *tb* on the performance of the model (equation 2.6) on a single chip from the Latin Square dataset. We found that the best performance of the model was obtained at *ΔG_{ss-cutoff}* = -3.6 Kcal/mol and a *tb* = 0.35 (Fig. 2.4).

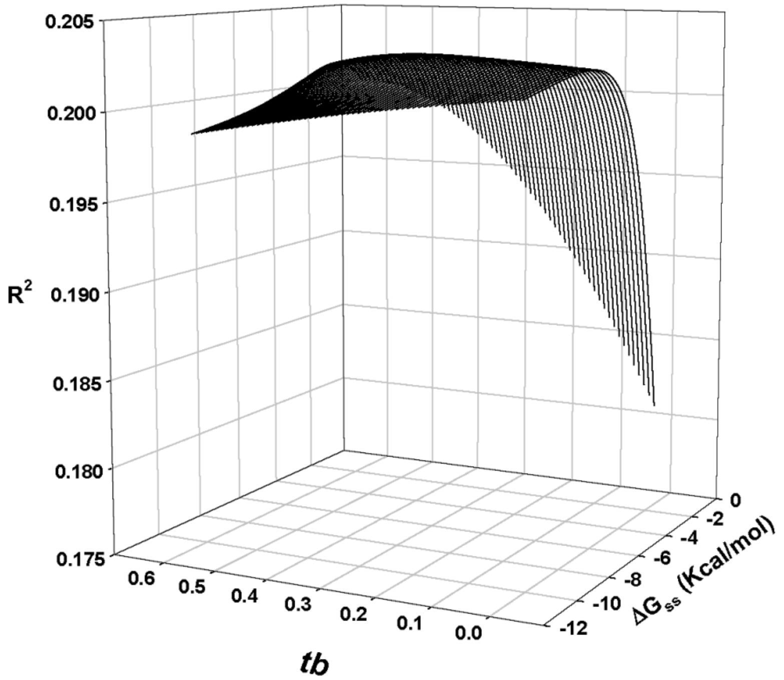

FIGURE 2.4 Optimizing $\Delta G_{ss\text{-}cutoff}$ and *tb*.
Effects of changing the values of $\Delta G_{ss\text{-}cutoff}$ and *tb* on the performance of the position-dependent secondary-structure attenuated affinity model using the human genome U133 Latin Square Experiment 2 Replicate 1 PM probes.

To summarize, we define our position-dependent secondary-structure attenuated affinity model (PSAA) as equation 2.6**,** where *B* is the raw probe intensity, *M* is the median intensity of the array, *l* is letter index, *k* is the position of *l* along the probe, *A* is the per-site per letter affinity, *S* a Boolean variable equal to 1 if the probe sequence has *l* at *k* and zero otherwise, and *θ* is:

a. 1, if the probe $\Delta G_{ss} > 0$ Kcal/mol.

b. 1, if the probe $\Delta G_{ss} \leq 0$ Kcal/mol and *l* is not involved in secondary structure hydrogen bonding.

c. $\begin{pmatrix} 0.35, & if\,\Delta G_{ss} \leq -3.6 \\ \frac{0.65}{3.6}\langle \Delta G_{ss} \rangle_j + 1, & if -3.6 < \Delta G_{ss} \leq 0 \end{pmatrix}$ and *l* is involved in secondary structure hydrogen bonding.

Here, the involvement of each probe base in secondary structure hydrogen bonding is based on its minimum energy structure.

When we consider the folded probe presented in Fig. 2.2, equation 2.6 reads:

$$\begin{aligned}
&\ln\langle B/M \rangle = \\
&\theta\ ((S_{1G} \times A_{1G}) + (S_{1A} \times A_{1A}) + (S_{1T} \times A_{1T}) + (S_{1C} \times A_{1C})) + \\
&\theta\ ((S_{2G} \times A_{2G}) + (S_{2A} \times A_{2A}) + (S_{2T} \times A_{2T}) + (S_{2C} \times A_{2C})) + \\
&\theta\ ((S_{3G} \times A_{3G}) + (S_{3A} \times A_{3A}) + (S_{3T} \times A_{3T}) + (S_{3C} \times A_{3C})) + \\
&\cdots\cdots\cdots\cdots \\
&+\theta\ ((S_{25G} \times A_{25G}) + (S_{25A} \times A_{25A}) + (S_{25T} \times A_{25T}) + (S_{25C} \times A_{25C}))
\end{aligned}$$

$$\ln\langle B/M \rangle = A_{1C} + A_{2G} + 0.64A_{3A} + \cdots + A_{25C}$$

The model defined in equation 2.6 was fitted to all the datasets (Table 2.1). The fitting was done on the PM and MM probes separately. Table 2.1 shows a comparison between the native Naef and Magnasco model [77] and our position-dependent secondary-structure attenuated affinity model. We see that including probe secondary structure information improved the fit of the native Naef and Magnasco model [77] by 1-3%, depending on the chip and probe type. Note that all the models (equations 2.2, 2.3, 2.4, 2.5 and 2.6) perform better on the MM probes due to the higher background noise present in the MM signal.

2.4.3 Gains in performance cannot be trivially explained by additional free parameters

We note that there are two distinct kinds of free parameters in our model. The 100 free parameters from the original Naef and Magnasco model (equation 2.5) are calculated for each chip by linear least squares fit. We have added two free parameters in equation 2.6, $\Delta G_{ss\text{-}cutoff}$ and *tb*. These parameters were determined from one of the Latin Square dataset chips from the curves shown in Fig. 2.4 and were held constant for all the datasets in this chapter. Given that our fits contain between 173,262 and 496,468 data points (Table 2.1), it seems unlikely that the improvements in performance could be explained by the addition of the free parameters $\Delta G_{ss\text{-}cutoff}$ and *tb*. Nonetheless, to further rule out this possibility, we refolded the Latin Square data set probes with either a completely random sequence (generated with an equal probability of A, C, G and T) or a shuffled sequence. Then we fed equation 2.6 the original probe sequence (i.e. the right *l* at *k*) along with the new ΔG_{ss} and the new minimum energy structure that resulted from folding the random or shuffled probe sequence. For the random sequence case, the performance of the original Naef and Magnasco model [77] was severely degraded as shown in Figure 2.5.

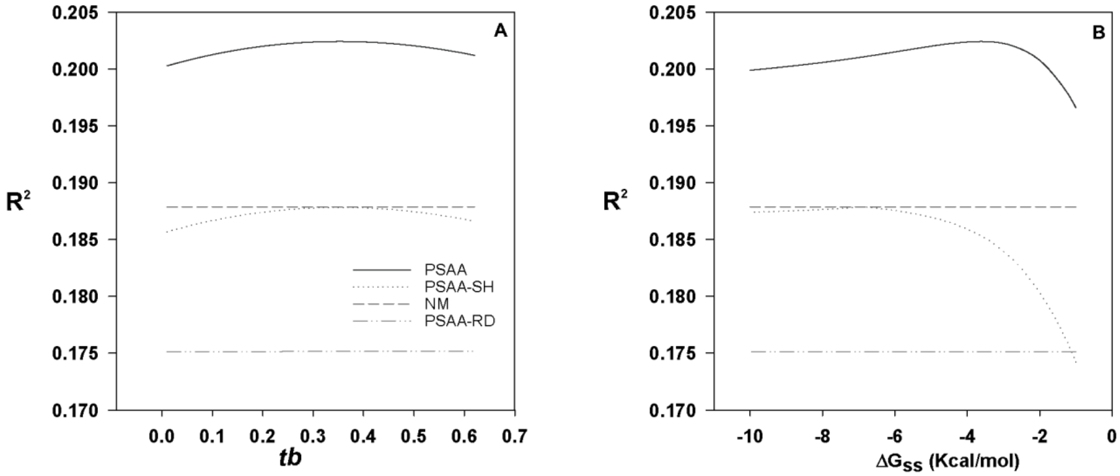

FIGURE 2.5 Effects of changing *tb* or $\Delta G_{ss\text{-}cutoff}$ on the performance of the PSAA model.
Effects of changing (A) the values of *tb* while holding $\Delta G_{ss\text{-}cutoff}$ = -3.6 or (B) the values of $\Delta G_{ss\text{-}cutoff}$ while holding *tb* = 0.35 on the performance of PSAA: the position-dependent secondary-structure attenuated affinity model (equation 2.6). Data shown are for the human genome U133 Latin Square Experiment 2 Replicate 1 PM probes. NM: Naef and Magnasco [77] model (equation 2.5). The suffixes (-SH) and (-RD) indicates the R^2 after generating the minimum folding energy (ΔG_{ss}) and the minimum energy structure from shuffled and random sequences, respectively (see section 2.4.3 for explanation).

For the shuffled sequences, the probe's base composition is not affected, but the position of each base has been changed due to the shuffling process. For the shuffled sequences, the fit of the model dropped down to that of the original Naef and Magnasco model [77]. These results on shuffled and random sequence show that the presence of the two additional free parameters $\Delta G_{ss\text{-}cutoff}$ and *tb* cannot by themselves explain the improved performance over the original Naef and Magnasco model [77]. This strongly supports our argument that the gain in the *r-squared* values of our model came from including probe secondary structure information and do not arise trivially from the addition of free parameters.

2.5 Discussion

In the absence of a clear understanding of the microarray hybridization mechanisms and the frequent use of probes that fold into stable secondary structure under the hybridization conditions on microarrays, a model is needed to explain or approximate the effects of such behavior on microarray signal. Using simple linear models, we saw a modest relationship ($R^2 < 10^{-3}$) between probe intensity and its ΔG_{ss} or S_L. We propose as a more powerful alternative to two parameter linear models, a modification of the Naef and Magnasco model [77] to include probe secondary structure effects on the background intensity. Our model works by equating an increase in secondary structure with a decreased contribution to a linear least square fit. If a particular base is involved in secondary structure hydrogen bonding (Fig. 2.2), we assign it a low θ score depending on the overall ΔG_{ss} of the probe (Eq. 2.7). Consequently, this base contribution is attenuated in the Naef and Magnasco model [77]. Consider, for example, the third adenine base in the folded probe presented in Fig. 2.2, in the Naef and Magnasco model [77] its contribution to the brightness is A_{3A}. Based on the predictions of hybrid-min-ss, this base is involved in secondary structure hydrogen bonding and we therefore expect a reduced contribution to the intensity caused by background binding. We therefore attenuate its contribution to the brightness by θ, and its contribution now is $0.64A_{3A}$ instead of A_{3A}. Results attenuated by θ have more power than the original Naef and Magnasco model [77] over a wide range of Affymetrix datasets (Table 2.1).

The secondary structure information used here is based on the minimum folding energy (ΔG_{ss}) and the minimum energy structure, as predicted by an energy minimization algorithm [92] that uses the nearest-neighbor parameters [28] to predict secondary structure of single-stranded DNA molecules in solution. In the absence of clear understanding of the effects of the geometric constraints of attaching one end of the DNA probe to the chip surface on its secondary structure, the nearest-neighbor parameters represent a reasonable approximation for microarray [66, 80]. We are also fully aware that single-stranded DNA molecules are highly dynamic and each molecule is likely to exist in an ensemble of structures. Based on that, predicting the minimum folding energy (ΔG_{ss}) and the minimum energy structure for any single-stranded DNA molecule can be different when using different prediction algorithms, even when the same folding conditions are used. The results presented here are based on the minimum folding energy (ΔG_{ss}) and the minimum energy structure calculated using UNAFold [92]. It has been shown that the differences in the predicted minimum folding energy (ΔG_{ss}) and the minimum energy structure between different prediction algorithms are small [103-104]. Consequently, we would expect similar results no matter which of the currently popular secondary structure prediction algorithms were used.

The results presented in this chapter suggest that, on average, 1-3% of all the intensities on Affymetrix GeneChip microarrays can be explained by probe secondary structure independent of any target information. Given that not all the probes form stable secondary structure (50% of the human genome U133 Latin Square dataset probes, for example have predicted $\Delta G_{ss} > 0$), the 1-3% enhancement over the original model is quite satisfactory, and represent a step forward in understanding the factors that affect the on-chip hybridization process.

The current design of GeneChip microarrays devotes half of the chip to MM probes. The sole purpose of these probes is to estimate the background noise portion present in the PM signal to enhance the chip ability to detect differently expressed genes. Advances in the ability to correctly estimate background noise on Affymetrix GeneChip microarrays based on probe sequence information may in the future eliminate the need of MM probes on these arrays offering more space to interrogate more genes on the same array.

CHAPTER 3: BACKGROUND CORRECTION USING DINUCLEOTIDE AFFINITIES IMPROVES THE PERFORMANCE OF GCRMA

3.1 Abstract

High-density short oligonucleotide microarrays are a primary research tool for assessing global gene expression. Background noise on microarrays comprises a significant portion of the measured raw data, which can have serious implications for the interpretation of the generated data if not estimated correctly. We introduce an approach to calculate probe affinity based on sequence composition, incorporating nearest-neighbor (NN) information. Our model uses position-specific dinucleotide information, instead of the original single nucleotide approach, and adds up to 10% to the total variance explained (R^2) when compared to the previously published model. We demonstrate that correcting for background noise using this approach enhances the performance of the GCRMA preprocessing algorithm when applied to control datasets, especially for detecting low intensity targets. Modifying the previously published position-dependent affinity model to incorporate dinucleotide information significantly improves the performance of the model. The dinucleotide affinity model enhances the detection of differentially expressed genes when implemented as a background correction procedure in GeneChip preprocessing algorithms. This is conceptually consistent with physical

This chapter is adapted from Gharaibeh *et al*. [70]

models of binding affinity, which depend on the nearest-neighbor stacking interactions in addition to base-pairing.

3.2 Introduction

Affymetrix GeneChip arrays are one of the most popular gene expression array systems used by researchers worldwide [83]. The purpose of an expression microarray experiment is to measure the abundance of each known transcript in the sample under investigation. Abundance is inferred from the signal generated by a set of 11-20 probe pairs. Each pair is composed of a perfect match probe (PM), which exactly complements a region on the transcript, and a mismatch probe (MM), which is identical to the PM probe except at the 13th base, where the reverse complement nucleotide is introduced [105]. The fluorescent signal from each probe, however, includes background noise that not only measures the transcript abundance, but also non-specific binding (NSB) and autofluorescence of the chip surface. MM probes were originally introduced by Affymetrix to measure background noise. It has been shown by many groups that MM probes contain significant amount of the PM signal and are therefore unreliable as estimators of background noise [84-86].

A gene expression experiment using the Affymetrix GeneChip system usually involves a design step, a preprocessing step, an inference step and finally, a validation step [106]. The preprocessing step is of special importance; preprocessing transforms the raw fluorescence signals from each probe in a probeset into a composite gene expression value. The main goal of the preprocessing step is to remove non-biological variation from the raw data [106]. Usually, the preprocessing step in Affymetrix GeneChip array analysis includes three main treatments of the raw data. A background adjustment step separates the specific signal from the non-specific signal. A probe-level normalization step then removes non-biological variation between arrays. Finally, a summarization step generates a single expression value for each gene from its corresponding probeset. The method described in this manuscript is an implicit physical model that modifies the background adjustment step.

Background noise and non-biological variation of the signal generated from each probe are common phenomena in GeneChip microarray experiments [81, 87]. The differences in the signal produced can be attributed to many sources: optical noise, cross-hybridization, dye-related contributions and probe sequence composition. Many preprocessing algorithms have been developed in an attempt to correct for these artifacts [88]. According to Allison *et al*. [106] there is no clear winner among the available preprocessing algorithms. However, GCRMA [68], a modification of RMA [107], often performs as well as or better than other algorithms [88, 108-110]. GCRMA incorporates probe sequence composition into background adjustment, following the physical model of Naef and Magnasco [77]. The model describes a probe affinity that is dependent on its base composition and the position of each base along the probe and suggests that probe sequence can significantly affect the intensity of the signal generated from that probe, independent of the concentration of its target.

Performance assessment of GCRMA has been done using both spike-in [95, 108, 110-111] and real [109] datasets followed by quantitative real time PCR confirmation . So far, a number of reports have been published recommending the use of GCRMA for detecting differentially expressed genes and estimating relative expression, emphasizing its outstanding performance in detecting low-intensity, differentially expressed genes [110-111]. When comparing microarray analysis algorithms, Irizarry *et al*. [88] have argued for an approach that balances accuracy and precision. Irizarry *et al*., define accuracy as the ability of the algorithm to detect the relative expression of a transcript without bias to its abundance (concentration). They define precision as low variance; this is characterized by a steady performance on replicates of the same sample. GCRMA is among the few preprocessing algorithms that scores well in both accuracy and precision [110].

In this study, we modified the portion of GCRMA derived from the model of Naef and Magnasco [77] to calculate probe affinity using position-specific dinucleotide information. The dinucleotide is a fundamental chemical unit that contributes a well-understood component to nucleic acid duplex stability and to the free energy of duplex formation during hybridization [28, 48]. We applied the new model to different datasets, and achieved an improved fit to microarray data with R^2 increasing by 5-10%. Then, we tested the downstream effect of our modified background model on the performance of GCRMA in detecting differentially expressed genes, when used to analyze two publicly available control datasets: the human genome U133 Latin Square dataset [94] and the golden spikein dataset [95].

In both datasets, application of the dinucleotide model in background correction improved the detection of differentially expressed genes. Therefore, we propose that probe affinity be modeled based on dinucleotide composition of the probe instead of the original single nucleotide approach.

3.3 Methods

3.3.1 Datasets

The U133 Latin square dataset

This dataset is composed of 14 experiments (three technical replicates for each experiment) in which 42 transcripts are spiked at a concentration range of 0.125-512 pM following a Latin square design. The dataset files were downloaded from Affymetrix web site [94]. For Affycomp analysis, all probesets were included. For the 14 2X comparisons the following probesets were excluded following Affymetrix recommendations: 209374_s_at, 205397_x_at, 208010_s_at. In addition, we excluded any probesets with a name starting with AFFX- that was not included in the 42 true positive spikeins.

The Golden spikein dataset

This dataset has more spikein genes than the Latin Square dataset, but consists of only six microarrays, 3 C (control) and three S (spikein) [95]. The S pool contains cRNA at concentration equal to or higher than the C pool [95]. Each pool was hybridized to the Affymetrix *Drosophila* array (three technical replicates for each hybridization). Probesets measuring spikein transcripts were determined based on the analysis of Schuster *et al*. [111]. We considered all probeset that measure differentially expressed genes to be true positives (a total of 1353 probesets).

Several issues have been raised concerning the use of the Golden spikein dataset in validating GeneChip preprocessing algorithms [112-114]. However, the analysis of Pearson [115] shows clearly that the Golden spikein dataset can be used to validate and compare the performance of GeneChip preprocessing algorithms.

3.3.2 Model implementation

The single nucleotide model was implemented in Perl [78], the dinucleotide model was implemented in Java. All the models were fitted using the least squares method. The fitted parameters for the dinucleotide model for each of the two datasets were used to generate an affinity.info matrix for that dataset. This affinity.info matrix was used in GCRMA analysis later on. Affinity.info matrix generation was done using a local R script following the steps found in GCRMA source code (available here [116]). The Java code for the dinucleotide model is provided here [116].

3.3.3 Data analysis

All analysis steps were performed using R [102] version 2.5.0 and Bioconductor [117] unless otherwise indicated.

Expression summaries were generated using the full model of GCRMA version 2.8.1. The commands used to generate the summaries for GCRMA-NN, GCRMA-L and GCRMA-R can be found here [116]. The affinity.info matrix for the U133 Latin square dataset and the Golden spikein dataset affinity.info matrix can be found here [116].

Affycomp analysis was done using a locally installed Affycomp 1.14.0 package. All expression summaries were converted back from the log scale to the original scale and formatted to a comma-delimited text files using a local Perl script. Metrics generation for the expression summaries was done using a local R script following the directions of the package maintainers. The following metrics were used to evaluate the performance of each algorithm (definitions are according to Affycomp website [118]): Median SD is the median standard deviation across replicates. It measures the consistency of the algorithm (variance across the range of expression levels); the lower the median SD the more consistent the algorithm. Null log-fc IQR and null log-fc 99.9% are the interquartile range and the 99.9th percentile of the log fold changes from probesets, for genes that should not change. A perfect score is 0 for both metrics. Signal detect slope is the slope obtained from regressing expression values on nominal concentrations in the spikein data. Signal detect R^2 is the R squared obtained from regressing expression values on nominal concentrations in the spikein data. Low.slope, med.slope and high.slope are as in signal detect slope, but for probesets targeting low (<4 pM), medium (4-32 pM) and high (> 32 pM) spikeins, respectively. Obs-intended-fc and Obs-(low)int-fc slopes are slopes obtained from regressing observed log fold changes against nominal log fold changes for all probesets, and for those with nominal concentration less than 2 pM, respectively. Low, med and high AUC reflect the area under the ROC curve (with up to 100 false positives) for spikeins with low, medium and high intensities, standardized so that optimum is 1, respectively. Weighted avg AUC is the weighted average of the previous three ROC curves with weights related to amount of data in each class (low, medium and high).

ROC curve generation was implemented in Java and cyber *t* analysis was done in R. Detailed description of the implementation and the analysis can be found here [98].

3.4 Results

3.4.1 Dinucleotide affinity model

Naef and Magnasco [77] model probe affinity (probe hybridization effect) based on sequence composition as follows:

$$\ln\left\langle B/M\right\rangle = \sum_{k=1}^{25}\sum_{l\in(A,T,C,G)} S_{lk}A_{lk} \qquad \text{Eq. 3.1}$$

where *B* is the raw probe intensity, *M* is the median intensity of the array, *l* is the nucleotide index (A,C,G or T), *k* is the position of *l* along the probe (note that *k* has a range of 1 to sequence length, that is 25 for GeneChip probes), *S* is a Boolean variable equal to 1 if the probe sequence has *l* at *k* and zero otherwise, and *A* is the per-site-per- nucleotide affinity. As an example, consider the following sequence: CGAC, for which equation 3.1 reads:

$$\begin{aligned}&\ln\left\langle B/M\right\rangle = \\ &(S_{1G}\times A_{1G})+(S_{1A}\times A_{1A})+(S_{1T}\times A_{1T})+(S_{1C}\times A_{1C})+ \\ &(S_{2G}\times A_{2G})+(S_{2A}\times A_{2A})+(S_{2T}\times A_{2T})+(S_{2C}\times A_{2C})+ \\ &(S_{3G}\times A_{3G})+(S_{3A}\times A_{3A})+(S_{3T}\times A_{3T})+(S_{3C}\times A_{3C})+ \\ &(S_{4G}\times A_{4G})+(S_{4A}\times A_{4A})+(S_{4T}\times A_{4T})+(S_{4C}\times A_{4C})\end{aligned}$$

$$\ln\left\langle B/M\right\rangle = A_{1C}+A_{2G}+A_{3A}+A_{4C}$$

Equation 3.1 is a simple model that has four free parameters for each probe base (100 free parameters for a 25-base probe). The values of these 100 free parameters are generated by linear least squares fit. Given the large number of probes on each chip (about half a million for the human genome U133 chip, for example) over-fitting is not a concern.

Figure 3.1 shows the 25 parameters (term A in Equation 1) of the four nucleotides as a function of their position along the probe for the U133 Latin square dataset (parameters derived from a single chip are shown in panel A and an average of the parameters across all the 42 chips is shown in panel B). A similar pattern of parameters have been obtained fitting equation 1 to other Affymetrix datasets (data not shown and [77]). These fitted per-site-per-nucleotide affinities imply that the signal generated from each probe will be affected by the probe sequence. Consider two probes interrogating two transcripts, which are present in identical concentration. In such a case, a probe containing many adenines (A) will produce a lower signal intensity than the probe with many cytosines (C), especially if the As or Cs are concentrated at or near the center of the probe (position 13).

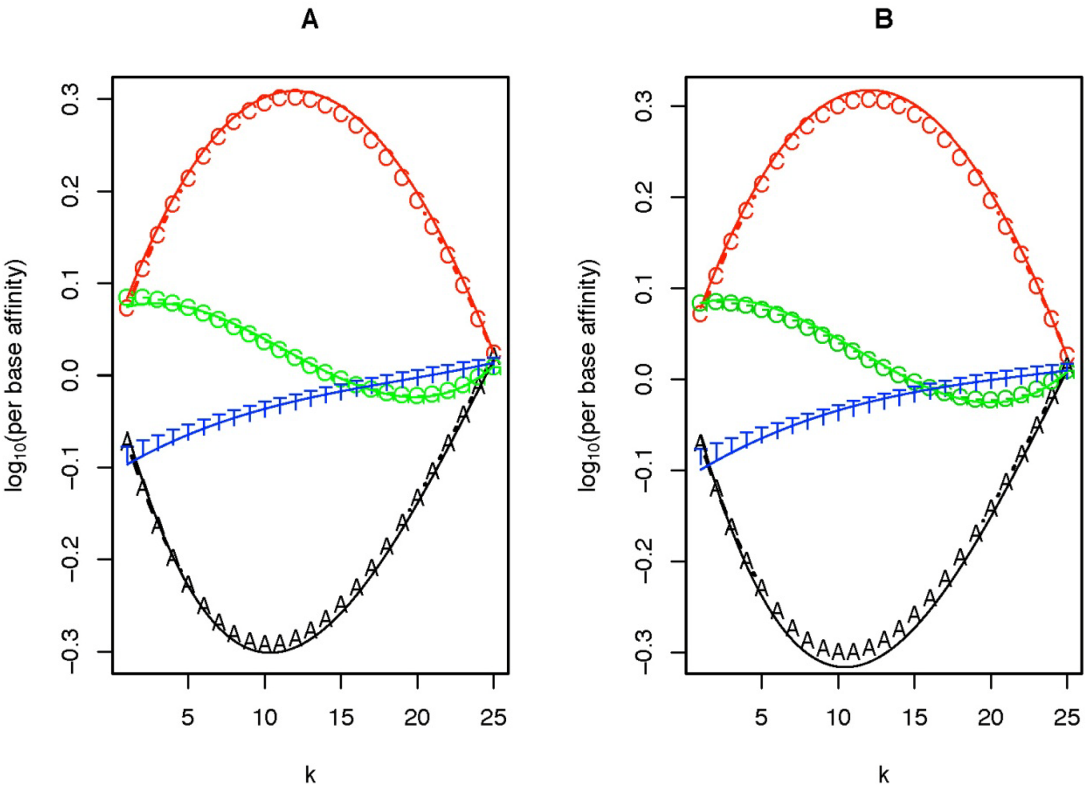

FIGURE 3.1 Affinity parameters calculated using single nucleotide model. Affinity parameters calculated using equation 3.1 for the human genome U133 Latin Square. Panel A is for Experiment 11 Replica 2 and panel B shows an average of the parameters across all the 42 chips. K represents the position of each nucleotide along the probe length. Affinity parameters calculated using equation 3.2 are shown as solid lines. Higher affinity (Y-axis) indicates brighter signal.

The model defined in equation 3.1 can also be expressed as a polynomial of degree 3, thus reducing the free parameters from 100 to 16 as shown below:

$$\ln\left\langle B/M\right\rangle = \sum_{k=1}^{25} \sum_{l\in(A,T,C,G)} \sum_{t=0}^{3} S_{lk} A_{lt} k^{t} \qquad \text{Eq. 3.2}$$

By assuming the affinities can be modeled as a third order polynomial function of position, the number of free parameters in the model can be reduced from 100 to 16 with little loss of predictive accuracy as the polynomial generated with 16 parameters (Fig. 3.1 solid lines) closely matches the 100 independently estimated parameters (Fig. 3.1 symbols) and the R^2 of both models are similar (Fig. 3.2).

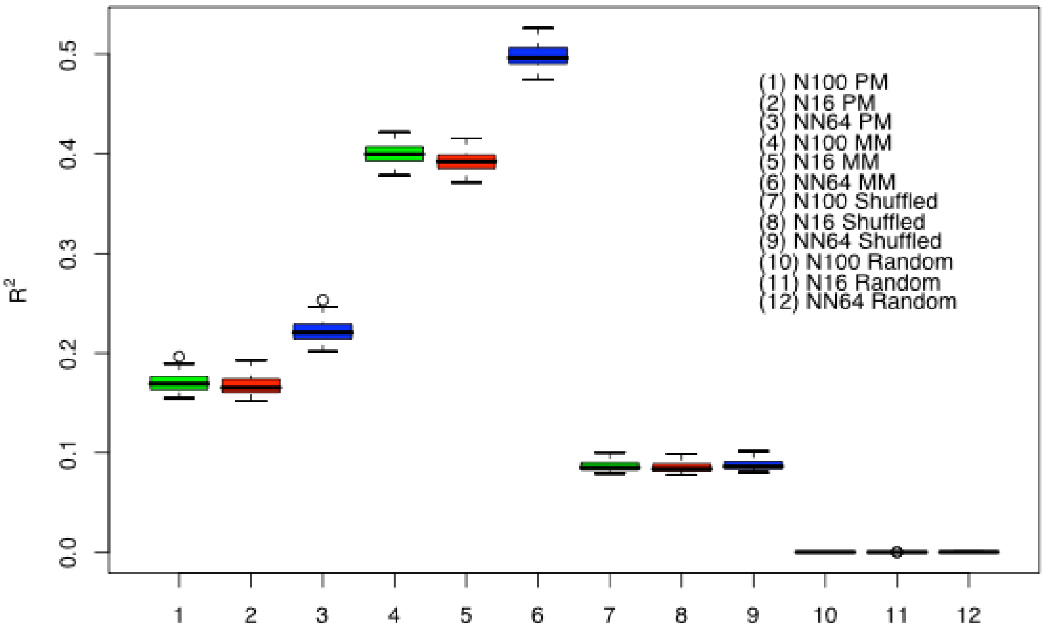

FIGURE 3.2 Box plots of the R^2 of the single nucleotide model and the dinucleotide model.

Box plots showing the R^2 of the single nucleotide model (N) (using the 100 free parameters (N100), equation 3.1, and the 16 free parameters (N16), equation 3.2) and the dinucleotide model with 64 free parameters (NN 64), equation 3.4 on the 42 Latin square chips. PM indicates the fit was done on the perfect match probes, MM indicates the fit was done on the mismatch probes, shuffled indicates the fit was done on the shuffled probe sequences and random indicates the fit was done on randomly generated probe sequences.

In the dinucleotide model, we follow a similar strategy to the above, but we model composition-biased probe affinity using dinucleotides (pairs of adjacent bases), which are a fundamental chemical unit in physical models of nucleic acid folding and hybridization rather than single nucleotides. The dinucleotide model is as follows:

$$\ln\left\langle B / M \right\rangle = \sum_{k=1}^{24} \sum_{l \in NN} S_{lk} A_{lk} \qquad \text{Eq. 3.3}$$

where *B* is the raw probe intensity, *M* is the median intensity of the array, *l* is the NN nucleotide pair (AA, AC AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG or TT), *k* is the position of *l* along the probe (note that *k* has a range of 1 to sequence length minus one, that is 24 for GeneChip probes), *S* is a Boolean variable equal to 1 if the probe sequence has *l* at *k* and zero otherwise, and *A* is the per-site-per-dinucleotide affinity. We then again assume that the per-site-per-dinucleotide affinity follows a polynomial of degree 3 as a function of the position *k* as outlined in equation 3.4:

$$\ln\left\langle B / M \right\rangle = \sum_{k=1}^{24} \sum_{l \in NN} \sum_{t=0}^{3} S_{lk} A_{lt} k^{t} \qquad \text{Eq. 3.4}$$

This reduces the number of free parameters from 384 (16 dinucleotides x 24 nucleotide positions, equation 3.3) to 64 (16 dinucleotides x 4 parameters, equation 4), which makes this approach computationally feasible. As an example, consider the following sequence: CGAC (three dinucleotides: CG for *k*=1, GA for *k*=2, and AC for *k*=3), for which equation 3.4 reads:

$$\ln\left\langle B/M\right\rangle =$$
$$(A_{CG0})+(A_{CG1}\times 1)+(A_{CG2}\times 1^2)+(A_{CG3}\times 1^3)+$$
$$(A_{GA0})+(A_{GA1}\times 2)+(A_{GA2}\times 2^2)+(A_{GA3}\times 2^3)+$$
$$(A_{AC0})+(A_{AC1}\times 3)+(A_{AC2}\times 3^2)+(A_{AC3}\times 3^3)+$$

$$\ln\left\langle B/M\right\rangle = 4A_{CG}+15A_{GA}+40A_{AC}$$

Note that we do not explicitly fit the stacking energies of the NN pairs; rather we explicitly fit the NN pairs' affinities along the probe sequence position. The fitted per-site-per-dinucleotide affinities are shown in Fig. 3.3 for the Latin square dataset.

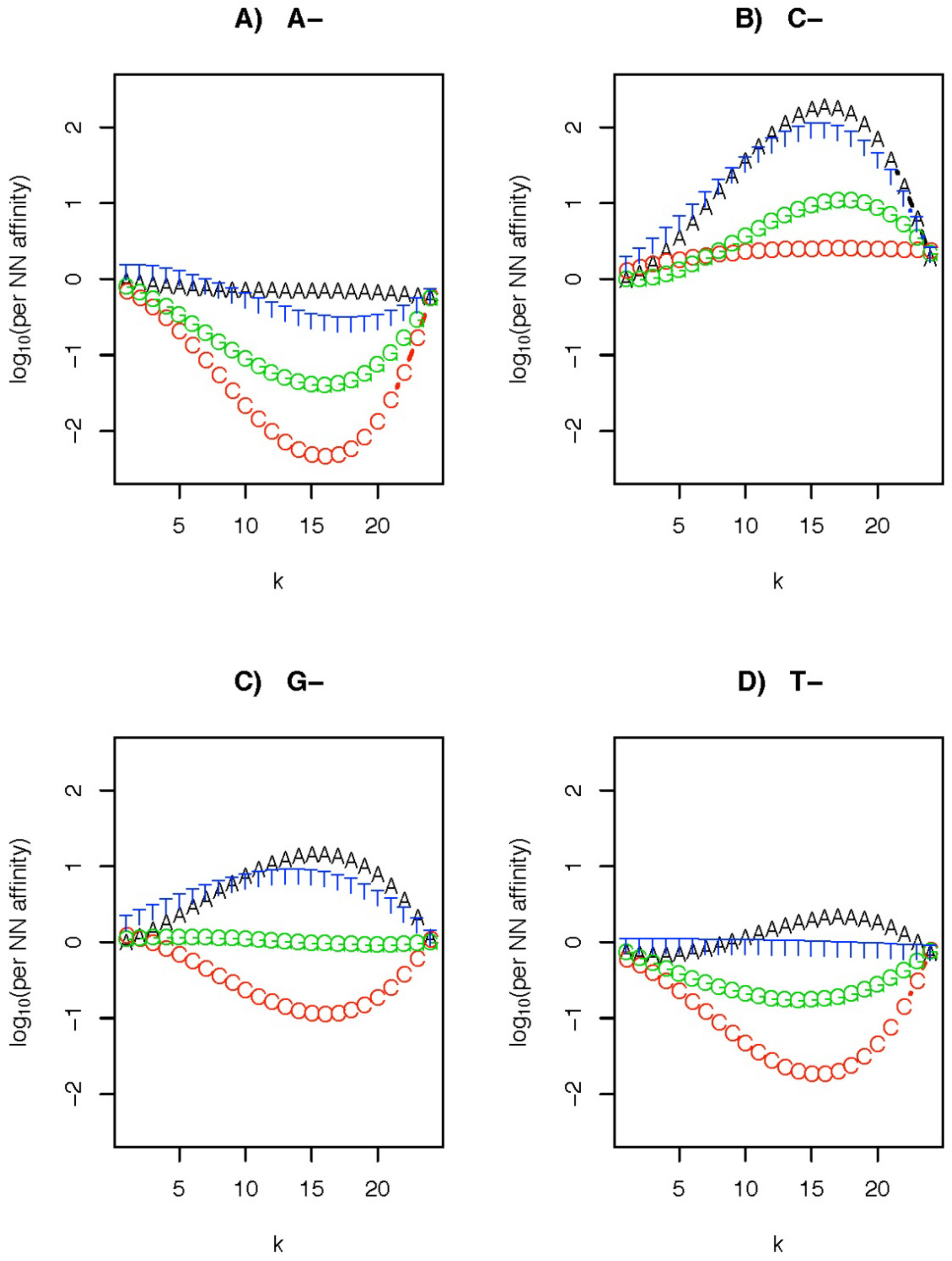

FIGURE 3.3 Affinity parameters calculated using dinucleotide model.
Affinity parameters calculated using equation 3.4 for the human genome U133 Latin Square. Affinity parameters are averaged across all the 42 chips; parameters for any single chip resemble those shown here. The first letter of each dinucleotide is indicated at the top of the figure, the second letter is indicated on the connected lines. K represents the position of each dinucleotide along the probe length. Higher affinity (Y-axis) indicates brighter signal.

Parameters obtained from other datasets are similar to the Latin square dataset parameters (data not shown). The figure shows that a probe with many AN (N=A, C, G, T) pairs (Fig 3.3A) tends to have much lower intensity than a probe with many CN pairs (Fig 3.3B) especially when those pairs are located at or near the probe center. This is broadly what we expect from the single nucleotide model. However, examining the effect caused by second nucleotide in each NN pair shows a pronounced effect for certain dinucleotides, which cannot be captured in the single nucleotide model. This can be seen in Fig. 3.3C and 3.3D. GA and GT rich probes are significantly brighter than GC rich probes, and TA rich probes are brighter than TC and TG rich probes.

The model defined in equation 3.4 was fitted to a number of datasets (see Methods section of this chapter). Fitting was performed on the PM and MM probes separately. Table 3.1 shows a comparison between the native Naef and Magnasco [77] affinity model (single nucleotide model, equation 3.1) and our dinucleotide affinity model (equation 3.4). We see that the dinucleotide model gives a better fit to microarray data by 5-10% on average (Table 3.1 and Fig. 3.2), depending on the chip and probe type. Note that both models perform better on the MM probes due to the higher background noise present in the MM signal.

TABLE 3.1 Dinucleotide model performance on different datasets.
R^2 of Naef and Magnasco [77] model (Single nucleotide) and the dinucleotide model for the five data sets used in this study. Results presented as average R^2 ± SD

Data set	nc[a]	np[b]		Single nucleotide model (Eq. 3.1)	Dinucleotide model (Eq. 3.4)[c]
Latin Square [94]	42	248152	PM	0.17±0.01	0.22±0.01
		248152	MM	0.40±0.01	0.50±0.01
Golden spikein [95]	6	195994	PM	0.20±0.02	0.22±0.02
		195994	MM	0.46±0.02	0.51±0.02
Leukemia [96]	72	201800	PM	0.49±0.06	0.55±0.07
		201800	MM	0.60±0.04	0.69±0.04
Etoposide response	60	496468	PM	0.05±0.04	0.08±0.06
[98]		496468	MM	0.11±0.06	0.16±0.08
BK knockout	20	496468	PM	0.09±0.04	0.13±0.04
[99-100]		496468	MM	0.29±0.050	0.36±0.06

[a] nc: number of chips.
[b] np: number of probes.
[c] The differences in R^2 between single nucleotide model and dinucleotide model are all statistically significant ($P < 10^{-3}$) using paired one-sided Wilcoxon and *t* tests.

Given that our fits contain between 195,994 and 496,468 data points (Table 3.1), it seems unlikely that the improvements in performance of our model could be explained by the additional free parameters (64 for our model vs. 16 for the original Naef and Magnasco model). Nonetheless, to rule out this possibility, we fitted both the single nucleotide model (N) (using the 100 free parameters and 16 free parameter version of the Naef and Magnasco model, equation 3.1 and 3.2, respectively) and the dinucleotide model (NN) with 64 free parameters (equation 3.4) to the Latin Square dataset using completely random probe sequences (generated with an equal probability of A, C, G and T). We also performed the same test on shuffled probe sequences in which the probe's base composition is not affected, but the position of each base has been changed due to the shuffling process. The results of this analysis are shown in Figure 3.2. We see that the R^2 of the shuffled and random probe sequences are nearly identical, no matter which method is used. The presence of additional free parameters in our model, therefore, cannot by itself explain the improved performance over the Naef and Magnasco model. This strongly supports our argument that the gain in the *r-squared* values of the NN model comes from including dinucleotide information and does not arise trivially from the addition of free parameters.

3.4.2 Background adjustment using dinucleotide affinity model

Using a more accurate estimate of background noise should improve the quality of Affymetrix GeneChip data. Given the better fits observed using the dinucleotide affinity model, we expected it to improve the analysis results to some degree when applied to control datasets. We tested the downstream effects of using this model on the quality of microarray data. We chose to implement the model within GCRMA [68], since it already has the single nucleotide model implemented in its background correction procedure, and therefore the two models could be directly compared.

In GCRMA, Wu *et al*. [68] model the signal intensity generated from each probe as:

$$PM = O_{PM} + N_{PM} + S, \qquad \text{Eq. 3.5}$$
$$MM = O_{MM} + N_{MM} + \phi S$$

where O is the optical noise, N is the background noise of non-specific binding, and S is the signal generated from specific binding between the probe and its intended target. The parameter ϕ reflects the fact that for some probe pairs, the MM signal may contain specific signal. The background components $log(N_{PM})$ and $log(N_{MM})$ are assumed to follow a bivariate distribution with means of $\mu_{pm}= h(\alpha_{PM})$ and $\mu_{mm}= h(\alpha_{MM})$, where h is a smoothing function and α (probe affinity) is defined by equation 3.1. In this paper, we make these same assumptions, but we derive α using equation 3.4.

We reasoned that GCRMA with background correction using the dinucleotide model, which we will subsequently refer to as GCRMA-NN, would perform better than the native GCRMA model. It is important to clarify that GCRMA offers two options for background correction, the first of which uses a precomputed α (called reference affinity) from the authors' own non-specific binding (NSB) experiments, while the second computes α directly from the data (called local affinity). In the following sections, we compare GCRMA-NN (where α is computed directly from the data using equation 3.4) to GCRMA-L (GCRMA with local affinity) and GCRMA-R (GCRMA with reference affinity).

3.4.3 Application of the dinucleotide affinity model to the Human Genome U133 Latin square dataset

We obtained expression measures for the Human Genome U133 Latin square dataset after processing it with GCRMA-R, GCRMA-L and GCRMA-NN. The three expression measures were evaluated using two approaches. The first approach is based on Affycomp [118-119], a performance evaluation tool for preprocessing algorithms (see below). The second approach is based on the number of true positives captured for all the 14 2X comparisons of the Latin square dataset at a cutoff of four false positives after using the cyber *t* test [120]. Cyber *t* is a popular variant of the *t* test, in which a weighted standard deviation replaces the conventional standard deviation and an adjusted number of degrees of freedom is used instead of the conventional degrees of freedom.

Performance of GCRMA-R, GCRMA-L and GCRMA-NN as reported by Affycomp based on 14 metrics is shown in Table 3.2.

TABLE 3.2 Affycomp scores for GCRMA-L, GCRMA-R and GCRMA-NN.
Fourteen Affycomp metrics for the U133 Latin square dataset rounded to two decimal points.

Metric*	GCRMA-L	GCRMA-R	GCRMA-NN	Perfect score
Median SD	0.06	0.06	0.07	0
null log-fc IQR	0.05	0.03	0.08	0
null log-fc 99.9%	0.62	0.61	0.64	0
Signal detect slope	0.99	1	0.98	1
Signal detect R^2	0.89	0.91	0.91	1
low.slope	0.49	0.48	0.55	1
med.slope	1.05	1.06	1.02	1
high.slope	0.97	0.97	0.96	1
Obs-intended-fc slope	0.99	1	0.98	1
Obs-(low)int-fc slope	0.48	0.47	0.53	1
low AUC	0.44	0.45	0.50	1
med AUC	0.87	0.87	0.86	1
high AUC	0.85	0.86	0.83	1
weighted avg AUC	0.55	0.56	0.59	1

*A brief description of each metric is provided under the Methods section of this chapter.

One notable performance enhancement of GCRMA-NN over GCRMA-L and GCRMA-R is a 3-4% increase in the weighted average area under the curve (AUC) (Table 3.2). This is a receiver operator characteristics (ROC) based metric, in which the absolute log-ratios for the expression summaries, for every comparison of any two pairs of the 14 arrays (92 comparisons), are sorted. After that, the number of true and false positives is found, and then the number of true positives at 100 false positives is determined for each pair of arrays. Finally, the resulting values are averaged over the three concentration groups (low, med and high), weighted by the number of probesets in each group and a score is recorded. Note that a perfect algorithm will have a score of 1, where all the true positives are captured before any false positive is recorded.

Examining Table 3.2 shows that the increase comes mainly from the AUC for low intensity targets (low AUC entry in Table 3.2). The low intensity genes make up most of the genes in a typical Affymetrix experiment [110] and are also the hardest to detect. Algorithms that perform inference generally can detect large changes involving highly expressed genes. It is much more difficult to detect changes in the more frequently observed genes that produce low intensities on the array. GCRMA-NN enhanced the detection of low intensity targets, while maintaining similar values for the medium and high intensity ones. The enhancement in detecting low intensity targets is also evident in the form of an increase in the low detection slope (low.slope entry in Table 3.2).

In the crucial category of low intensity genes, we argue that our algorithm outperforms most of the algorithms submitted to Affycomp, including GCRMA-R and GCRMA-L. The Affycomp webpage currently contains data for 88 algorithms for analyzing Affymetrix microarrays. For each of these algorithms, Affycomp defines accuracy as the slope obtained from regressing expression values on nominal concentration. An algorithm with a perfect accuracy would have a slope of 1, reflecting a perfect correspondence between nucleotide concentration and signal. Affycomp defines precision as the 99.9% percentile of the log fold changes of null (true negative) probesets across arrays. A perfect algorithm would have a precision of 0 reflecting a fold change of 1 (i.e. no change). Figure 3.4 is a plot of precision vs. accuracy for the Latin Square dataset for the 88 algorithms submitted to the Affycomp webpage. In Figure 3.4A, we see that when looking at overall accuracy vs. precision, the GCRMA-NN algorithm (blue dot) performs about as well as GCRMA-R (green dot) and GCRMA-L (red dot). However, for the crucial low intensity genes, for which inference is the most difficult, GCRMA-NN provides a better accuracy with no loss of precision (Fig. 3.4B).

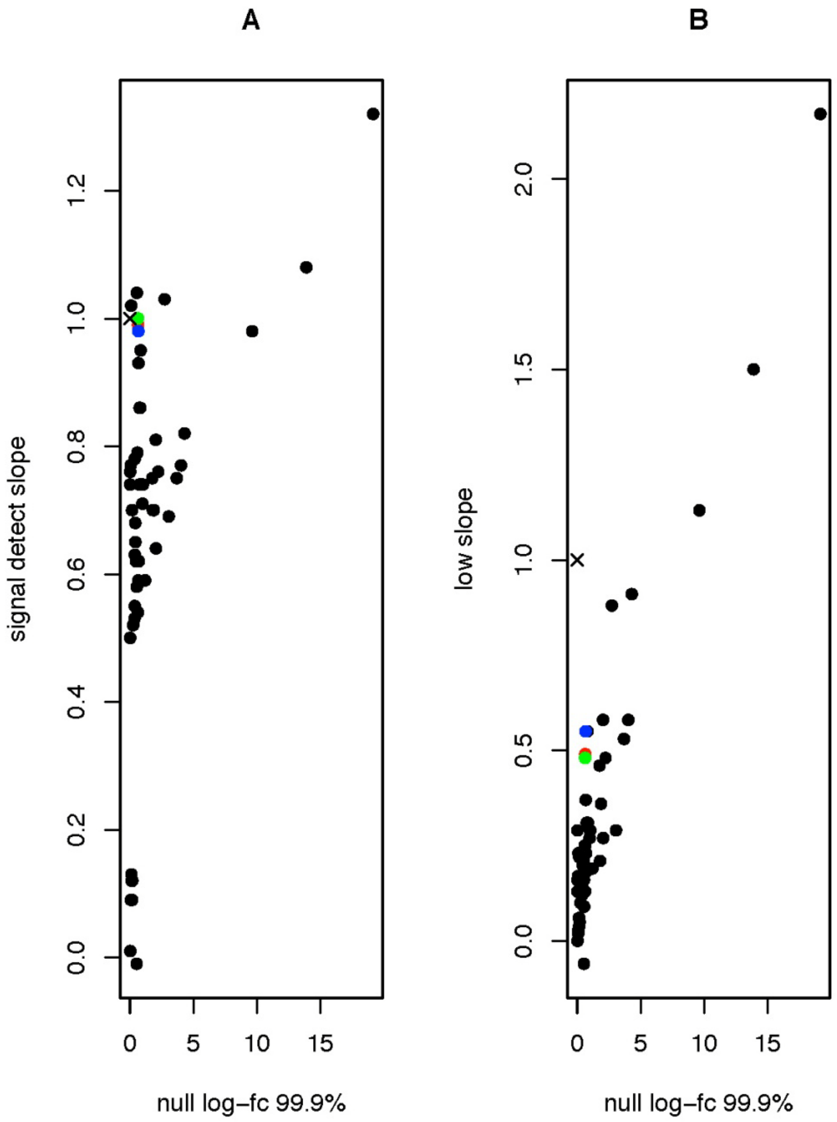

FIGURE 3.4 Accuracy and precision of GCRMA-R, GCRMA-L and GCRMA-NN.
A) Accuracy and precision of GCRMA-R (green dot), GCRMA-L (red dot) and GCRMA-NN (blue dot) compared to other preprocessing algorithms (black dots) submitted to Affycomp [118], information retrieved from Affycomp on November, 14th 2007. B) As A but for low expressed genes (< 4 pM). A perfect score is shown as an (×) on both panels. See Results section of this chapter for explanation.

Since the results of Affycomp suggest an improvement for the low intensity, hard to detect spikeins, we reasoned that inference performed with GCRMA-NN would be more successful than inference with GCRMA-R or GCRMA-L. We therefore applied GCRMA-NN, GCRMA-R and GCRMA-L to the U133 Latin square dataset. We considered only the 14 2X comparisons, in which the ratio of each spikein, between any two consecutive pair of arrays, is 2. Then we used the cyber *t* statistic [120] to generate a list of *P* values for the null hypothesis that the mean signal intensity in each comparison is the same. The lists were ordered, and for each of the 14 comparisons we generated an ROC curve. Figure 3.5 shows the average of these 14 ROC curves. For each ROC curve, we determined the number of true positives captured at an arbitrary cutoff of four false positives (vertical dashed line in Fig. 3.5A). The result of this analysis is summarized in Figure 3.5B. We see that GCRMA-NN outperforms GCRMA-R and GCRMA-L with a small but significant improvement. One-sided Wilcoxon and *t* tests reject the null hypothesis that GCRMA-NN is the same as GCRMA-R and GCRMA-L with all tests $P < 0.005$. These are consistent with the results we would have expected based on the Affycomp comparison (Table 3.2).

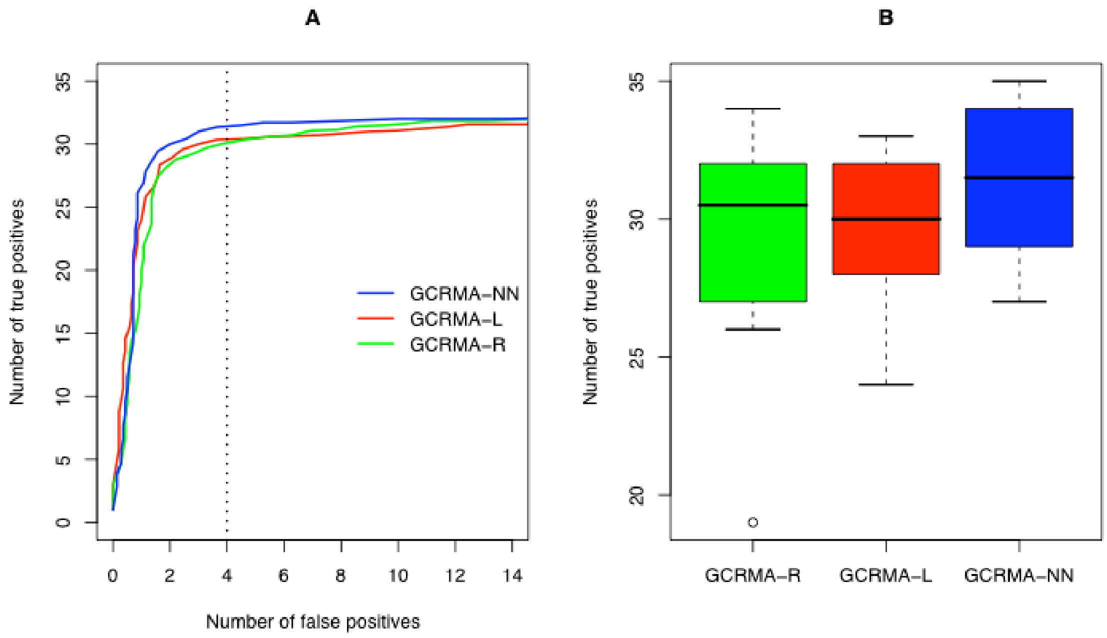

FIGURE 3.5 Performance of GCRMA-R, GCRMA-L and GCRMA-NN on the Latin square dataset.
A) ROC curves showing the average true positives and false positives across the 14 2X Latin square experiments following application of the cyber *t* test. B) The number of true positives captured for all the 14 2X Latin square experiments at a cutoff value of four false positives (dashed vertical line in panel A). The differences in panel B between GCRMA-R, GCRMA-L and GCRMA-NN are statistically significant (*P* < 0.005) using paired one-sided Wilcoxon and *t* tests.

3.4.4 Application of the dinucleotide affinity model to the golden spikein dataset

In order to ensure that our data were valid for more than one control data set, we next applied GCRMA-R, GCRMA-L and GCRMA-NN to the “golden spikein dataset” [95], which is not included in Affycomp. Figure 4.6 shows a ROC graph for the differentially expressed genes between the S and the C “golden spike” samples (see Methods section of this chapter) detected by GCRMA-R, GCRMA-L and GCRMA-

NN. As in the Latin Square data, the graph shows that GCRMA-NN is capable of capturing more true positives at lower false positive rate than both GCRMA-R and GCRMA-L. This supports our assertion that an improved background correction algorithm can have a noticeable effect on downstream analyses.

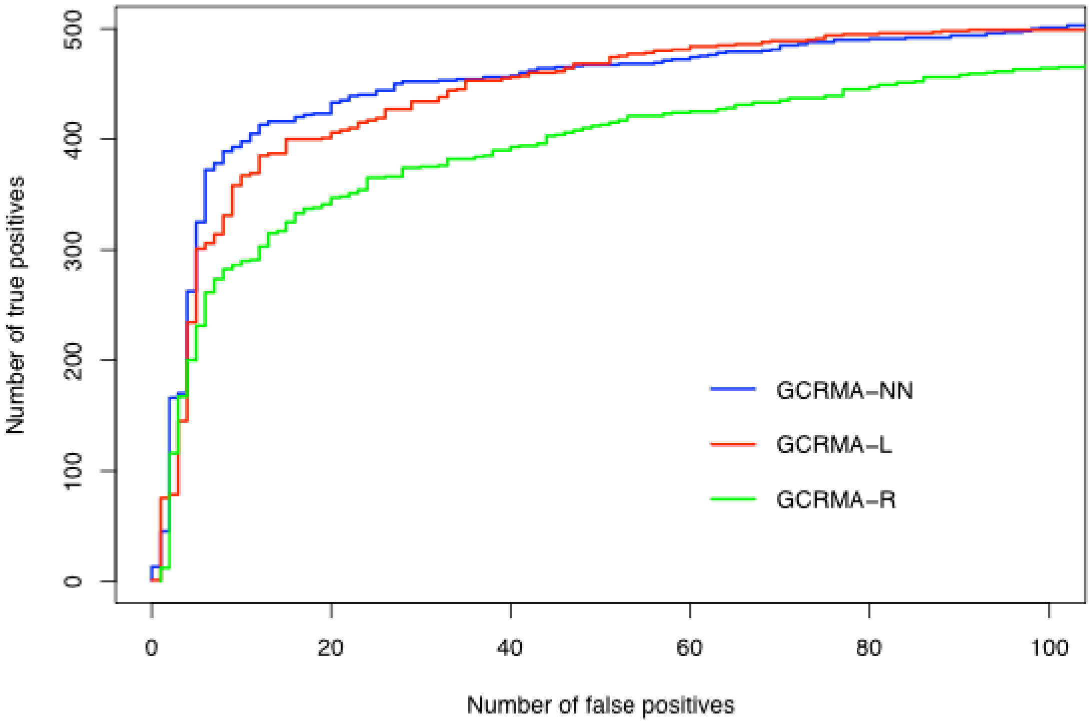

FIGURE 3.6 Performance of GCRMA-R, GCRMA-L and GCRMA-NN on the Golden spikein dataset.
ROC curves for the Golden spikein experiments C versus S after application of the cyber *t* test.

3.5 Discussion

Background estimation and correction are important steps in analyzing the data generated by GeneChip arrays. Improving algorithms for these steps increases the amount of true "signal" that we can detect from microarrays. Understanding background noise on GeneChip arrays, especially the part contributed by NSB signal, requires a deeper understanding of the behavior of on-chip hybridization. Given that we lack a detailed physical model of on-chip hybridization derived from first principles, an empirical model that estimates the specific and non-specific signal based on the data on the array and probe sequence is a useful tool for understanding the on-chip hybridization process.

Nucleic acid hybridization in solution is well approximated by the nearest neighbor model [121], which describes duplex formation as a function of the two adjacent nucleotides and their stacking orientation. This approach was used by Zhang *et al*. [69] to model the on-chip specific and nonspecific hybridization using the free energy formation for the adjacent nucleotides. Zhang *et al*. concluded that the on-chip hybridization parameters are different than the solution ones. Using a different approach to background correction, Naef and Magnasco [77] used single nucleotides to assign an overall affinity score for a probe based on its sequence away from the energy contributions of the dinucleotide pairs. This approach was used to perform background correction for the GCMRA algorithm [68] while the Zhang *et al*. approach was used to create the algorithm PerfectMatch [69]. PerfectMatch estimates the signal and the background at the same step while

GCRMA estimates background noise first then proceed to signal estimation. PerfectMatch is, therefore, much more computationally demanding than GCRMA as the parameter space searched by PerfectMatch is vast and is sampled with Monte Carlo methods. Direct comparison between GCRMA and PerfectMatch has proven controversial. Such a comparison is beyond the scope of this chapter, and can be found elsewhere [110, 112, 122].

In this chapter we combine some elements of GCRMA and PerfectMatch. We replace the single nucleotide model of Naef and Magnasco with a model in which the affinity of each probe is a function of its dinucleotide composition. Because we use GCRMA's approach of separating estimates of background and signal, we can use a linear model and avoid the Monte Carlo simulation approach of PerfectMatch [69]. Our approach is therefore both computationally more efficient and guarantees the best fit to the data. This approach enables us to examine the contribution of different dinucleotides at different positions to the raw probe signal (Fig. 3.3), rather than assigning one weight function to all the dinucleotides, as is done with PerfectMatch [69]. This allows our model to capture several important features of the background data such as the effect of the first versus the second nucleotide on probe affinity (e.g. CA vs. CG), and the effect of the stacking orientation (AC vs. CA). In general, we find that the dinucleotide approach has more power than the single nucleotide approach over a wide range of datasets (Table 3.1).

The mechanism that determines why particular dinucleotides affect probe affinities the way they do is, in some cases, unclear. However, we observe that the

NN model bears some similarities to the models of both Naef and Magnasco and Zhang *et al*. All three models emphasize the importance of the probe middle region; this is probably due to the surface attachment, as well as to the relative instability of the free end in RNA-DNA hybridization. The effect of the stacking orientation is in agreement with the findings of Zhang *et al*. [69]. The AN versus CN (where N refers to any of the four nucleotides: A, C, G, T; AN for example means AA, AC, AG and AT) asymmetry (Fig. 3.3A and 3.3B) is in agreement with Naef and Magnasco [77]. When comparing these affinity curves to the original Naef and Magnasco result, it is important to recognize that the NN model considers the affinity of dinucleotides rather than single nucleotides. Therefore, we do not necessarily expect to see the same asymmetry within CN or AN, i.e. there will be no asymmetry between CA and CC (Fig. 3.3B), or between AA and AC (Fig. 3.3A). The NN model, however, does show unexpected behavior for the GN and TN dinucleotides. While both G and T show slight asymmetry in the Naef and Magnasco model, the effect of these two nucleotides is magnified in the NN model. GN contributes positively to the signal but not when the second nucleotide is C (Fig. 3.3C). TN contributes negatively but not when the second nucleotide is A (Fig. 3.3D). This trend is partially explained by the fact that T forms fewer hydrogen bonds than G, therefore contributing negatively, while the G has stronger binding, thus contributing positively. This trend is not consistent, and appears to be dependent on the adjacent nucleotide. It could also be due to the biotin label present on the RNA target sequence.

When applied to two control datasets, GCRMA-NN showed improved performance (Figs. 3.5 and 3.6) especially on low intensity targets (Table 3.2; Fig. 3.4). We argue that this is due to better background correction for these targets; a higher percentage of low intensity signal will be made up of background, so it is therefore not surprising that better background correction will make more of a difference on low intensity targets. The detection of low intensity targets represents the most significant challenge to microarray analysis algorithms, which makes any enhancement in the detection of these targets significant.

In conclusion, incorporating dinucleotide information into a previously described probe affinity model increases the fit of the model by 5-10%. The dinucleotide affinities highlight the importance of the stacking orientation on probe behavior. This is in agreement with the physical models that describe hybridization binding affinities. The results presented here show that the affinity of any single nucleotide is affected by its neighbor, in addition to its location along the probe. Considering the second nucleotide offers more insights into the on-chip behavior of the four bases in relation to each other. Such insights are important to develop a better understanding of the on-chip hybridization process and therefore better analysis procedures. The model described here enhances the performance of an existing widely-used preprocessing algorithm for GeneChip data. We expect the same model to enhance the performance of preprocessing algorithm for other types of arrays, in particular those used for SNP analysis.

CHAPTER 4: APPLICATION OF EQUILIBRIUM MODELS OF SOLUTION HYBRIDIZATION TO MICROARRAY DESIGN AND ANALYSIS

4.1 Abstract

Mismatches in sequence between target and probe on microarrays affect the behavior and the quality of the data obtained from such probes. Although long oligonucleotide microarrays are in widespread use, there are a limited number of controlled studies of probe response to mismatched targets on these arrays, and the properties of mismatched duplexes are not well understood, as they are for 25-mer based platforms. The probe percent bound calculated using multi-state equilibrium models of solution hybridization is shown here to be useful in understanding the hybridization behavior of microarray probes up to 50 nucleotides in length.

We present a comprehensive analysis of the effects of single, double and triple central mismatches on the behavior of 50-mer probes at eight different target concentrations. The results show that differentiation between the perfect match signal and the mismatch signals is possible at medium (100 pM to 200 pM) target concentrations, and that this behavior is predictable using solution hybridization modeling methods. Both the models and the array platform are sensitive to the effects of single, double and triple central mismatches on hybridization of 50-mer probes at multiple target concentrations. We discuss the impact of these results on microarray design, optimization and analysis.

Our results highlight the importance of incorporating biophysical factors in both the design and the analysis of microarrays. We suggest use of the probe 'percent bound' predicted by equilibrium models of hybridization as a factor in predicting and assessing the behavior of long oligonucleotide probes.

4.2 Introduction

DNA microarrays [1] have revolutionized every area in biology [2]. Microarrays allow thousands of genes to be assayed at once, offering global views of biological processes by providing a systematic way to survey gene expression [105], DNA sequence variation [2], alternative splicing [123], and to rapidly perform cancer classification [124], genome annotation [125] and functional genomics assays [126]. The biology research community, both commercial and nonprofit, has invested heavily in microarray technology despite ongoing challenges with data quality and data interpretation.

A microarray chip is a surface, which provides support to tethered nucleic acids called probes [3]. These probes are designed to interact with a mixture of labeled nucleic acids called targets [3]. Chemically synthesized oligonucleotide probes vary from short (20-30mer) probes to long (50-70mer) probes [4]. The origin of the targets depends on the type of the experiment. For example, targets for gene expression arrays are derived from total mRNA extracted from the cell under specific conditions. For SNP (single nucleotide polymorphism) arrays, targets are derived from genomic DNA. When a labeled target binds to its complementary probe, a stable interaction results and a signal is detected. This signal is interpreted, for example, as reflecting transcript abundance, presence and/or absence in gene expression experiments [5-6], or used for genotype calling in genetic analysis and SNP experiments [127].

It has been shown that probes between 50-70 nucleotides in length can deliver higher sensitivity than shorter probes [21], due to their higher target affinities, and that longer probe lengths allow for the design of microarrays with fewer, but more selective probes [7, 128]. As a result, the use of long oligo microarrays is widespread, but model development for analysis of hybridization signal from these arrays has received relatively little attention. Most of the available modeling, optimization and analysis studies have been done on short oligonucleotide microarrays [66, 129-130]. Our long-term goal is to address the deficiencies in modeling of long-oligo microarrays, and we report a significant step toward that goal below.

Like many applications in molecular biology, microarrays rely on stable duplex formation between an oligonucleotide probe and a potential target in a genomic DNA or mRNA pool. Accurate interpretation of the signal relies on the specificity of hybridization, the discrimination that can be achieved between completely complementary hybrids and those with some degree of mismatch [131]. While the biophysics of short oligonucleotide binding in solution is well understood [48], the binding properties of longer, tethered oligonucleotides are less well characterized [48]. Understanding the effects of mismatches at various positions in long probes (50-70 nucleotides) on the extent of probe-target hybridization is critical for any microarray experiments, because microarray results are often interpreted as if they are quantitative, even when all of the variables that may affect hybridization are not well understood.

The hybridization interference that results from the presence of mismatches between probe and target represents a data analysis challenge and a potential opportunity at the same time. Mismatches may degrade the sensitivity and the specificity of the probe by reducing the affinity of the probe for its intended target, or, if present in sufficient number, mismatches may lead to false positive results through cross hybridization with unintended targets [132]. On the other hand, the properties of mismatched oligonucleotides are important for applications including interspecies microarray hybridization [133], detection of SNPs across the genome, and background noise estimation [105].

The effect of a small number of mismatches on the hybridization of short oligonucleotides (20-30 nucleotides) has been well investigated [69, 85, 134-136], and they are well suited to discriminating small sequence differences. However, the limit of detection of mismatches in the hybridization of long oligonucleotides, 50-mers particularly, has not been comprehensively investigated. Instead, most investigations have focused on determining the minimum number of base pairs required for detection of significant signal contributions from non-specific hybridization [132, 137]. Published studies of long oligonucleotide probes up to this point use a very limited number of sequences, only deal with more than three mismatches in a long probe, or only investigate the outcome of a mismatch between probe and target at a single target concentration [129, 132]. This study extends our understanding of perfect match (PM) vs. mismatch (MM) binding behavior to longer oligo lengths.

A common feature of the available studies on both short and long oligonucleotide probes is the assumption that probe-target hybridization follows a simple nearest-neighbor two-state model. It has been shown that microarray hybridization is more complicated and that factors such as probe folding, target folding, probe-probe interaction, target-target interactions and competition from other similar sequence strands limit the applicability of the traditional two-state model to microarray hybridization modeling [76, 90, 138]. Multi-state models where all of the above mentioned factors are taken into consideration are suggested for a more accurate modeling of nucleic acid hybridization [48].

Available thermodynamic modeling studies of long oligos have previously been aimed at comparing solution hybridization free energy (nearest neighbor) parameters and microarray hybridization free energy parameters [80, 139-140]. Such studies provide valuable information about the molecular interactions that take place on different microarray platforms. These studies provide a proof of concept for using solution hybridization parameters as a valid approximation for microarray hybridization, enabling the use of solution hybridization parameters in predicting the behavior of microarray probes.

Application of nearest-neighbor based models of solution hybridization to predict the behavior of long oligonucleotide probes has been limited by the fact that the model is thought to be most accurate for probes with length ≤ 40 oligonucleotides [141]. Recently, several groups have shown that solution hybridization parameters based on the nearest-neighbor model can be applied to short oligonucleotides [139-140]. Hooyberghs *et al*. [80] showed that the nearest-neighbor parameters of solution hybridization and microarray hybridization are well correlated (*r*= 0.839) for probes of 30 oligonucleotides in length. To our knowledge, application of multi-state solution hybridization models to predict the behavior of oligonucleotide probes, above the 40nt length at which accuracy of solution models is predicted to decline, has not been previously demonstrated. Extension of these models to long-oligonucleotide platforms is the focus of our present research.

In this study, we designed and modeled the binding behavior of a set of ten 50-mer probes, each of which has six counterparts with different introduced mismatches. We subsequently refer to the base set of 10 probes as perfect match (PM) probes, and their permuted counterparts as mismatch (MM) probes. The full set of 70 probes was spotted on a standard epoxysilane-coated glass slide substrate, and the resulting arrays hybridized to a mixture of well-defined targets. By creating mismatches as permutations of the surface-bound probes rather than permutations in the target, we are able to unambiguously separate and directly compare the signal from a perfectly matched duplex and an analogous duplex with one, two or three central mismatches. This experiment confirms the accuracy of current computational models of multi-state solution hybridization for the microarray surface context, with the caveat that the experimental system is still somewhat simplified relative to the complexity of a typical microarray experiment. Below, we extract the signal and then model the effects of central single, double and triple mismatches on the hybridization signal intensity detected at 50-mer probes under different target concentrations. Finally, we propose a simple model to predict signal intensity based on computational models of solution hybridization. This model uses the probe percent bound calculated based on a multi-state equilibrium model of solution hybridization, as implemented in the OMP (Oligonucleotide Modeling Platform) software [142], as a predictor.

4.3 Methods

4.3.1 Probe design and selection

Perfect match probes (PM) were designed using a two-stage process of rapid sequence screening followed by biophysical modeling (Fig. 4.1). Briefly, the design process used Yoda [27] as the primary selection tool. In order to produce probes having a natural sequence composition, probes were generated based on known gene sequences from the genome of *Brucella suis*. Yoda-generated probes were then screened for secondary structure formation potential using the program *hybrid-ss-min* from the UNAFold package [143]. Finally, hybridization behavior of probes that passed this step was simulated using OMP against a 50-mer perfect match target.

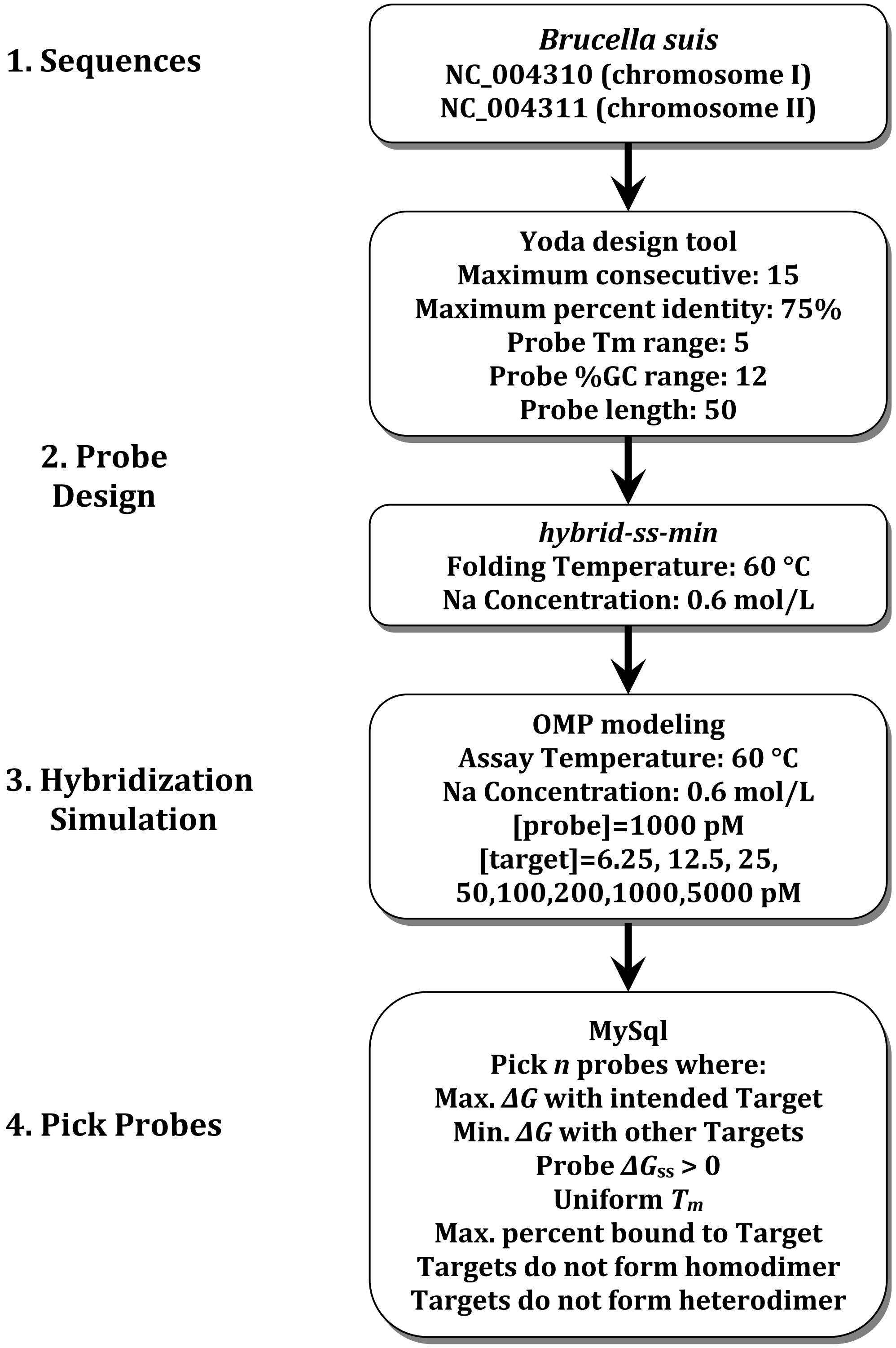
1. Sequences
Brucella suis
NC_004310 (chromosome I)
NC_004311 (chromosome II)
2. Probe Design
Yoda design tool
Maximum consecutive: 15
Maximum percent identity: 75%
Probe Tm range: 5
Probe %GC range: 12
Probe length: 50
hybrid-ss-min
Folding Temperature: 60 °C
Na Concentration: 0.6 mol/L
3. Hybridization Simulation
OMP modeling
Assay Temperature: 60 °C
Na Concentration: 0.6 mol/L
[probe]=1000 pM
[target]=6.25, 12.5, 25, 50,100,200,1000,5000 pM
4. Pick Probes
MySql
Pick n probes where:
Max. ΔG with intended Target
Min. ΔG with other Targets
Probe ΔG_{ss} > 0
Uniform T_m
Max. percent bound to Target
Targets do not form homodimer
Targets do not form heterodimer

FIGURE 4.1 Probe design schema.

Simulation results for all probe candidates were stored in a MySQL database and the most optimal ten probes were selected using an *ad hoc* multi-criterion sort. Six mismatch (MM) counterparts were designated for each of the ten probes. The first MM has a single mismatch at position 24, the second MM has a single MM at position 25 and the third MM has a single MM at position 26. The first double-mismatch probe has two permutations at positions 24 and 25, the second at positions 25 and 26. The last MM probe has triple-mismatches at positions 24, 25 and 26. All permutations were generated using a local Perl script. Six probes from *Arabidopsis thaliana* genome were designed using the same approach, to serve as negative controls.

4.3.2 Fabrication of microarray slide

All probes were synthesized with amino-C6 linkers at the 5` end by Operon Biotechnologies (Huntsville, AL). Microarray slides were manufactured by ArrayIt (Sunnyvale, CA). Probe purity and concentration were verified spectrophotometrically and with PAGE. Each probe was spotted in six replicate spots on each slide. After preliminary hybridization tests with multiple spotting buffers, a buffer containing 10 μM probe concentration was chosen as the optimal spotting concentration for the current experiment.

4.3.3 Preparation of target mixture

Perfect match (PM) targets (50-mer length) for the ten original PM probes were synthesized and Cy3 labeled at the 5` end by Operon Biotechnologies (Huntsville, AL). All targets were HPLC purified to ensure length and labeling uniformity and verified spectrophotometrically and with PAGE for purity and concentration. Target Oligos were re-suspended to 100 µM concentration in 2X SSC. A target master mixture was made by adding all ten targets into a single solution. This solution was aliquoted and then diluted to a series of concentrations using a target buffer solution (0.4 mg/ml salmon sperm, 4X SSC, 0.5% SDS) which had been heated to 95 °C for 5 minutes, and then chilled on ice for 10 minutes before the addition of the oligos. The following target concentrations were used: 5000 pM, 1000 pM, 200 pM, 100 pM, 50 pM, 25 pM, 12.5 pM and 6.25 pM.

4.3.4 Array hybridization

The slides (two technical replicates for each concentration) were placed in a HS 4800 Pro Hybridization Station (Tecan, Mannedorf, Switzerland), which had been preheated to 55 °C. All wash solutions were also preheated by the hybridization station. The slides were then wetted by a brief rinse with a Hyb Wash solution (0.5X SSC, 0.005% SDS) so that the slide was not dry when it receives the blocking buffer. The slides were blocked with BlockIt solution (ArrayIt, Sunnyvale, CA) for 30 minutes. The slides were then washed again for 2 minutes with the Hyb Wash solution. 60 µL of target solution was then added and the slides were hybridized for 18 hours at 55 °C. Slides were subject to mechanical agitation at medium intensity (1.1 minutes agitation with 3.5 minutes break) during hybridization. Then the slides were washed three times in the Hyb Wash solution for 30 seconds, then washed for a minute with the Hyb Wash solution and cooled to 50 °C. The slides were then washed with TE for 30 seconds and cooled to 45 °C, washed with 0.5X TE for 30 seconds and cooled to 40 °C, washed with 5% alcohol (Sigma Aldrich, St. Louis, MO) for 1 minute and cooled to 30 °C, and finally washed twice with ddH_2O for 40 seconds and cooled to 25 °C. After these washes the slide was dried under nitrogen for 3 minutes.

4.3.5 Image acquisition and data analysis

Slides were scanned with the 532nm laser, a 575nm filter, 10μm resolution, an over sampling factor of 2 and a 150 PMT gain in the LS Reloaded Scanner (Tecan, Mannedorf, Switzerland). Images were saved as Tagged Image File Format (tif) and then analyzed using SPOT (CSIRO, Sydney, Australia, http://www.hca-vision.com/product_spot.html), with the segmentation option set to 'seeded region growing'. The quality of each array and its spots were determined according to He *et al*. [132]. The raw intensities were loaded into the LIMMA [144] package (version 2.12.0) of Bioconductor [117] using the *read.maimages* function. LIMMA was also used for between-array (quantile) normalization for each pair of technical replicates. The analysis presented in this work was done using R (version 2.6.1) [145].

4.3.6 OMP hybridization simulation

Hybridization simulations for probe design and signal intensity prediction were done using OMP DE (version 1.1.0.2089) running on Red Hat Enterprise Linux 4. In this case, the complete system of 76 probes (including negative controls) and ten 50-mer targets could be simulated simultaneously, as the sequence lengths of the system was manageable by OMP in a reasonable period of time. For each probe, the following data were collected from the OMP output: ΔG, ΔH, ΔS, T_m, basepair count and probe percent bound (PPB). All the parameters used in the hybridization simulations are shown explicitly in Figure 4.1. Target concentrations for signal prediction are listed in section 4.3.3 above.

4.3.7 Langmuir isotherm fitting

The Langmuir isotherm is a chemical adsorption model [146] that has been applied successfully to short oligonucleotide microarrays [62, 66-67]. The model is simply a hyperbolic response function in the form of:

$$I_j = A\frac{c_j}{K + c_j} + bg \qquad \text{Eq. 4.1}$$

where I_j is the signal intensity from the probe at target concentration j. A, K and bg are the model fitting parameters, c is the target j^{th} concentration in pM. This model has three free parameters (A, K and bg) fitted to eight different concentrations (5000 pM, 1000 pM, 200 pM, 100 pM, 50 pM, 25 pM, 12.5 pM and 6.25 pM). The fitting parameter K is the probe affinity constant, A is the saturation intensity (assuming no cross-hybridization, i.e. $bg = 0$) and bg is a background component [62, 71, 73, 146]. The model was fitted using the *nls* function of R (version 2.6.1) [145].

4.3.8 Predicting signal intensity using probe percent bound

We aim to explain the response of each probe in the experiment according to the physical chemistry of hybridization. Based on OMP simulations, we developed a simple linear model with only two free parameters to predict the signal intensity.

Our model makes use of probe percent bound (PPB) as a predictor of the signal generated from each probe, following the equation below:

$$I_j = B_0 + B_1 \langle \%Bound \rangle_j + \varepsilon_j \qquad \text{Eq. 4.2}$$

where I_j is the signal intensity from the probe at target concentration *j*. *%Bound* is the PPB of the probe at target concentration *j*. B_0 and B_1 are free parameters and ε_j is an error term. OMP percent binding predictions were computed in the presence of all targets (competitive hybridization), at eight different target concentrations: 5000 pM, 1000 pM, 200 pM, 100 pM, 50 pM, 25 pM, 12.5 pM and 6.25 pM. This model has two free parameters (B_0 and B_1) fitted to eight PPB values. The model was fitted using the *lm* function of R (version 2.6.1) [145].

4.3.9 The U133 Latin square dataset

This dataset is composed of 14 experiments (three technical replicates for each experiment) in which 42 transcripts are spiked at a concentration range of 0.125-512 pM following a Latin square design. The dataset files were downloaded from Affymetrix web site [94].

4.3.10 Code and data

The code and data used in this chapter are available as an R package and can be downloaded from [147]. The complete set of figures for this study can be reproduced using this package.

4.4 Results

4.4.1 Effect of central mismatches on signal intensity

It is well known that mismatches in the central part of a short (25-35mer) probe have the most destabilizing effect on the probe-target duplex. This effect has not been examined thoroughly for longer oligonucleotides, and the concentration dependence of the effect has not been studied. Although the presence of one or a few mismatches in a long probe will not abolish hybridization, it is likely to reduce hybridization efficiency [148]. We seek to understand how much efficiency reduction such mismatches will introduce in a 50-mer probe at different target concentrations. In Figure 4.2, we examine the effect of single-, double- and triple-MM on the hybridization signal intensity at eight different target concentrations.

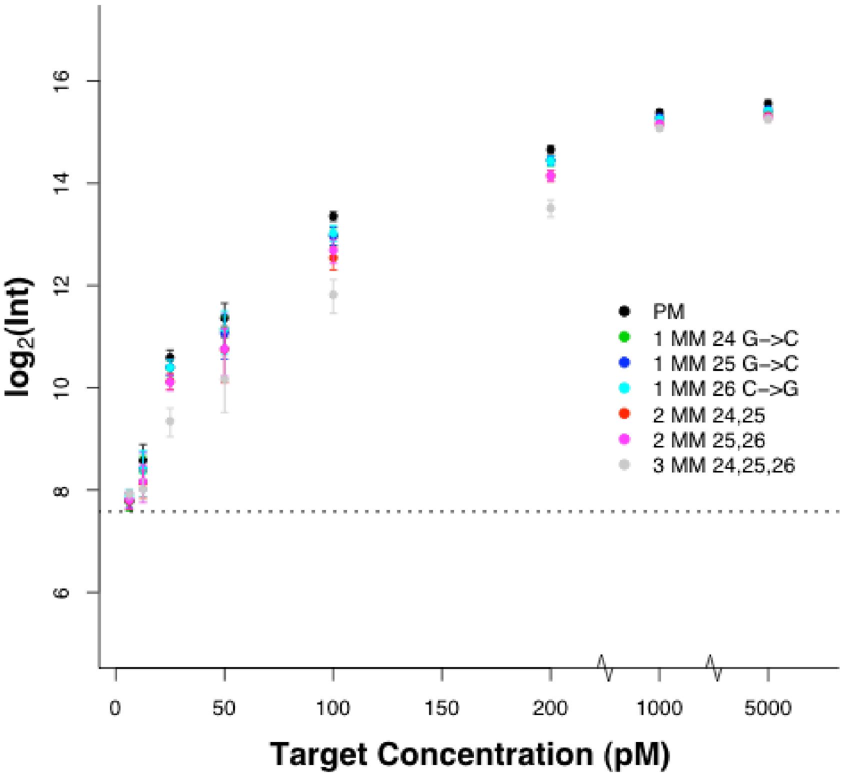

FIGURE 4.2 PM and MM probes responses to eight target concentrations. Hybridization signal intensities for PM probe 5005 (black) and its six MM counterparts: 5035 (green), 5036 (blue), 5037 (cyan), 5038 (red), 5039 (pink), 5040 (gray) under the eight target concentrations used in this study. The type, identity and location of the mismatch are indicated beside the symbol of each probe: 1MM: single-MM, 2MM: double-MM, and 3MM: triple-MM. Dotted line shows the average background signal across eight concentrations.

Figure 4.2 shows an example of the differences in signal intensity between PM, single-, double- and triple-MM for probe 5005 and its six mismatch probes. Noticeably, even three mismatches in the middle of the probe did not abolish the hybridization signal. Figure 4.2 also shows a correlation between the target concentration and the effect of mismatches. At the lowest target concentration examined (6.25 pM), no visible differences can be seen. The three categories of MMs have approximately the same signal intensity as the PM probe but they are marginally above the background signal. At medium target concentrations (12.5-200 pM), the differential effect of mismatches is clearly noticeable. The signal intensity is reduced with increasing number of mismatches (PM > single-MM > double-MM > triple-MM). At higher target concentrations (1000 pM and 5000 pM) the signal intensities obtained from each MM probe are close to that of the PM probe, and no visible effect of the presence of mismatch on the generated signal intensity can be seen. It should be noted that the above analysis was also conducted using raw intensities (no normalization) and similar trends were observed for all probe sets.

Based on the analysis shown in Figure 4.2, we wanted to test if the differences in the signal obtained from the PM and the three different categories of mismatches are statistically significant. Using one-sided *t-test*, we tested the null hypothesis that the mean signal intensity of each PM probe is lower than the mean signal intensity of its: (A) single-MM counterparts (each PM probe has three single-MM probes, which results in 10 x 3 tests), (B) double-MM counterparts (each PM probe has two double-MM probes, which results in 10 x 2 tests), (C) triple-MM counterpart (each PM probe has one triple-MM probe, which results in 10 x 1 tests). The result of this analysis, for all the probes, is presented in Figure 4.3 and indicates that the observed differences are significant.

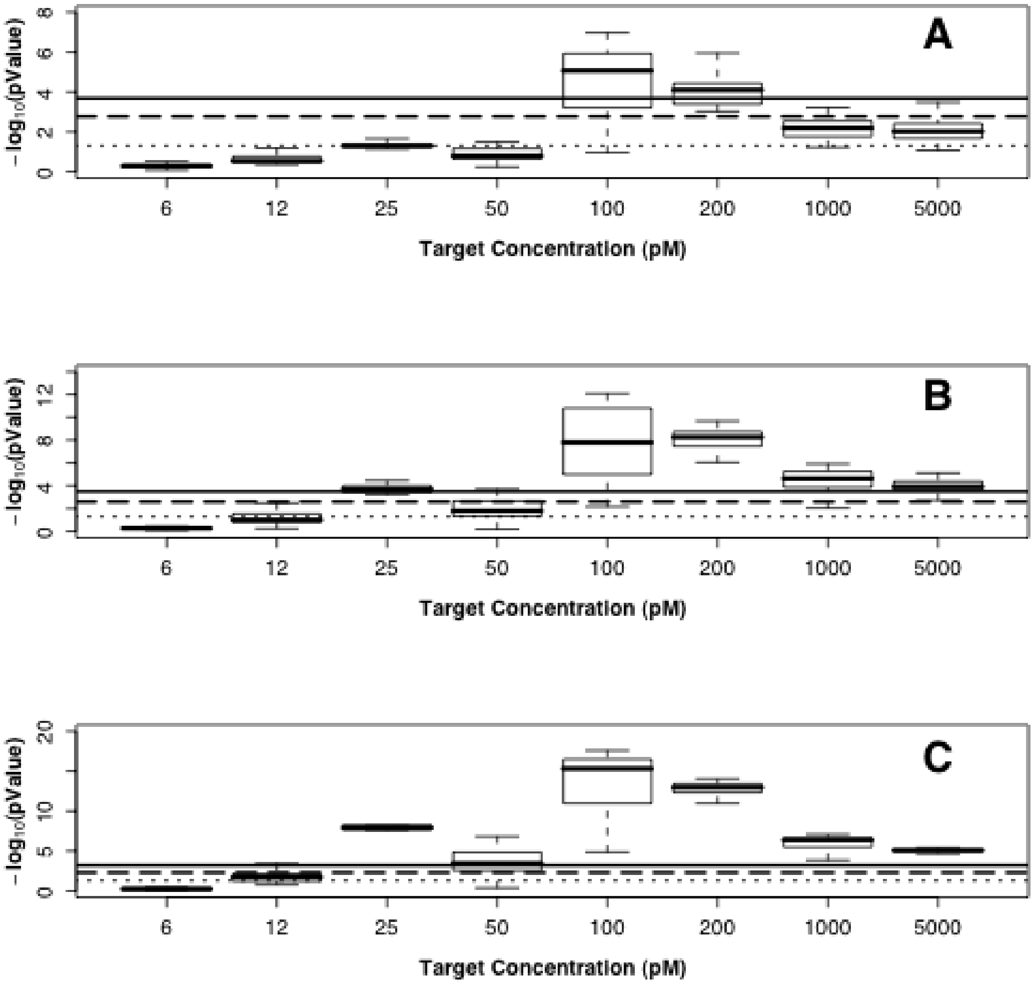

FIGURE 4.3 Single-MM, double-MM and triple-MM signal can be differentiated from the PM signal.

Box plots for the *p*-values of testing the null hypothesis (A) PM signal intensity < single-MM signal intensity, (B) PM signal intensity < double-MM signal intensity and (C) PM signal intensity < triple-MM signal intensity for all probes under the eight target concentrations used in this study. Dotted line indicates P = 0.05 level, dashed line indicates P = 0.05/(30 or 20 or10) level (Bonferroni correction using the number of hypothesis tests conducted at each concentration; 30 comparisons for single-MM probes or 20 comparisons for double-MM probes or 10 comparisons for triple-MM probes) and solid line indicates P = 0.05/(240 or160 or 80) level (Bonferroni correction using the sum of all the conducted tests; 8 concentrations x 30 comparisons for single-MM probes or 8 concentrations x 20 comparisons for double-MM probes or 8 concentrations x 10 comparisons for triple-MM probes). All the tests were done using a one-sided *t*-test. A Wilcoxon signed-rank test returned similar results.

In all cases the more mismatches the greater the discrimination, as expected. And the results were most clearly seen where there was sufficient but not saturating signal. The result also shows a dependency on the target concentration, in particular, target concentrations in the range of 100-200 pM yielded the most significant results. Figure 4.3A shows box plots of *p*-values for the difference between PM and single-MM probes. Although Figure 4.2 shows no visible differences in the signal intensity between PM and single-MMs at the eight target concentrations, we were able to confidently differentiate between PM signal and single-MM signal at target concentrations of 100 and 200 pM. Figures 4.3B and 4.3C show box plots of *p*-values for the difference between PM and double- and triple-MM, respectively. Most of the observed differences in Figure 4.2 are statistically significant. We were able to differentiate between PM signal and double- and triple-MM signals at all the tested concentrations except at extremely low target concentrations 6.25, 12.5 and 50 pM (PM and double-MM) and 6.25 and 12.5 pM (PM and triple-MM).

Similar results were obtained when comparing single-MM vs. double-MM, single-MM vs. triple-MM and double-MM vs. triple-MM. Target concentrations of 100 and 200 pM showed the most significant difference, while 6.25 and 12.5 pM showed no statistically significant difference in mean signal intensity. Comparing the mean signal intensity obtained from single-MM at three different positions (24, 25 and 26) did not show any significant difference. Signal intensity obtained from double-MM at two different positions (24 + 25 and 25 + 26) did not also show any significant difference.

Surprisingly, given the results of prior studies in 25-mers, which showed strong sequence dependence of binding affinity [62, 70, 77-78], there seem to be no dependency on the identity of the mismatch. For all the 70 probes studied, we grouped probes based on the identity of the mismatch and compared their mean signal intensity. Changing A→T, T→A, G→C or C→G did not affect the generated signal significantly. Similarly, no differences were detected between double-MM regarding the identity of the mismatch.

4.4.2 50-mer probe signal intensities show nonlinear response over target concentrations

To study the hybridization characteristics of 50-mer probes, we fitted the data to equation 4.1. Figure 4.4 shows, as an example, the responses of probe 5003 and its six MM probes to eight different target concentrations. The responses are typical of the Langmuir isotherm model and show that the model captures the physical chemistry of hybridization with $R^2 \geq 0.97$. Figure 4.4 also shows a clear distinction between PM, single-, double- and triple-MM probes. The model can predict the response of each probe type at different target concentration with good discrimination between PM, single-, double- and triple-MM probes. This trend was observed for all the probes investigated in this study.

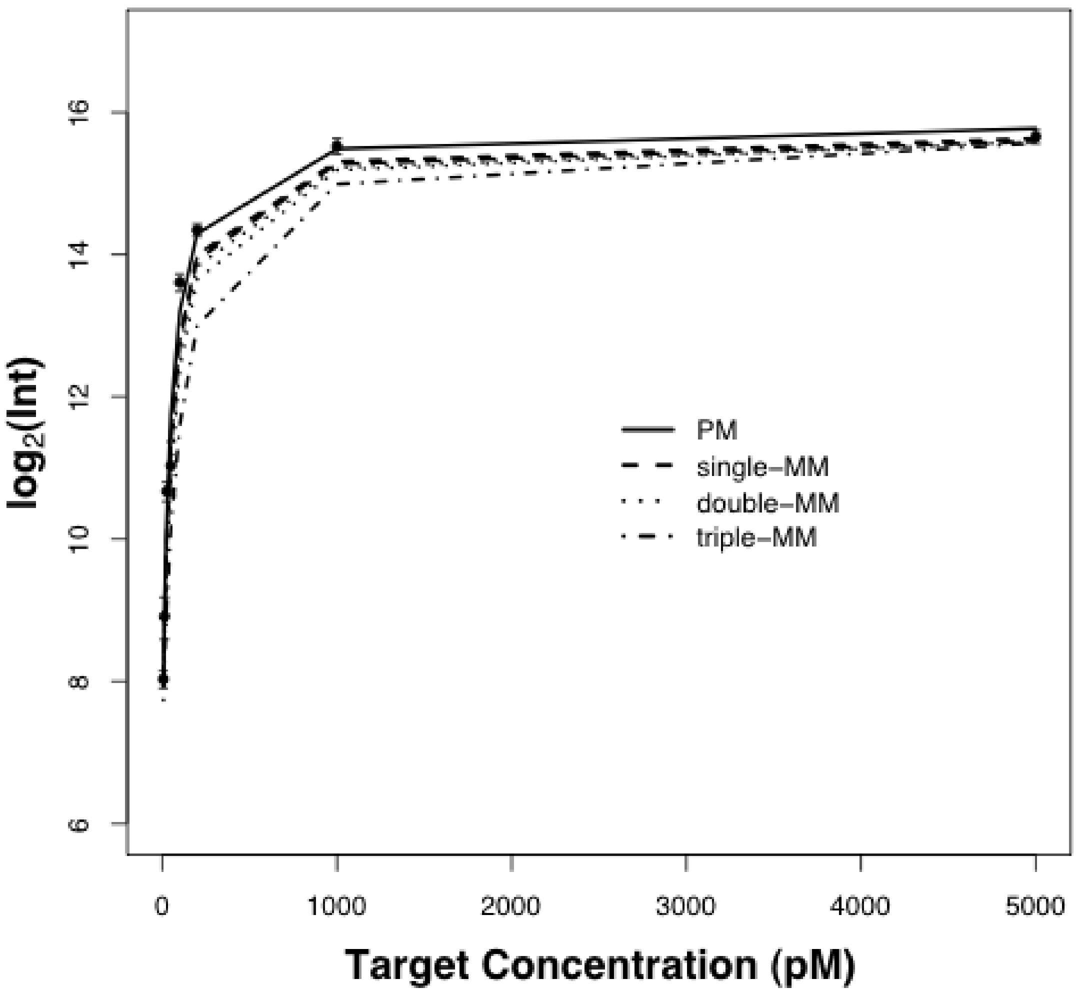

FIGURE 4.4 Signal intensity versus target concentration.
Signal intensity versus target concentration for PM probe 5003 and its six MM counterparts. Points represent the observed PM intensities and lines represent Langmuir Fit (equation 4.1) for PM (solid), single-MM (dashed), double-MM (dotted) and triple-MM (dot-dashed) probes. The observed MM intensities are omitted for clarity.

We did not observe any significant differences between the three types of single-MM probes; they have similar responses and cannot be differentiated from each other but can be differentiated from the other classes of probes in the set (Fig. 4.4). That is, single-MM probes appear to respond to target concentrations with very similar curves despite the fact that they have different mismatches at different locations. The two types of double-MM probes also show similar response curves and cannot be differentiated from each other but can be differentiated from other classes of probes in the set (Fig. 4.4). Finally, triple-MM probes have distinctive response curves, which can be differentiated from the PM, single- and double-MM probes. This trend is seen for all of the ten sets of probes under investigation.

Figure 4.4 also shows the chemical saturation of the probes. All the probes (PM and MM) saturate between 1000 and 5000 pM target concentration, with different final saturation intensities. As expected, the PM probe has the highest saturation intensity followed by the single-MM probes, then the double-MM probes and lowest for the triple-MM probe. To examine the response of all probes used in this study, we collected the saturation intensity (the fitted parameters $A + bg$ in equation 4.1) and the affinity constant (the fitted parameter K in equation 4.1) for each group of probes (PM, single-MM, double-MM and triple-MM), and we present them in Figure 4.5.

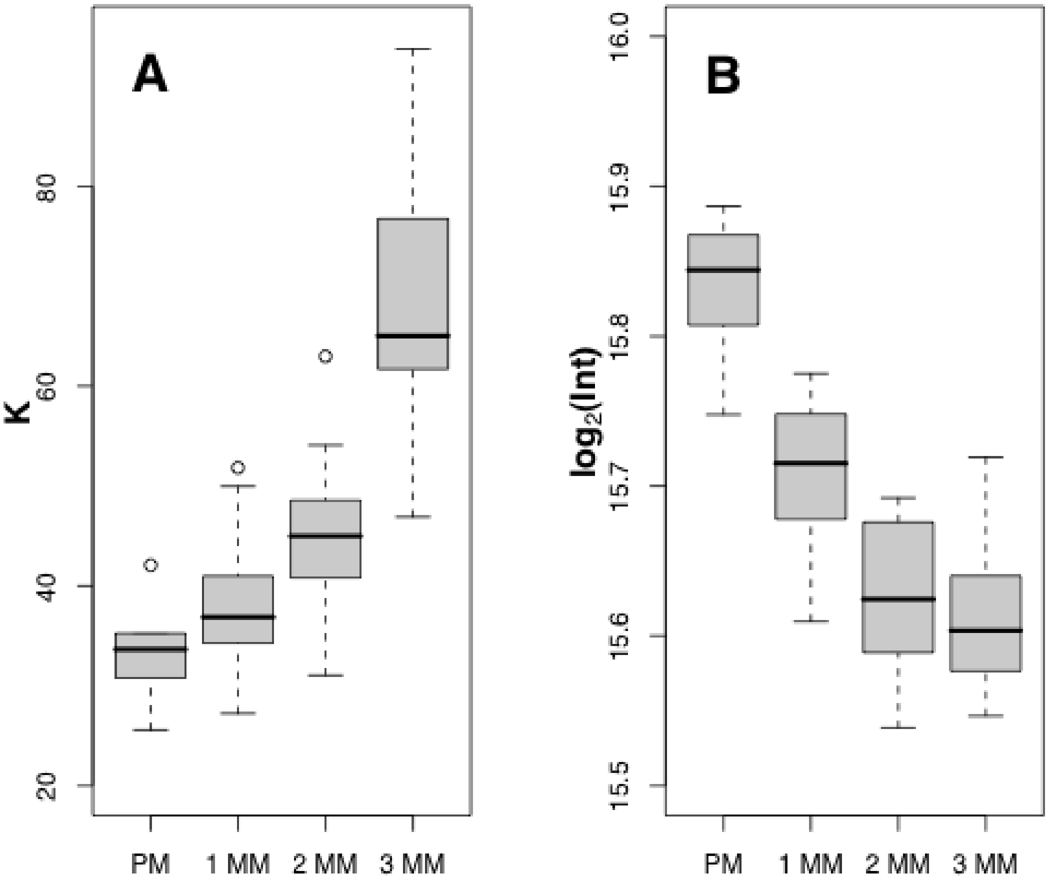

FIGURE 4.5 Langmuir parameters comparison.
Box plots of the fitted Langmuir parameters (A) affinity constant (the fitted parameter K in equation 4.1) and (B) saturation intensity (the fitted parameters A + bg in equation 4.1). PM: Perfect Match, 1MM: single-MM, 2MM: double-MM, and 3MM: triple-MM.

Figure 4.5A shows the affinity constant K for the four groups of probes. The figure clearly indicates how the number of mismatches affects the affinity of the probe. A small K value implies high probe affinity, whereas a large K value implies low probe affinity. As expected, PM probes have the smallest K values, followed by the single-MM probes with a slightly larger K values, the difference is statistically significant ($P < 0.05$, one-sided Wilcoxon test). The double-MM probes show significantly larger K values than the PM probes and the single-MM probes ($P < 2 \text{ X } 10^{-3}$, one-sided Wilcoxon test). The largest K values of all the probe groups are for the triple-MM probes ($P < 5 \text{ X } 10^{-4}$, one-sided Wilcoxon test). As expected, PM probes have the highest affinity to targets followed by single-MM probes then double-MM probes, and the triple-MM probes show the lowest affinity for their targets. This is emphasized in Figure 4.5B. High affinity probes hybridize better with their targets than low affinity probes and as a result show higher saturation intensities. PM probes show significantly higher saturation intensities than the MM probes ($P < 3 \text{ X } 10^{-5}$, one-sided Wilcoxon test). The saturation intensity decreases as the number of mismatches increase; double-MM probes ($P < 4 \text{ X } 10^{-5}$, one-sided Wilcoxon test) and triple-MM probes ($P < 5 \text{ X } 10^{-4}$, one-sided Wilcoxon test).

4.4.3 Relationship between signal intensity and ΔG

The results shown so far suggest detectable and significant differences between the four groups of probes. We aim to explain and predict these differences based on the physical chemistry of the hybridization and link them to probe design and optimization processes. We first choose free energy of probe-target duplex formation (ΔG) as our explanatory factor. ΔG has been used in several studies to explain the differences observed among short oligonucleotide probes [66, 71]. Moreover, ΔG can be easily calculated from probe sequence [48] and is one of the most frequently used parameters in microarray probe design [4].

We investigated the effects of mismatches on the signal intensity for each MM probe group compared to the perfect match group. We examined the relative signal intensity of the three groups of MM probes and find it, as expected, to decrease when the number of mismatches increases. The relative signal intensities across the eight target concentrations for single-MM probes, double-MM probes and triple-MM probes were: 85%, 70% and 48%, respectively. On the other hand, relative values of predicted ΔG for single-MM probes, double-MM probes and triple-MM probes were: 95%, 88% and 84%, respectively. We calculated relative intensities at each concentration separately and found them to be in agreement with the values reported above. Based on that, we can safely conclude that ΔG, generally, affects signal intensity. The lower the predicted ΔG value, the stronger the affinity between the probe and its target, which results in higher probe signal. But ΔG alone does not explain the variation seen at each separate concentration. To account for that, we tested how much of the difference in signal intensity at each target concentration can be explained by ΔG using a simple linear model like the one presented in equation 4.2, but using ΔG instead of *%Bound*. Examining the correlation between *log* signal intensity at each target concentration and predicted ΔG of duplex formation reveals a weak correlation between them (R^2 between 0.05-0.28) at the first seven concentrations. However, we find a stronger correlation (R^2 = 0.79) between *log* signal intensity and ΔG at the highest target concentration (5000 pM) (Fig. 4.6).

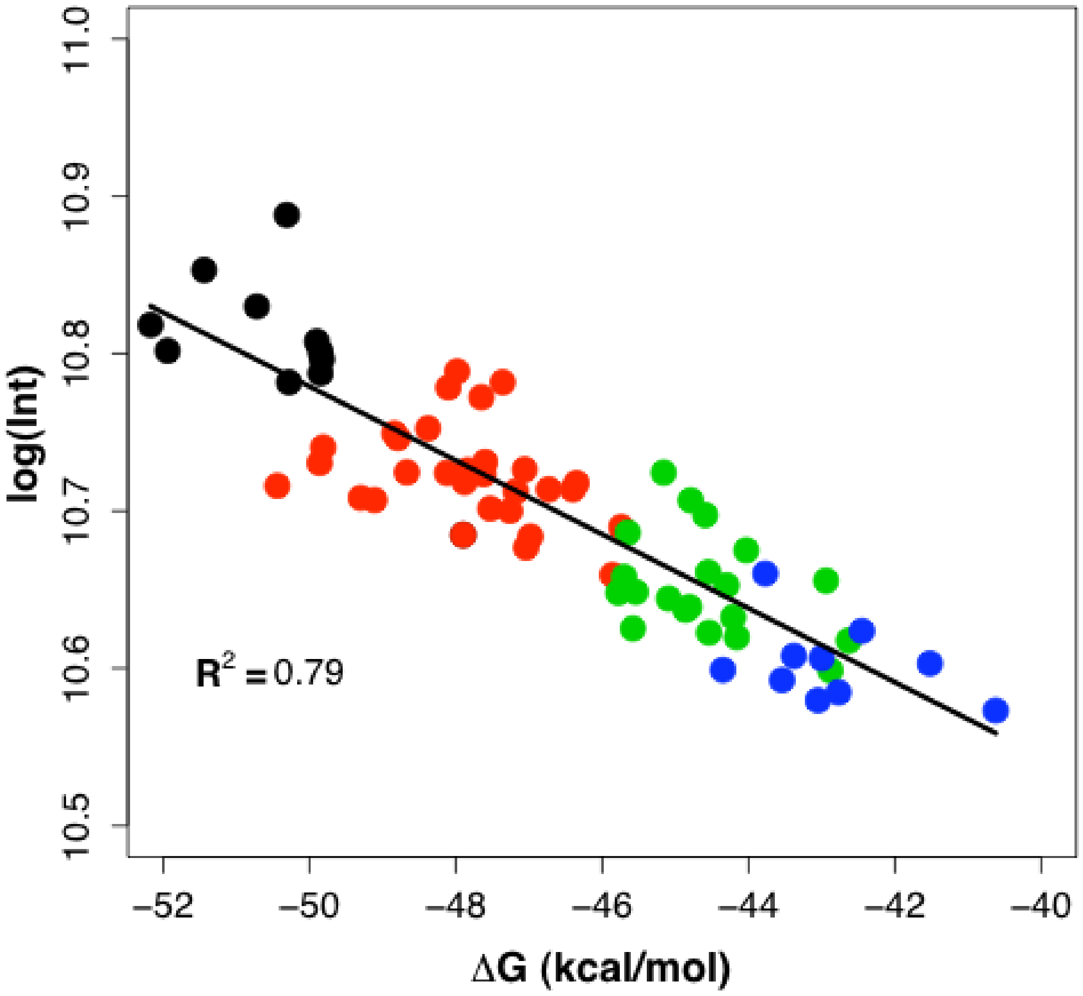

FIGURE 4.6 Relationship between probe-target *ΔG* and signal intensity. Relationship between probe-target free energy of duplex formation (*ΔG*) and hybridization signal intensity, at a target concentration of 5000 pM. Each dot represents one probe: PM probes are shown in black, single-MM probes are shown in red, double-MM probes are shown in green and triple-MM probes are shown in blue. The regression line is shown as a solid black line.

This observation suggests that in principle ΔG is a good, but insensitive, estimator of probe signal variations. It is important to emphasize that probes respond to different target concentration with a wide range of signal intensity, while probe-target ΔG changes little over the same range of probe and target concentrations, since the probe concentration is the driver and is constant and relatively high. It is important to mention that we are not claiming ΔG as unusable, rather we point out that ΔG alone is insensitive to variations in target concentration.

4.4.4 Explaining probe signal intensity variation using predicted probe percent bound (PPB)

Given the results above, ΔG alone cannot be used to explain the variations in signal intensity between different probes at different target concentrations. We shift our focus to probe percent bound (PPB), a quantity calculated by nucleic acid modeling software packages such as OMP which perform multi-state equilibrium modeling of each nucleic acid molecule and its binding partner or partners under the prevailing hybridization conditions. PPB can be defined, in microarray terms, as the percentage of each probe molecule that exists as a heterodimer with its target under the relevant hybridization conditions (see [142] for more details). PPB, like ΔG, can be easily calculated for each probe under different target concentrations, but unlike duplex ΔG, this quantity is highly sensitive to variations in target concentration when the probe saturation concentration (the highest target concentration after which no significant increase in signal intensity can be detected from each probe) or the approximate probe concentration is known for each probe.

We developed a simple linear model (equation 4.2), which relates probe observed signal intensity to the probe's PPB value. The model was fitted to data from the four probe groups (PM, single-MM, double-MM and triple-MM). Figure 4.7 shows a typical example of the results obtained for the PM probes. Figure 4.7A shows PPB vs. probe signal intensity represented by probe 5006 and Figure 4.7B shows a summary of R^2 of the fitted model (equation 4.2) and the p-value given the null hypothesis that the B_1 parameter in equation 4.2 is equal to zero for all the PM probes.

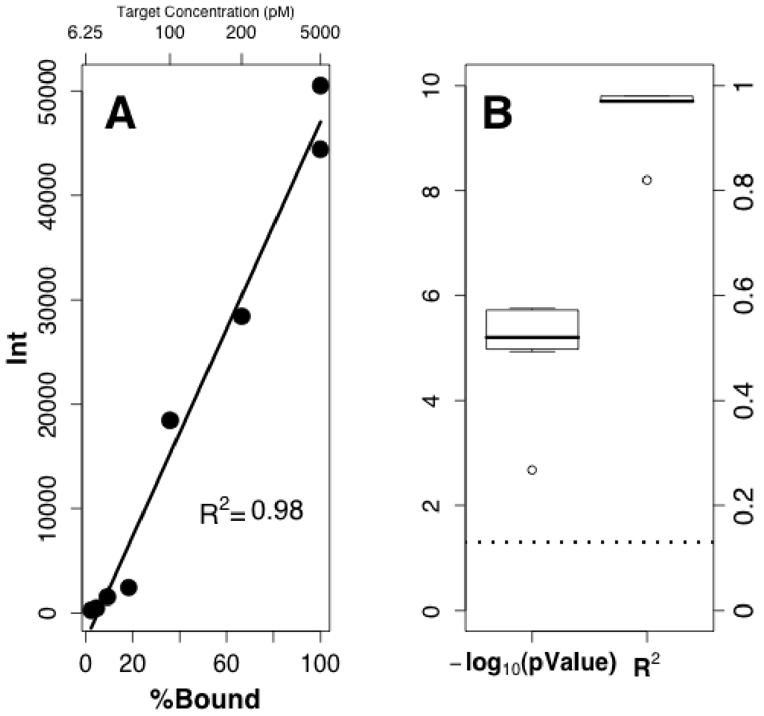

FIGURE 4.7 Probe percent bound is a sensitive predictor for both PM probe behavior and signal intensity.
(A) Relationship between probe signal intensity and predicted percent bound (PPB) at each target concentration for probe 5006 from the PM group. Points represent observed intensities, and the solid line represents the fit of the model (equation 2).
(B) Box plots for the obtained R^2 and p-values of the null hypothesis that the B_1 parameter in equation 4.2 is equal to zero from all PM probes. Dotted line indicates $P = 0.05$.

Clearly, PPB captures the variations in signal intensity according to target concentrations. Low target concentrations yield low PPB and therefore low signal intensity. Fitting the data to the model (equation 4.2), we found a strong correlation between PPB and signal intensity ($R^2 = 0.98$). This means that the variations in signal intensity can be explained by PPB alone. Fitting the model to intensity values from all ten PM probes (Fig. 4.7B) confirms that PPB is capable of capturing the variation in signal intensity with target concentration for all of the ten PM probes. Moreover, the null hypothesis, that the B_1 parameter in equation 4.2 is equal to zero, is rejected with high confidence (Fig. 4.7B). The same analysis was performed on single-MM probes, double-MM probes and triple-MM probes. The results for single-MM probes are presented in Figure 4.8. As in the case of PM probes, we see excellent correlation between PPB and signal intensity. The analyses of double-MM probes and triple-MM probes produced similar results and are shown in Figure 4.9 and Figure 4.10, respectively.

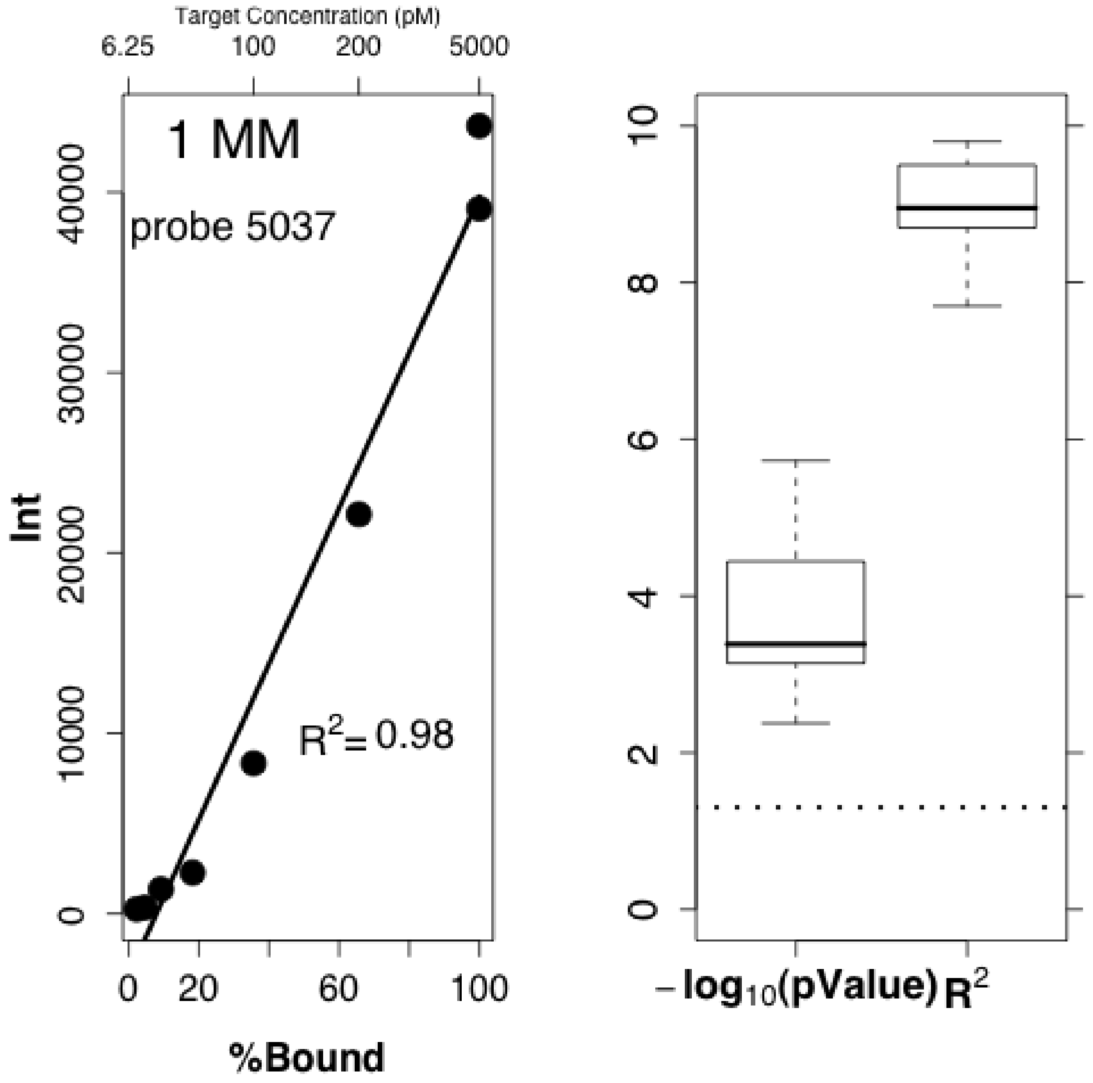

FIGURE 4.8 Probe percent bound is a sensitive predictor for both single-MM probe behavior and signal intensity.

(A) Relationship between probe signal intensity and predicted percent bound (PPB) at each target concentration for probe 5037 from the single-MM group. Points represent observed intensities, and the solid line represents the fit of the model (equation 4.2). (B) Box plots for the obtained R^2 and *p*-values of the null hypothesis that the B_1 parameter in equation 4.2 is equal to zero from all single-MM probes. Dotted line indicates $P = 0.05$.

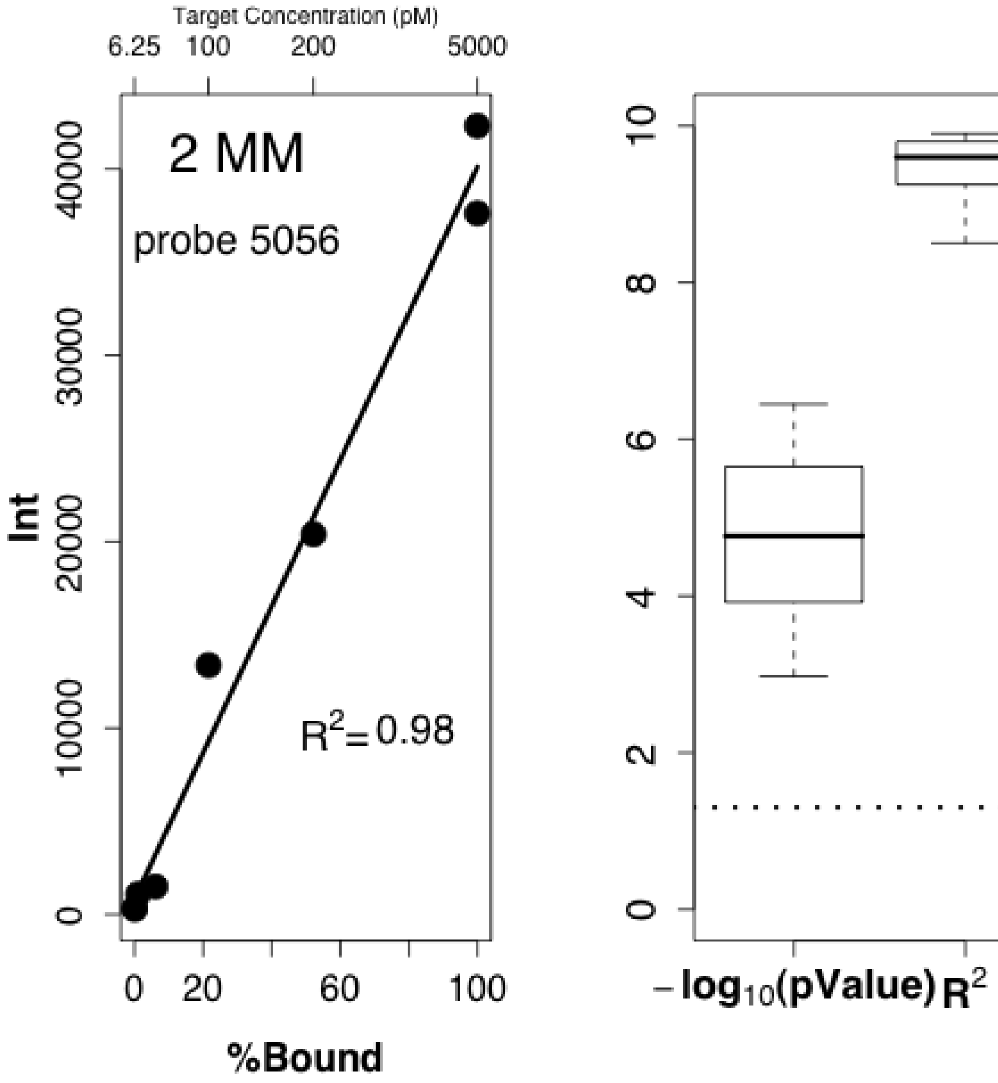

FIGURE 4.9 Probe percent bound is a sensitive predictor for both double-MM probe behavior and signal intensity.

(A) Relationship between probe signal intensity and predicted percent bound (PPB) at each target concentration for probe 5056 from the double-MM group. Points represent observed intensities, and the solid line represents the fit of the model (equation 4.2). (B) Box plots for the obtained R^2 and p-values of the null hypothesis that the B_1 parameter in equation 4.2 is equal to zero from all double-MM probes. Dotted line indicates $P = 0.05$.

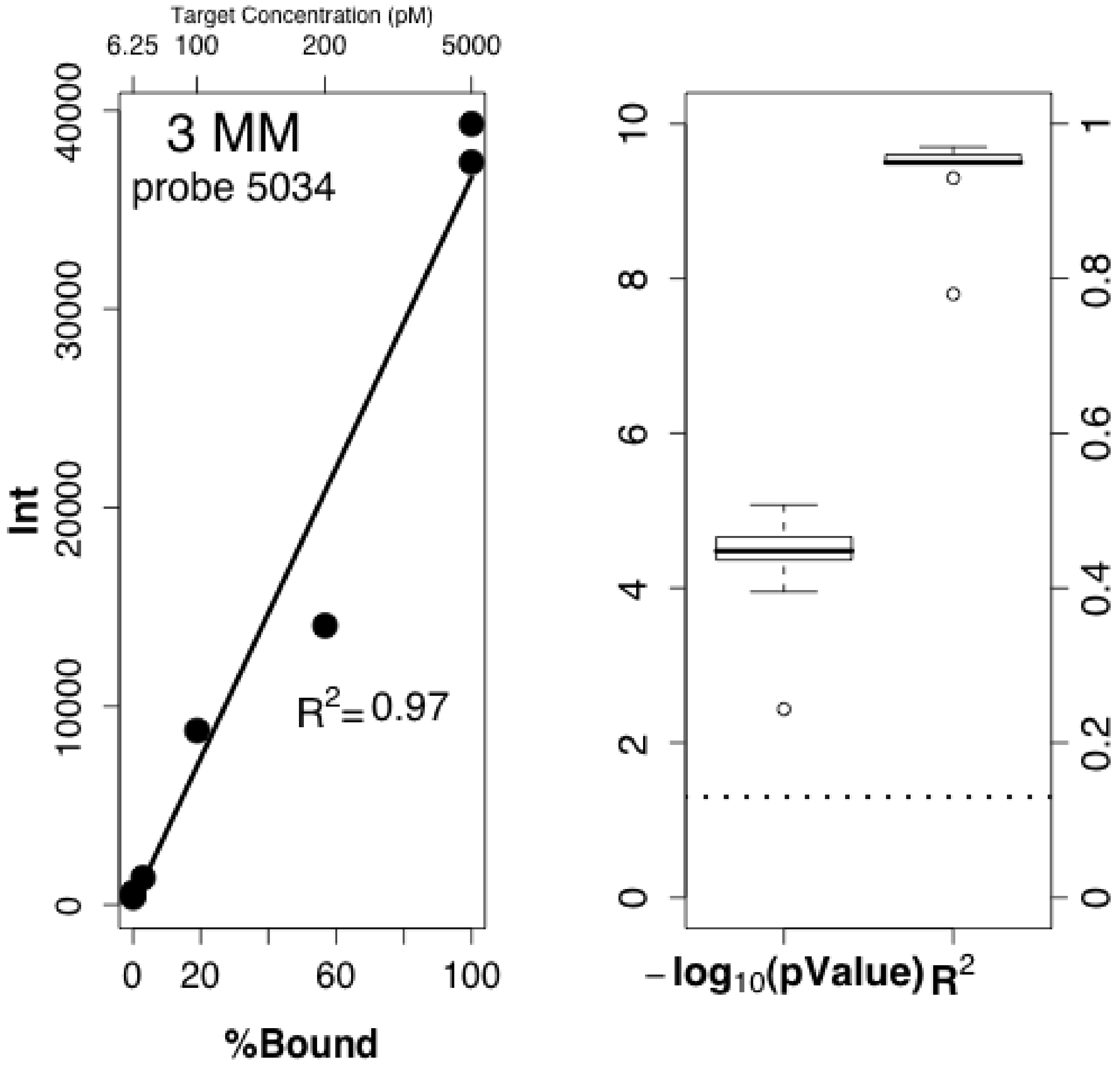

FIGURE 4.10 Probe percent bound is a sensitive predictor for both triple-MM probe behavior and signal intensity.

(A) Relationship between probe signal intensity and predicted percent bound (PPB) at each target concentration for probe 5034 from the triple-MM group. Points represent observed intensities, and the solid line represents the fit of the model (equation 4.2). (B) Box plots for the obtained R^2 and *p*-values of the null hypothesis that the B_1 parameter in equation 4.2 is equal to zero from all triple-MM probes. Dotted line indicates $P = 0.05$.

4.5 Discussion

Detection of low numbers of mismatches in long oligonucleotides. Given the large number of possible locations and permutations for mismatches, its not feasible to study all types of mismatches that can be made to a 50-mer probe. Our initial modeling (data not shown) and other studies [66, 129-130] have demonstrated that permutations at the middle of the probe will have a more significant effect on discrimination than permutations at the ends. Our current study focuses on single-, double- and triple-MM introduced at the central region of the probe. The identity of the mismatches follows Affymetrix GeneChip mismatch probe rules of homomeric transversions, i.e. A→T, T→A, G→C and C→G, which provides a good starting point for future modeling studies, and facilitates comparisons with the mismatch modeling and prediction results reported in the literature for Affymetrix GeneChip arrays.

Most of the studies available on mismatches in long oligonucleotides use very limited numbers of probe-target pairs. They deal with a relatively high and fixed number of mismatches (i.e. four mismatches in 50-mer probe) or measure signal only at a single target concentration [129, 149]. Here, we try to expand the available data by using a larger number of probe-target pairs, and we examine the behavior of different numbers of mismatches, over a range of target concentrations, in a consistent experimental context. The use of a range of concentrations is of special importance, since not all genes or genomic segments are present at the same concentration in the microarray context. Probes, especially those that have different sequence composition and consequently different thermodynamic properties, also respond differently to different concentrations of their targets. Based on that, rules for probe performance deduced under one concentration may not hold true under another concentration. The results presented in Figure 4.2 and 4.3 confirm the importance of target concentration. For single-MM probes, PM signal can be differentiated from MM signal only at higher target concentrations.

Probes with one mismatch tend to have similar responses even if they differ in the identity and the position of the mismatch. This lack of specificity is probably a result of the length of the probe and its ability to accommodate one base change, and the low cost in terms of ΔG of having one mismatch (an average of 5% less than the PM). On the other hand, none of the single-MM probes showed a signal higher than the PM signal, analogous to the common MM > PM phenomenon seen in short probes used for GeneChip arrays [77]. We attribute that to two causes: the first is the absence of targets at extremely low concentration; the second is our method of target labeling, which is not sequence content dependent. We examined the MM > PM phenomenon on GeneChips using the Latin square dataset and we see a correlation between target concentration and the severity of the MM > PM phenomenon. Most of the MM > PM signal is seen under target concentrations < 2 pM. The second cause is likely the end-labeling protocol used in our study, which is not biased by sequence content.

For the double-MM and triple-MM probe-target pairs, we also notice a dependency on target concentration. Generally, we can detect double-MM and triple-MM signal differences at lower concentrations than we can detect single-MM differences. Medium target concentrations (100 and 200 pM) gave the most statistically significant results. The identity and the exact position of the mismatch in double-MM probes, made no significant difference in the detected signal, as observed with the single-MM probes.

The comparatively lower signal intensity with increasing number of mismatches is due to the destabilizing effect of mismatches and the high cost of the presence of double- and triple-MM in the probe, in terms of binding energy. Duplexes containing a double mismatch had nearest-neighbor ΔG 12% less than the corresponding perfect match duplex, and triple mismatches reduced ΔG by 16%. The signal intensity difference between perfect match and triple-MM probes is similar to what is reported elsewhere in the literature, for probes of the same length and same number of mismatches [132], and provides a confirmation of earlier studies done with very few example probe-target pairs.

Predicting microarray outcomes using multi-state equilibrium models of solution hybridization. The widespread use of array applications demands an in-depth understanding of the dynamics of array hybridization. Regardless of the array application, the quality of the results depends heavily on appropriate interpretation of the detected signal. Understanding the factors that affect this signal is of importance to both the scientists who design the array and those who use it. While the current experiment is greatly simplified relative to, for example, a genomic expression array, and factors such as internal structure and diffusion of large target molecules have been eliminated, the results of this study demonstrate that probe percent bound (PPB), as predicted using a standard computational method for equilibrium modeling of systems of interacting oligonucleotides, is an excellent predictor of detected signal intensity on the microarray surface. This is significant because the current state of the art in hybridization modeling does not include well-determined models of surface effects. If predicted solution behavior of oligonucleotide molecules can be used as a reasonable approximation of surface behavior, the interpretation of microarray results becomes immediately better informed by our understanding of hybridization chemistry.

It has long been assumed that solution models can be used as a proxy for behavior at the microarray surface under the right conditions, but there has been relatively little experimental confirmation of this assumption [80, 129]. The results presented in this study serve as a proof of principle for the validity of using PPB as a sensitive predictor for both probe behavior and signal interpretation. Given the straightforward computational procedure for calculating PPB, it is not difficult or computationally expensive to consider this factor in probe design and selection. PPB can also be incorporated into array analysis algorithms to enhance the quality of the results. The only limitation of PPB calculation is its dependence on the probe saturation concentration or the approximate probe concentration. However, for most commercial microarray platforms, this quantity is either published, or an approximation can be provided by the array manufacturer or determined from independent optimization results reported in the literature (see [71] for an example). For other platforms, the probe saturation concentration or the approximate probe concentration can be deduced by a simple target concentration titration experiment. The results obtained regarding PPB support previous research regarding the importance of incorporating biophysical factors in both the design and analysis of microarrays [62, 64, 66-67, 77].

Our primary goal in this experiment was to understand and explain the differences between perfect match and mismatched duplex concentrations on the microarray via modeling of biophysical parameters. Therefore we also examined the relationship between probe-target binding energy (ΔG) and signal intensity. The results suggest that ΔG alone is not as capable of explaining all the observed differences in signal with the same accuracy as does probe percent bound (PPB). The correlation between ΔG and the observed signal intensity increases as the target concentration increases (Fig. 4.6). This is in agreement with a recent modeling study presented by Li *et al.* [71] in which a good correlation between ΔG and signal intensity is shown at the highest target concentrations. This result supports the notation of a universal fundamental principle for on-chip hybridization, but it does not accurately represent the concentration dependence of hybridization chemistry in general.

Fitting the Langmuir isotherm model to our data showed clearly that both perfectly matched and mismatched probes reach chemical saturation at high target concentrations (Fig. 4.4). Again, this emphasizes the importance of target concentration in both experimental design and modeling, and shows that long oligonucleotide probes have a physical hybridization profile similar to their shorter kin. This opens the door for further improvements in the analysis of long oligonucleotide probes. Given the advantages of using physical models in the analysis of microarray data, approaches like the one developed by Abdueva *et al*. [67] can now be applied to long oligonucleotide arrays. This will enhance the quality of the results obtained from these arrays and will also help shed more light on the physical principles of on-chip hybridization. Examining the fitted parameters of the Langmuir isotherm models for the four groups of probes (Fig. 4.5) confirms the similarities of their physical hybridization profile to that of shorter probes. Hekstra *et al.* [62] reported similar results when comparing the fitted parameters between PM and single-MM probes when the Langmuir isotherm model is applied to Affymetrix GeneChip probes. This supports the concept of a fundamental common ground of on-chip hybridization.

Impact on array design and analysis. Target concentration is, of course, the principal unknown quantity in standard array applications, and determining exact target concentration as an input for biophysical modeling at the design stage of a microarray experiment is not possible. However, the results of this experiment are significant from an experimental design point of view. During the process of probe design and selection, most probe design applications that consider concentration in their design process model relatively high target concentrations, usually 50 nM or 1 μM [13, 21, 27]. Based on the results obtained in this experiment for the array platform used, it is likely that these concentrations are too high and will not give the best result; instead, the entire array design will be optimized for response to high copy number targets. We therefore recommend the use of either 100 or 200 pM concentration should be considered in biophysical modeling to support probe selection for array platforms like the one used in this study. For other array platforms, the optimal target concentration can be determined based on simple optimization procedure, after taking probe saturation concentration or approximate probe concentration into account.

From the position of enabling improved retrospective analysis of experimental results, given consistent concentration-dependent behavior, it should be possible to project target concentration from intensity across the experiment either based on a spike-in calibration mixture or from calibration functions based on predictions over a range of concentrations combined with multi-state solution hybridization modeling. In fact, we have developed a new model for array analysis that shows promising results when applied to the array platform used here. The model is able to predict target concentration based on probe signal intensity and PPB, with an average R^2 between nominal and predicted target concentrations ≥ 0.8. We mention this result here to give the reader an understanding of the practical application of the work reported in this paper. However, we are in the process of testing the model on other array platforms, and the complete results will appear in a subsequent manuscript.

In conclusion, we have demonstrated that single mismatches in long oligonucleotide probe/target pairs are detectable, and that additional mismatches give progressively larger decreases in intensity, but that with respect to the three central positions in a 50-mer, the position of those mismatches does not give a diagnostic intensity difference. One of the many factors an analyst would like to know, in order to understand hybridization behavior on long oligonucleotide microarrays, is the effect of mismatches, including their presence, number, nucleotide identity and location. This study explores the limit of observable mismatch effects in 50mers; at all but the lowest concentrations of target, single, double and triple mismatches between the probe and target are detectable and significant. Our second significant finding is that binding predictions, in terms of predicted probe percent bound (PPB), derived from multi-state solution hybridization models, are strongly correlated with microarray signal, at least in cases where many variables have been controlled. While it remains to be seen whether this finding will hold for more complex probe-target systems, this study suggests that appropriately parameterized solution models of hybridization will accurately represent interactions on the DNA microarray surface.

This study presents additional experimental evidence that the use of proper thermodynamic modeling yields probes that have better performance in terms of specificity, sensitivity, noise and bias [44]. It demonstrates that even with the limitations of current probe design tools and array analysis algorithms, satisfactory results can be achieved when the hybridization of probe and target is understood and appropriate biophysical factors are taken into account during the design and the analysis steps. This study also highlights the validity of solution nucleic acid hybridization modeling and prediction approaches in the microarray context. Although the current approaches are not perfect, they provide a good starting point for further developments. Most of the tools used in this study provided satisfactory results within a reasonable margin of error. As more enhancements to this field are introduced, our ability to relate observed microarray signal directly to predicted hybridization behavior will help researchers in many disciplines overcome the current limitations of microarray technology.

CHAPTER 5: ACCURATE ESTIMATES OF MICROARRAY TARGET CONCENTRATION FROM A SIMPLE SEQUENCE-INDEPENDENT LANGMUIR MODEL

5.1 Abstract

Microarray technology is a commonly used tool for assessing global gene expression. Many models for estimation of target concentration based on observed microarray signal have been proposed, but in general these models have been complex and platform-dependent. We introduce a universal simple model, characterized by only three free parameters. We find that this model, which ignores all sequence-based features of DNA probes, yields excellent predictions across different microarray platforms, including Affymetrix, Agilent, Illumina and a custom microarray developed in our lab. We demonstrate that a generalized 3-parameter Langmuir model can equal or even outperform models that explicitly incorporate sequence properties. In doing so, we eliminate the need for approaches that incorporate detailed models of the sequence. From a microarray design perspective, the results obtained here suggest that with a "spiked-in" concentration series targeting as few as 5-10 genes, reliable estimation of target concentration can be achieved for the entire microarray.

5.2 Introduction

DNA microarrays [1] are a primary research tool for assessing global gene expression. Structurally, a microarray is a solid surface on which nucleic acid strands (probes) are attached. Functionally, they operate on the principle of nucleic acid complementarity between the attached probes and components of the target mixture (a mixture of labeled nucleic acids). The result is formation of a stable duplex, from which a signal is detected at each probe if there is a complementary molecule present in the labeled target mixture. This signal is then used in further analysis and inference steps.

Models that attempt to estimate target concentrations on microarrays can be, generally, divided into two main categories: The first includes models that rely on the Langmuir isotherm [62, 66-67, 150], which in its simplest form is a hyperbolic response function in the form of:

$$I = d + a\,c \,/\, (b + c)$$

where I is the signal intensity from a given microarray probe at target concentration c, and a, b and d are the model fitting parameters. The model has three free parameters (a, b and d) fitted to different target concentrations. The fitting parameter a is the saturation intensity (assuming $d = 0$), b is the target concentration that saturates half of the probes, and d is the background component [62]. Some of the models in this category predict these parameters from probe sequence composition [62] or probe/target and target/target binding energy [66, 150]. Other models [67] fit the data to the Langmuir isotherm and obtain a, b and d

for each probe using a non-linear minimization approach. In all of these models, each probe is characterized by its own *a*, *b* and *d*. If a microarray has *n* probes or probesets, then there are 3*n* parameters. Once the three parameters are determined, target concentration is predicted by inverting the isotherm. A second category of models depend on competitive hybridization chemistry [40, 69, 71] to predict probe signal intensity, which is translated either to expression level or absolute target concentration. Those models are based on the thermodynamics of hybridization, and parameterized based on in-solution DNA hybridization [28, 48]. They rely on individual probe properties and consequently are prone to over-parameterization.

We developed a simple probe-property-independent model to predict absolute target concentration on different microarray platforms, including Affymetrix, Agilent, Illumina and a locally developed custom microarray. Our predictions of target concentration on these microarray platforms outperform previous models. We report the first approach that works on multiple array platforms, using fewer parameters than most other models.

5.3 Methods

5.3.1 The Langmuir isotherm

The Langmuir isotherm is a hyperbolic response function in the form of:

$$I_j = a\frac{c_j}{b+c_j} + d \qquad \text{Eq. 5.1}$$

where I_j is the signal intensity from the probes at target concentration *j*. *a*, *b* and *d* are the model fitting parameters, and *c* is the target j^{th} concentration in pM. This model has three free parameters (*a*, *b* and *d*) fitted to different concentrations, depending on the dataset used. The fitting parameter *a* is the saturation intensity (if there is no cross-hybridization, i.e. $d = 0$), *b* is the target concentration that saturates half of the probes, and *d* is the background component [62]. The model was fitted using the *nls* function of R [151]. In contrast with commonly used approaches, the three parameters were obtained by fitting the model to data from a number of probes (training probes) and not specifically to individual probes.

5.3.2 Estimation of target concentration

To estimate target concentration ($\hat{x}$), we used the approach described by Burden *et al*. [146] with a slight modification:

$$\hat{x} = \left\{ \begin{array}{l} X,\ if\, I > \hat{a} + \hat{d} \\ \hat{b}(I - \hat{d}) / (\hat{a} + \hat{d} - I),\ if\, \hat{d} < I < \hat{a} + \hat{d} \\ Y,\ if\, I < \hat{d} \end{array} \right\} \qquad \text{Eq. 5.2}$$

where $\hat{a}$, $\hat{b}$ and $\hat{d}$ are the fitted parameters of equation 1 above. *X* is an arbitrarily chosen large concentration, assigned when the probe has signal intensity above the Langmuir saturation limit. *Y* is an arbitrarily chosen small concentration, assigned when the probe has signal intensity below the predicted background limit. *X* and *Y* were set above the largest target concentration and below the smallest target concentration in each dataset, respectively.

In this report we divide spike-ins into three categories: low, medium and high, following McCall *et al*. [152]. For Figures 5.1, 5.2, 5.3, 5.4, 5.6, 5.7 and 5.8, we do not estimate target concentrations for all spike-ins in the low concentration category. These data are provided in Figure 5.5.

5.3.3 Datasets

Four control datasets were used to evaluate the performance of the model. The first three datasets (Affymetrix HGU133A GeneChip Latin Square dataset, Agilent 4x44K Whole Human Genome Oligo Microarray control dataset and Illumina's Human-6 v2 Beadchip control dataset) are part of the External RNA Control Consortium [153]. Full descriptions of those datasets, along with the raw data can be found here [152] and in references therein. The last dataset is an ArrayIt 50-mer control dataset spotted on a standard epoxysilane-coated glass slide substrate, described previously in [154].

5.3.4 Algorithms and data manipulation

Two previously described algorithms (for Affymetrix GeneChip) were used in this study for performance evaluation. The first algorithm is developed by Abdueva *et al*. [67] and the second is developed by Li *et al*. [71]. The source code and data for both algorithms were obtained from the authors. Signal intensities were normalized using quantile normalization [155] for the Abdueva *et al*. procedure. For Li *et al*., signal intensities were used without normalization and prepared according to the author's instructions [71]. Briefly, the raw signal intensities from 355 probes corresponding to 19 transcripts (out of 42 transcripts) fitted the authors filtering procedure and were used for estimating target concentration. Probe intensity was taken as the average across technical replicates. For our model, all signal intensities were used without normalization (unless indicated). The signal intensity of each probe was taken as the average signal across technical replicates. R^2 and *slope* values presented here were calculated using the *lm* function of R [151] using the default settings, except that the intercept term was omitted.

5.3.5 Training set and number of probes

For the purpose of determining the values of $\hat{a}$, $\hat{b}$ and $\hat{d}$, the global average model (GLAM) requires a training set of known spike-ins. In this report we follow a standard *N choose K*, where *N* is the total number of spike-ins and *K* is the number of probes or probesets included in the training set. *K* has a value between one and *N* minus one. To illustrate, consider Figure 5.9, which shows the results on Affymetrix U133A control dataset. In this dataset there are 42 spike-ins, thus *K* (*x axis* of Fig. 5.9) has a range from 1 to 41. When there are more than 42 possible combinations, we choose 42 at random and we take them as a representative for all possible combinations. To illustrate this mechanism, consider the first box plot in Figure 5.9, it represents all the 42 combinations of 42 choose 1. We call this leave-41-out, which means that GLAM was trained on one probeset and predicted the remaining 41 spike-ins. The second box plot is for 42 choose 2, since there are 861 different combinations, we shuffle the list of all the 42 probesets, then we choose two probesets at random and run the model, we repeat this process 42 times, thus each box plot in Figure 5.9 have 42 data points. The last box plot is for 42 choose 41, we call this leave-one-out, which means that the GLAM was trained on 41 probesets and predicted the concentration of the remaining spike-in.

5.4 Results

As an alternative to commonly used approaches (see introduction) for estimating target concentrations on microarrays, we tested the performance of a model in which the three parameters of the Langmuir isotherm were fit to all the data from each microarray. Instead of characterizing each probe or probeset with its own *a*, *b* and *d*, we characterize a group of experiments with one *a*, *b* and *d*, thus reducing the number of free parameters to three for each microarray. This model, which we call the global average model (GLAM), has the advantage that, unlike other algorithms [8, 67, 71] it can be fit with spike-in dose-response data from a small number of genes and make predictions for the entire microarray. Also, unlike most other algorithms [69, 81, 110] it is applicable to microarrays that don't have multiple probes per probeset. We tested the performance of GLAM on control datasets from each of the most popular microarray platforms. We describe our results for each of these control datasets below.

5.4.1 Estimation of target concentration on the Affymetrix platform

The Affymetrix U133A Latin square control dataset has 42 transcripts spiked in at concentration range of 0.125-512 pM in a Latin square fashion [88, 152]. We apply the GLAM model presented in equation 5.1 (see Methods) to this dataset. We obtained $\hat{a}$, $\hat{b}$ and $\hat{d}$ by fitting the model to training set composed of three randomly chosen probesets (Fig. 5.1A and 5.2A; red symbols). Figures 5.1 and 5.2 show that GLAM is able to recover absolute target concentration with R^2 of 0.99.

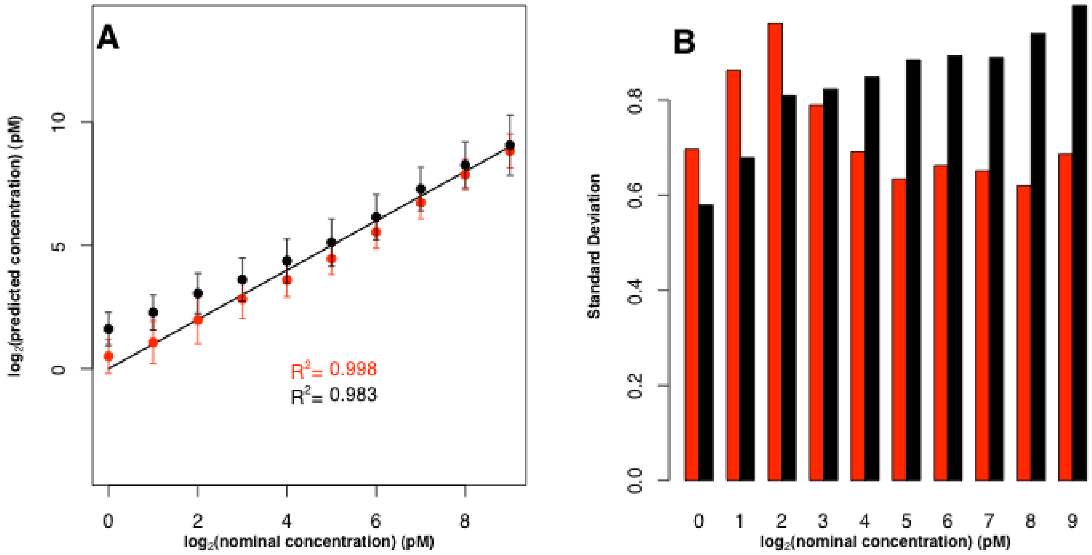

FIGURE 5.1 Estimation of transcript concentrations on the Affymetrix platform for 19 transcripts chosen by Li *et al*..
(A) Results obtained from a training set of three probesets for GLAM (red) and those obtained from Li *et al*. approach (black). Error bars are standard deviations. The solid line is the identity line (*x*=*y*). (B) Comparison of error bar lengths for each concentration for our approach (red) and Li *et al*. approach (black).

To evaluate the consequences of ignoring probe specific effects we compared the performance of GLAM to other algorithms. Figure 5.1A compares GLAM to absolute target estimates from Li *et al*. [71].

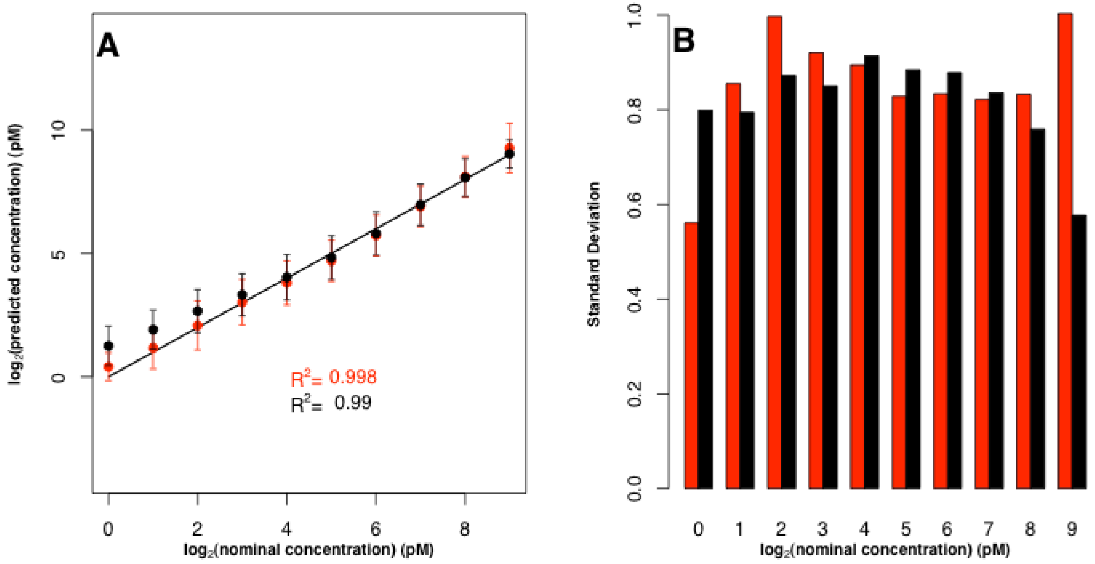

FIGURE 5.2 Estimation of transcript concentrations on the Affymetrix platform for all 42 transcripts in the Latin Square dataset.
(A) Results obtained from a training set of three probesets for GLAM (red) and those obtained from Abdueva *et al*. approach (black). Error bars are standard deviations. The solid line is the identity line (*x*=*y*). (B) Comparison of error bar lengths for each concentration for our approach (red) and Abdueva *et al*. approach (black).

Their approach depends on competitive hybridization chemistry, and target concentration is determined by the following equation [71]:

$$\hat{T} = \frac{\bar{S}}{A} + \frac{k_d \gamma}{Ap / \bar{S} - k_d / k_b - 1} \qquad \text{Eq. 5.3}$$

where $\hat{T}$ is the predicted target concentration, $\bar{S}$ is the observed signal intensity after scanner bias and background subtraction, A is the detection coefficient of fluorescence, k_d is the probe affinity coefficient, γ is a cross-hybridization factor, p

is the total number of probes in molar concentration units and k_b is the binding rate for target molecules [71].

Their estimates are based on a subset of 19 transcripts, which were selected based on target sequence alignment matching and probe signal intensity, and sorted based on probe thermodynamic properties. We estimate transcript concentrations for these 19 transcripts using GLAM, choosing three randomly selected probesets as a training set (Fig. 5.1). The results show that both approaches are able to recover target concentration with high R^2 (0.998 for GLAM and 0.983 for Li *et al*. [71]). Absolute target concentrations obtained using our approach show a *slope* of 0.958 and those obtained using the approach of Li *et al*. [71] have a *slope* of 1.045. The *slope* value describes the accuracy of the predictions [88]; a value of 1 is considered to be the perfect score. Values below or above 1 indicate underestimation or overestimation, respectively. Although the Li *et al*. model attempts to control for factors that might increase error (such as scanner bias), the errors between the two approaches are comparable (Fig. 5.1B) with GLAM having slightly higher errors for target concentrations less than 4 pM, but much lower errors for target concentrations > 4 pM.

Abdueva *et al*., [67] developed a Langmuir-based approach similar to GLAM. The main difference between the two is probe effects. In the Abdueva *et al*. [67] approach; $\hat{a}$, $\hat{b}$ and $\hat{d}$ are estimated for each probe, and the final transcript concentration is calibrated based on *log* predicted saturation intensity and *log* non-specific intensity of the probe. Those two values are predicted from probe thermodynamic properties, based on sequence content. In GLAM, $\hat{a}$, $\hat{b}$ and $\hat{d}$ are global, based only on a training set. We applied both approaches to the U133A Affymetrix control dataset, and estimated the absolute concentrations of the 42 transcripts. The results are presented in Figure 5.2. Both approaches perform well (Fig. 5.2A) but GLAM has a higher R^2 (0.998) than the Abdueva *et al*. approach (0.990) despite using significantly fewer free parameters. Examining the *slope* of the predicted target concentrations shows that GLAM predictions have a *slope* of 0.997, while the Abdueva *et al*. [67] predictions' *slope* is 1.007. Both approaches have similar error values, as shown in Figure 5.2B.

The Abdueva *et al*. [67] predictions are based on normalized signal intensities, while ours are based on raw signal intensities. To provide a fair comparision, we also predicted target concentrations using GLAM on the quantile normalized signal intensities used by Abdueva *et al*. [67] (Fig. 5.3). Similar results are obtained, with slight differences in the length of error bars and a *slope* of 0.994 for GLAM. The use of normalization does not appear to affect our conclusions.

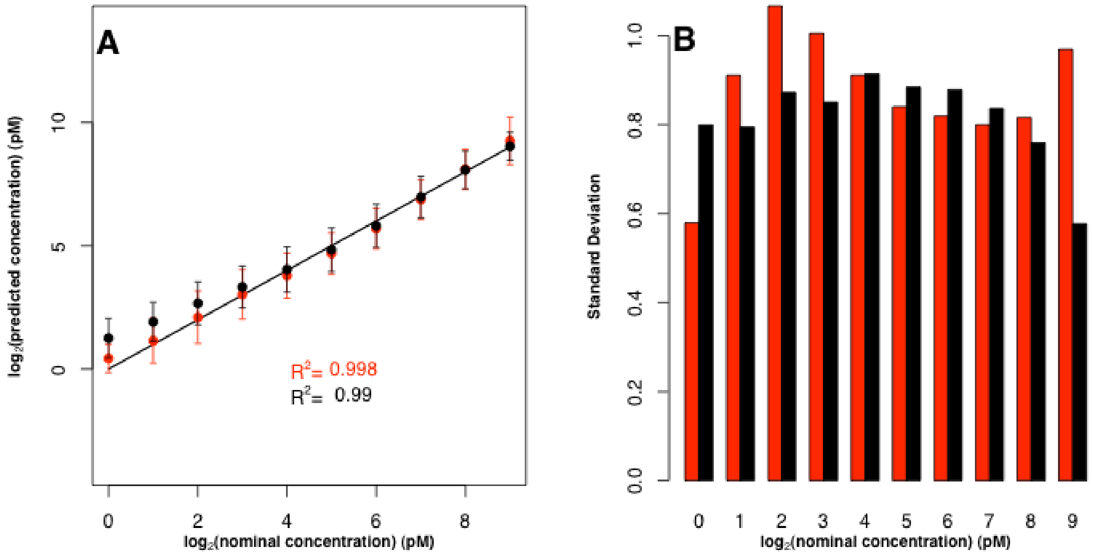

FIGURE 5.3 Estimation of transcript concentrations on the Affymetrix platform using quantile normalized signal intensities.
(A) Results obtained from a training set of three probesets for GLAM (red) and those obtained from Abdueva *et al*. approach (black). Error bars are the standard deviations of the 42 transcripts. The solid line is the identity line (*x*=*y*). (B) Comparison of error bar lengths for each concentration for our approach (red) and Abdueva *et al*. approach (black).

In the interest of complete assessment, we compared the performance of GLAM to the model of Abdueva *et al*. [67], but without calibrating the final transcript concentration (i.e. without including any probe properties). The only difference between these two approaches is that GLAM has a single set of parameters for *a*, *b* and *d* while Abdueva *et al*. model each probe individually. Figure 5.4 shows that removing probe properties causes the performance of Abdueva *et al*. model [67] to degrade, R^2 dropped to 0.843 and *slope* dropped to 0.53, while GLAM is unaffected.

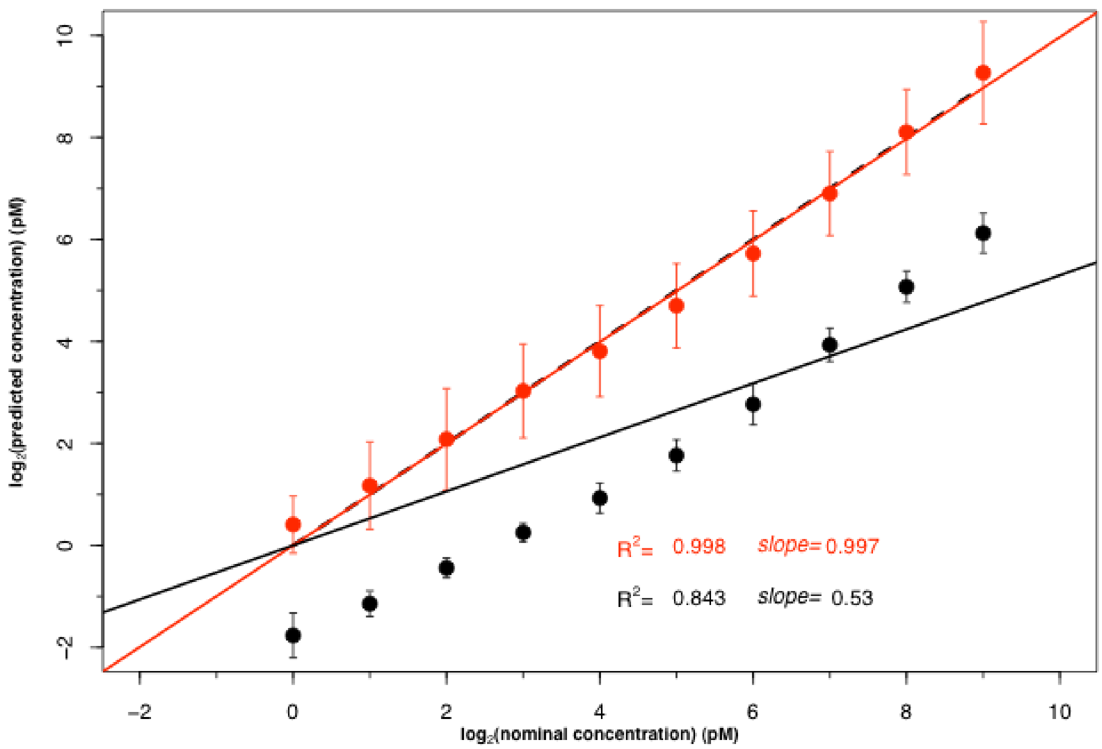

FIGURE 5.4 Performance comparison between GLAM (red) and probe-property-independent Abdueva *et al*. approach (black) on the Affymetrix platform. Results obtained from a training set of three randomly chosen probesets for GLAM. Error bars are the standard deviations of the 42 transcripts. The dashed line is the identity line (x=y); solid lines are the regression lines. R^2 and *slope* values are colored coded according to the schema above and indicated on the graph.

While our method yielded excellent predictions of absolute transcript concentration, we did not predict concentration for all transcripts in the low concentration category (Fig. 5.1 and 5.2). This is because there is a poor correlation between signal intensity and target concentration at the low end [152], and because the signal obtained from these targets can't be differentiated from background noise

[86]. Also, microarray scanner nonlinearity is at its worst at low intensity [71, 156].

We show the results of predicting the full range of concentrations in Figure 5.5.

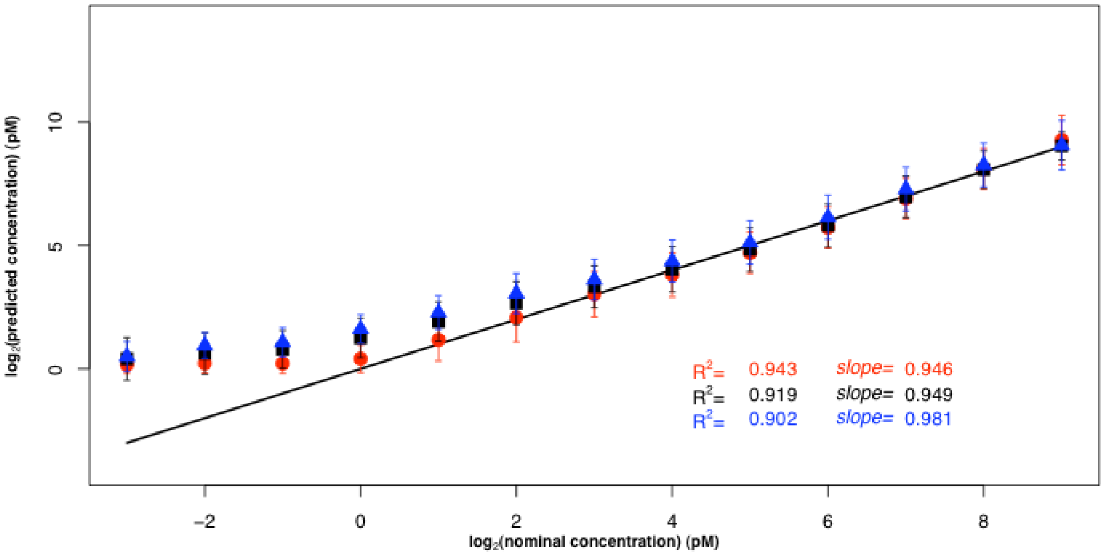

FIGURE 5.5 Estimation of transcript concentrations on the Affymetrix platform using the full range of concentrations (14 total).
Results obtained from a training set of three randomly chosen probesets for GLAM (red circles). Abdueva *et al*. approach results are shown as black squares and Li *et al*. approach results are shown as blue triangles. Error bars are the standard deviations of the 42 transcripts in the case of GLAM and Abdueva *et al*. approach and 19 transcripts in the case of Li *et al*. approach. The solid line is the identity line (*x*=*y*). R^2 and *slope* values are colored coded according to the schema above and indicated on the graph.

Although the three models show a decrease in terms of R^2 and *slope* values when low concentration transcripts are considered, the overall results do not affect our conclusions. All of the models, including GLAM, have the same difficulty predicting low target concentrations.

5.4.2 Estimation of target concentration on the Agilent platform

We next tested the applicability of GLAM to the Agilent platform, which has different probe and surface properties than the Affymetrix platform. A publically available Agilent control dataset is composed of ten transcripts spiked at ten concentrations [152]. We predicted transcript concentrations using GLAM, again without taking probe effects into consideration. Figure 5.6A shows the results using a summary of leaving-one-out procedure, where every nine probes were used as a training set and the resulting $\hat{a}$, $\hat{b}$ and $\hat{d}$ were used to estimate the concentration of the remaining tenth transcript. The average estimated concentrations agree well with the reported nominal concentrations with an R^2 of 0.999 and a *slope* of 0.997.

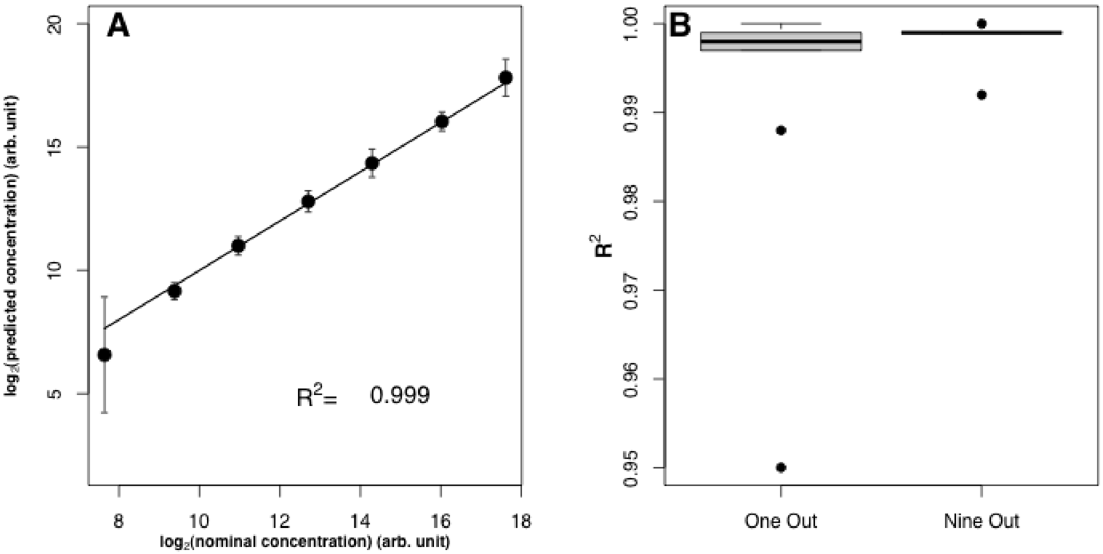

FIGURE 5.6 Estimation of transcript concentrations on the Agilent platform. (A) Results obtained from a comprehensive leave-one-out procedure. Error bars are the standard deviations of the ten transcripts. The solid line is the identity line (*x*=*y*). (B) Box plots of R^2 for the ten estimations of leave-one-out procedure and R^2 for five estimations of leave-nine-out procedure.

The above-mentioned R^2 was obtained by training GLAM on nine probes and predicting the remaining tenth transcript, which raised the possibility of overfitting. We then tested the effect of changing the fraction of total probes included in the training set, since a spike-in control procedure would be most useful if it could be trained on a small fraction of the array data. Figure 5.6B shows box plots of R^2 for the ten estimations of the leave-one-out procedure described above, and R^2 for five estimations of a leave-nine-out procedure. In the leave-nine-out procedure, $\hat{a}$, $\hat{b}$ and $\hat{d}$ were estimated from a training set of one probe and used to predict the concentrations of the remaining nine transcripts. The leave-nine-out procedure uses a small training set that is sensitive to the choice of probe for training. Probes showing non-Langmuir-like behavior can be avoided without explicit modeling and knowledge of their sequence, so five probes returning unphysical (negative) values for either $\hat{a}$, $\hat{b}$ or $\hat{d}$, were not used to predict target concentration. Transcript concentration estimation with parameters obtained from well-behaved single probes show excellent R^2 with a minimum of 0.992, depending on which probe was used for parameter estimation.

5.4.3 Estimation of target concentration on the Illumina platform

We also tested GLAM on an Illumina control dataset composed of 34 transcripts spiked at 11 different concentrations [152]. We follow the same procedures as with the Agilent platform, and the results are shown in Figure 5.7. Application of a comprehensive leave-one-out procedure (Fig. 5.7A) shows that our approach to estimating transcript concentration performs well on the Illumina platform; the average estimated concentrations show an R^2 of 0.992 and a *slope* of 1.165. The R^2 values obtained from 34 trials of the comprehensive leave-one-out procedure are shown in Figure 4B. Out of 34 probes, 18 probes returned unphysical values for one of the parameters and therefore were not used to train the model for target concentration prediction. We were able to use the remaining 16 probes in a leave-33-out procedure with excellent results, and the R^2 values are shown in Figure 4B.

It is clear from Figure 4A that GLAM underestimates transcript concentrations of 0.1 and 0.3 pM and overestimates transcript concentrations of 300 and 1000 pM on the Illumina platform.

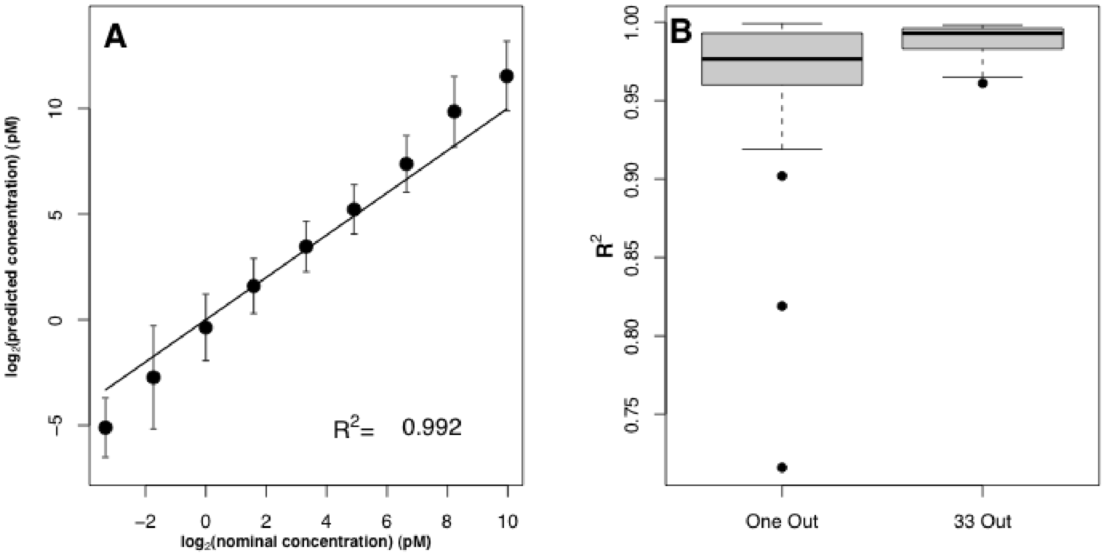

FIGURE 5.7 Estimation of transcript concentrations on the Illumina platform. (A) Results obtained from a comprehensive leave-one-out procedure. Error bars are the standard deviations of the 34 transcripts. The solid line is the identity line (*x*=*y*). (B) Box plots of R^2 for the ten estimations of leave-one-out procedure and R^2 for five estimations of leave-33-out procedure.

However, the regression slope values reported by McCall *et al.* [152] for this platform suggest that there is simply poor agreement between signal intensity and nominal spike-in concentration in those ranges.

5.4.4 Estimation of target concentration on a pin-spotted platform

Our pin-spotted array data set is a custom 50mer array that was developed in our laboratory and described in [154]. The platform is similar to many custom microarrays, where probes are contact-spotted using a robot. The platform differs from commercially available platforms in the attachment chemistry. The control experiment that uses this array has ten targets spiked at eight different concentrations [154]. We follow the same steps used for the above datasets and we estimate target concentrations for this dataset by obtaining $\hat{a}$, $\hat{b}$ and $\hat{d}$ using either a leave-one-out or leave-nine-out procedure.

Figure 5.8A shows the averaged predicted target concentrations for a leave-one-out procedure with an R^2 of 0.992 and a *slope* of 0.969. A leave-nine-out procedure (Fig. 5.8B) shows that even one probe was sufficient to retain $R^2 \geq 0.95$. Of the ten probes, one returned unphysical values for one of the parameters and was not used for estimating target concentration.

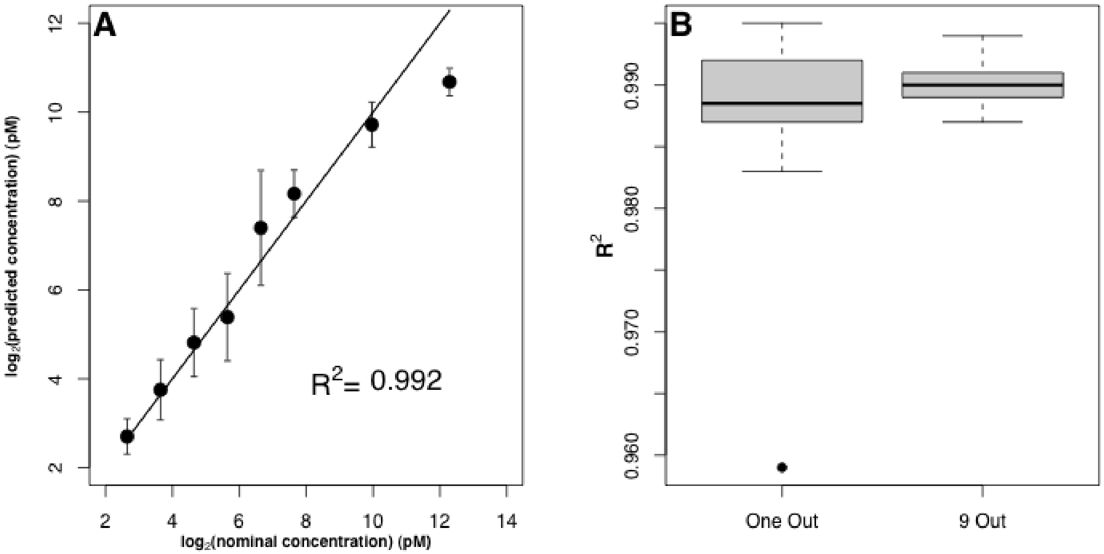

FIGURE 5.8 Estimation of transcript concentrations on the pin-spotted platform. (A) Results obtained from a comprehensive leave-one-out procedure. Error bars are the standard deviations of the ten transcripts. The solid line is the identity line (x=y). (B) Box plots of R^2 for the ten estimations of leave-one-out procedure and R^2 for five estimations of leave-nine-out procedure.

Although the R^2 and *slope* was lowest for this dataset, the model was able to produce acceptable target estimates. We believe the slight difference in model behavior for this platform was due to the different attachment chemistry, and to the presence of competing mismatch probes for each target in this dataset .

5.4.5 How many training probes are necessary for GLAM?

We examined the effect of varying the number of probes/probesets included in the training set on the performance of GLAM. We ran GLAM on the four datasets used in this study and considered all the possible numbers of training probes/probesets. The performance of GLAM in terms of R^2 is shown in Figure 5.9 for the Affymetrix

U133A Latin square control dataset. Figure 5.9 shows that five probesets are enough for GLAM to return reliable results.

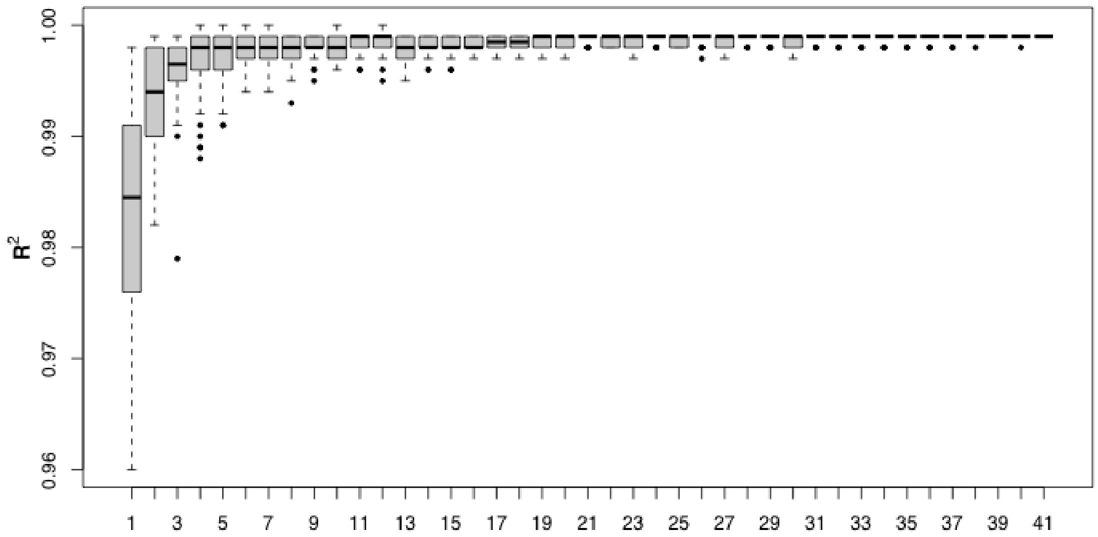

FIGURE 5.9 Effect of varying the number of probesets included in the training set on the performance of GLAM on the Affymetrix U133A control dataset.
Each Box plot shows the obtained R^2 (*y axis*) for 42 choose K (*x axis*) of the estimated target concentrations using the Affymetrix U133A control dataset. Each box plot has 42 data points (see Methods section of this chapter).

The effect of training set size on the performance of GLAM for the other three datasets is shown in Figures 5.10-5.12. Note that the *x-axis* in Figure 5.9 is the number of probesets (each probeset has 11 probes in general); while in Figures 5.10-5.12 *x-axis* is the number of individual probes.

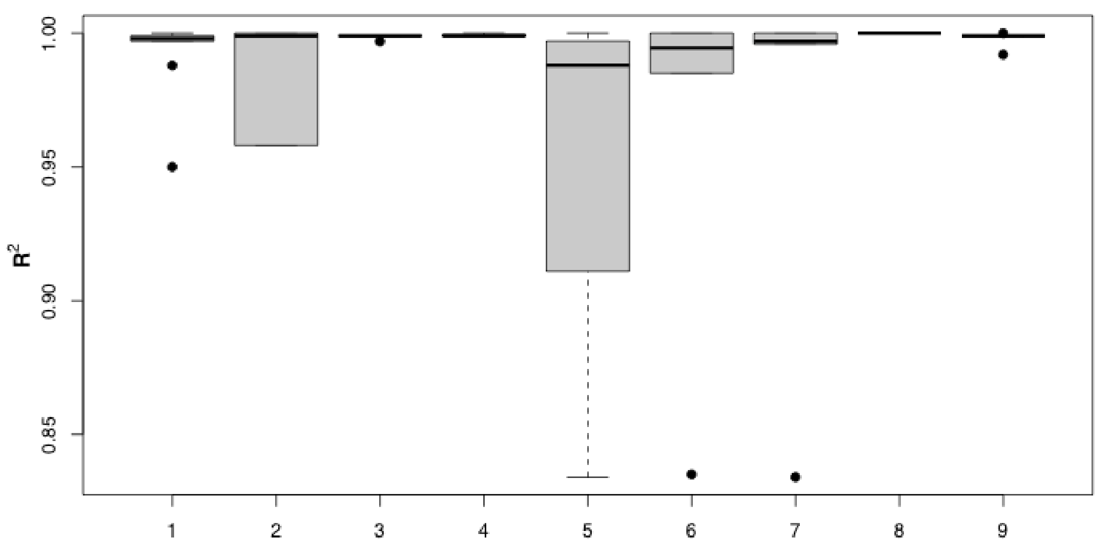

FIGURE 5.10 Effect of varying the number of probes included in the training set on the performance of GLAM on the Agilent control dataset.
Each Box plot shows the obtained R^2 (*y axis*) for 10 choose K (*x axis*) of the estimated target concentrations using the Agilent control dataset. Each box plot has 10 data points.

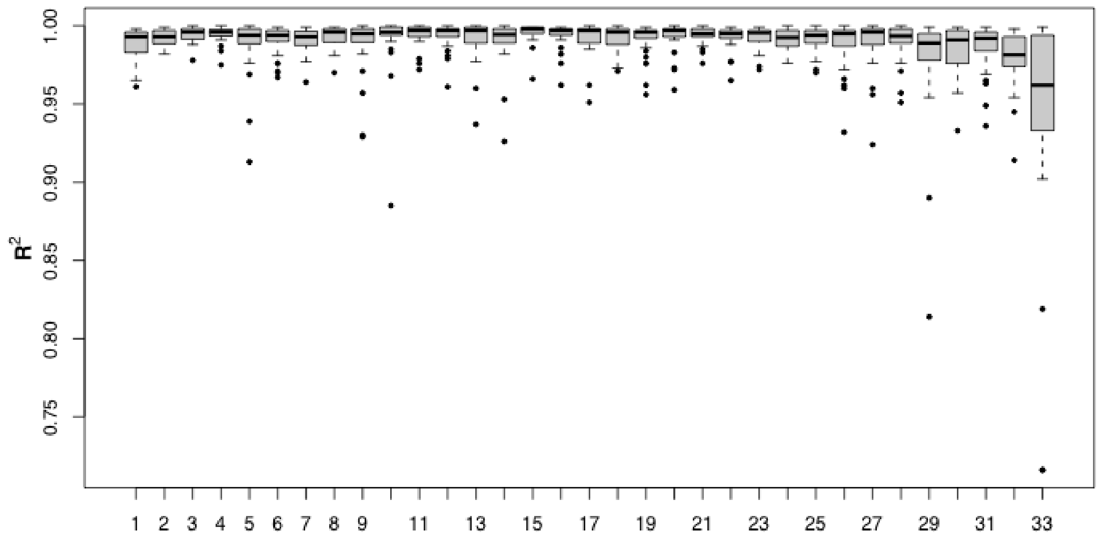

FIGURE 5.11 Effect of varying the number of probes included in the training set on the performance of GLAM on the Illumina control dataset.

Each Box plot shows the obtained R^2 (*y axis*) for 34 choose K (*x axis*) of the estimated target concentrations using the Illumina control dataset. Each box plot has 34 data points.

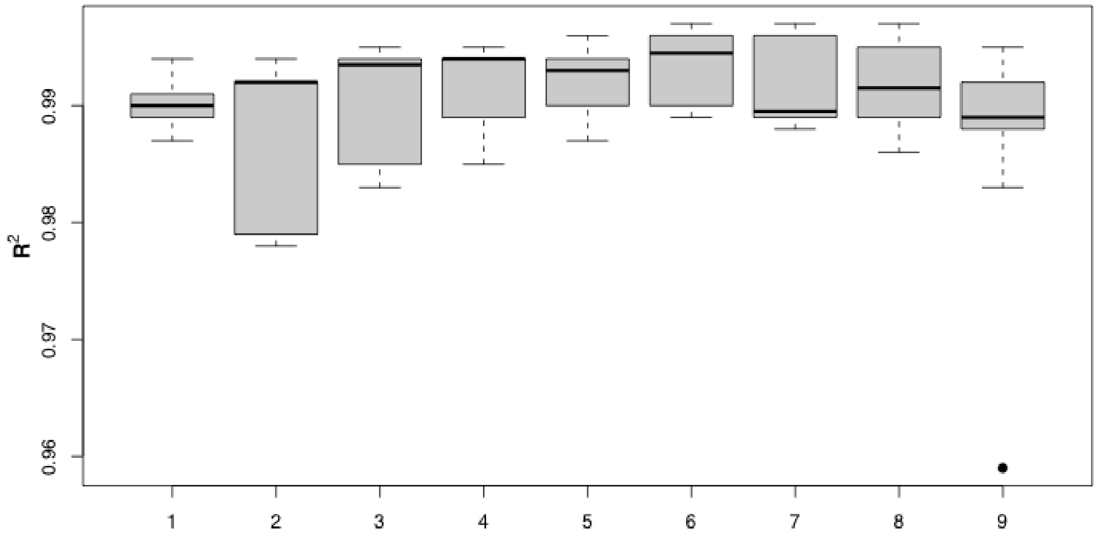

FIGURE 5.12 Effect of varying the number of probes included in the training set on the performance of GLAM on the pin-spotted control dataset.
Each Box plot shows the obtained R^2 (*y axis*) for 10 choose K (*x axis*) of the estimated target concentrations using the pin-spotted control dataset. Each box plot has 10 data points.

5.5 Discussion

Many approaches have been used to relate microarray probe properties to hybridization signal intensity [60, 79]. In this report, we showed that a simple model that captures array-wide binding parameters is comparable in performance to models that use per-probe parameters. We compared results from our GLAM approach to the results of two algorithms [67, 71], which have been demonstrated to be the best-performing of the Langmuir-based and hybridization chemistry-based algorithm types. Our results show that despite the differences in probe design and sequence, probe effects average out and may be modeled globally to recover specific transcript concentrations.

Obtaining $\hat{a}$, $\hat{b}$ and $\hat{d}$ for a training set of probes, then using those values to predict the behavior of other probes, implies that all probes have the same $\hat{a}$, $\hat{b}$ and $\hat{d}$,. We know from past studies that each probe has its own $\hat{a}$, $\hat{b}$ and $\hat{d}$, which are generally dependent on sequence composition [62]. Given that basic microarray probe design procedures usually require that all probes have similar GC content, on a carefully-designed array it may be sufficient to use global $\hat{a}$, $\hat{b}$ and $\hat{d}$ to parameterize the Langmuir isotherm. Fine-tuning parameters to reflect the differences of each probe based on its sequence composition and thermodynamic properties, or based on the observed response of each probe, may be unnecessary and likely leads to overfitting. What we find important is predicting these parameters from probes that show Langmuir-like response, and as we have shown, this should be enough to ensure reliable results (Figures 5.9-5.12). The number of probes or probesets used in the training set does not seem to affect the performance of our model as long as the probes included in the training set show Langmuir-like response or their number is sufficient to average the effect of other probes that do not follow Langmuir-like response (Figures 5.9-5.12). Using probes that do not follow Langmuir-like response to estimate $\hat{a}$, $\hat{b}$ and $\hat{d}$ (i.e. have negative values for any of these parameters) will degrade the performance of GLAM. This can be avoided by including more probesets in training GLAM or by using a set of probes that are demonstrated to follow a Langmuir-like response.

In conclusion, we have shown that using a simple form of the Langmuir isotherm model, with a minimum of parameters and assumptions and without explicit modeling of individual probe properties, we were able to recover absolute transcript concentrations with high R^2 on four different array platforms. To our knowledge, this is the first report to produce a working model that is equally valid for four of the most frequently used microarray array formats. Given the choice of models with equivalent performance, Occam's razor dictates that the model with the fewest free parameters is to be preferred. Our results therefore suggest that, despite considerable efforts by the bioinformatics community [60, 67, 70-71], the additional complexity introduced by models that attempt to use probe characteristics to improve estimates of absolute concentration is not justified by a corresponding increase in performance. Given consistent concentration-dependent behavior, it should be possible to project target concentration from intensity across the experiments based on a spike-in calibration mixture containing only few probes.

REFERENCES

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

2. Stoughton, R.B. 2005. Applications of DNA microarrays in biology. Annu Rev Biochem, 74:53-82.

3. Dufva, M. 2005. Fabrication of high quality microarrays. Biomol Eng, 22:173-184.

4. Kreil, D.P., Russell, R.R. and Russell, S. 2006. Microarray oligonucleotide probes. Methods Enzymol, 410:73-98.

5. Heller, M.J. 2002. DNA microarray technology: devices, systems, and applications. Annu Rev Biomed Eng, 4:129-153.

6. Murphy, D. 2002. Gene expression studies using microarrays: principles, problems, and prospects. Adv Physiol Educ, 26:256-270.

7. Tomiuk, S. and Hofmann, K. 2001. Microarray probe selection strategies. Brief Bioinform, 2:329-340.

8. Bishop, J., Blair, S. and Chagovetz, A.M. 2006. A competitive kinetic model of nucleic acid surface hybridization in the presence of point mutants. Biophys J, 90:831-840.

9. Koltai, H. and Weingarten-Baror, C. 2008. Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. Nucleic Acids Res, 36:2395-2405.

10. Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B. and DeRisi, J.L. 2003. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol, 4:R9.

11. Rimour, S., Hill, D., Militon, C. and Peyret, P. 2005. GoArrays: highly dynamic and efficient microarray probe design. Bioinformatics, 21:1094-1103.

12. Rouillard, J.M., Herbert, C.J. and Zuker, M. 2002. OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics, 18:486-487.

13. Rouillard, J.M., Zuker, M. and Gulari, E. 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res, 31:3057-3062.

14. Charbonnier, Y., Gettler, B., Francois, P., Bento, M., Renzoni, A., Vaudaux, P., Schlegel, W. and Schrenzel, J. 2005. A generic approach for the design of whole-genome oligoarrays, validated for genomotyping, deletion mapping and gene expression analysis on Staphylococcus aureus. BMC Genomics, 6:95.

15. Mrowka, R., Schuchhardt, J. and Gille, C. 2002. Oligodb--interactive design of oligo DNA for transcription profiling of human genes. Bioinformatics, 18:1686-1687.

16. Tolstrup, N., Nielsen, P.S., Kolberg, J.G., Frankel, A.M., Vissing, H. and Kauppinen, S. 2003. OligoDesign: optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling. Nucl Acids Res, 31:3758-3762.

17. Wang, X. and Seed, B. 2003. Selection of oligonucleotide probes for protein coding sequences. Bioinformatics, 19:796-802.

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

19. Chen, H. and Sharp, B. 2002. Oliz, a suite of Perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3' untranslated region. BMC Bioinformatics, 3:27.

20. Gordon, P.M.K. and Sensen, C.W. 2004. Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays. Nucl Acids Res, 32:e133.

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

22. Xu, D., Li, G., Wu, L., Zhou, J. and Xu, Y. 2002. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics, 18:1432-1437.

23. Kaderali, L. and Schliep, A. 2002. Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics, 18:1340-1349.

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

25. Rahmann, S. 2002. Rapid Large-Scale Oligonucleotide Selection for Microarrays. In: First IEEE Computer Society Bioinformatics Conference (CSB): 2002: IEEE Press.

26. Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G. and Fayard, J.M. 2004. ROSO: optimizing oligonucleotide probes for microarrays. Bioinformatics, 20:271-273.

27. Nordberg, E.K. 2005. YODA: selecting signature oligonucleotides. Bioinformatics, 21:1365-1370.

28. SantaLucia, J., Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A, 95:1460-1465.

29. SantaLucia, J., Jr., Allawi, H.T. and Seneviratne, P.A. 1996. Improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry, 35:3555-3562.

30. Schildkraut, C. 1965. Dependence of the melting temperature of DNA on salt concentration. Biopolymers, 3:195-208.

31. Breslauer, K.J., Frank, R., Blocker, H. and Marky, L.A. 1986. Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A, 83:3746-3750.

32. Le Novere, N. 2001. MELTING, computing the melting temperature of nucleic acid duplex. Bioinformatics, 17:1226-1227.

33. Nussinov, R., Shapiro, B., Le, S.Y. and Maizel, J.V., Jr. 1990. Speeding up the dynamic algorithm for planar RNA folding. Math Biosci, 100:33-47.

34. Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol, 132:365-386.

35. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. J Mol Biol, 215:403-410.

36. Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res, 31:3406-3415.

37. Smith, T.F., Waterman, M.S. and Fitch, W.M. 1981. Comparative biosequence metrics. J Mol Evol, 18:38-46.

38. Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet, 22:101-109.

39. Ramakrishnan, R., Dorris, D., Lublinsky, A., Nguyen, A., Domanus, M., Prokhorova, A., Gieser, L., Touma, E., Lockner, R., Tata, M., Zhu, X., Patterson, M., Shippy, R., Sendera, T.J. and Mazumder, A. 2002. An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. Nucleic Acids Res, 30:e30.

40. Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F.C., Shen, M.M., Lu, G., Fang, J., Liu, W.M., Ryder, T., Kaplan, P., Kulp, D. and Webster, T.A. 2003. Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci U S A, 100:11237-11242.

41. Suzuki, S., Ono, N., Furusawa, C., Kashiwagi, A. and Yomo, T. 2007. Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. BMC genomics, 8:373.

42. Ono, N., Suzuki, S., Furusawa, C., Agata, T., Kashiwagi, A., Shimizu, H. and Yomo, T. 2008. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. Bioinformatics, 24:1278-1285.

43. Wetmur, J.G. 1976. Hybridization and renaturation kinetics of nucleic acids. Annu Rev Biophys Bioeng, 5:337-361.

44. Luebke, K.J., Balog, R.P. and Garner, H.R. 2003. Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts. Nucleic Acids Res, 31:750-758.

45. Drmanac, R. and Drmanac, S. 2001. Sequencing by hybridization arrays. Methods Mol Biol, 170:39-51.

46. Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., Shah, N., Thomas, D., Fan, J.B., Gingeras, T., Warrington, J., Patil, N., Hudson, T.J. and Lander, E.S. 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat Genet, 24:381-386.

47. Drmanac, R., Drmanac, S., Chui, G., Diaz, R., Hou, A., Jin, H., Jin, P., Kwon, S., Lacy, S., Moeur, B., Shafto, J., Swanson, D., Ukrainczyk, T., Xu, C. and Little, D. 2002. Sequencing by hybridization (SBH): advantages, achievements, and opportunities. Adv Biochem Eng Biotechnol, 77:75-101.

48. SantaLucia, J., Jr. and Hicks, D. 2004. The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct, 33:415-440.

49. Borer, P.N., Dengler, B., Tinoco, I.J. and Uhlenbeck, O.C. 1974. Stability of ribonucleic acid double-stranded helices. J Mol Biol, 86:843-853.

50. Xia, T., SantaLucia, J., Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C. and Turner, D.H. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry, 37:14719-14735.

51. Peyret, N., Seneviratne, P.A., Allawi, H.T. and SantaLucia, J., Jr. 1999. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. Biochemistry, 38:3468-3477.

52. Allawi, H.T. and SantaLucia, J., Jr. 1998. Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. Biochemistry, 37:2170-2179.

53. Allawi, H.T. and SantaLucia, J., Jr. 1998. Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. Biochemistry, 37:9435-9444.

54. Allawi, H.T. and SantaLucia, J., Jr. 1998. Thermodynamics of internal C.T mismatches in DNA. Nucleic Acids Res, 26:2694-2701.

55. Allawi, H.T. and SantaLucia, J., Jr. 1997. Thermodynamics and NMR of internal G.T mismatches in DNA. Biochemistry, 36:10581-10594.

56. Arghavani, M.B., SantaLucia, J., Jr. and Romano, L.J. 1998. Effect of mismatched complementary strands and 5'-change in sequence context on the thermodynamics and structure of benzo[a]pyrene-modified oligonucleotides. Biochemistry, 37:8575-8583.

57. Blake, R.D., Bizzaro, J.W., Blake, J.D., Day, G.R., Delcourt, S.G., Knowles, J., Marx, K.A. and SantaLucia, J., Jr. 1999. Statistical mechanical simulation of polymeric DNA melting with MELTSIM. Bioinformatics, 15:370-375.

58. Bommarito, S., Peyret, N. and SantaLucia, J., Jr. 2000. Thermodynamic parameters for DNA sequences with dangling ends. Nucleic Acids Res, 28:1929-1934.

59. Bonnet, G., Tyagi, S., Libchaber, A. and Kramer, F.R. 1999. Thermodynamic basis of the enhanced specificity of structured DNA probes. Proc Natl Acad Sci U S A, 96:6171-6176.

60. Binder, H. and Preibisch, S. 2006. GeneChip microarrays—signal intensities, RNA concentrations and probe sequences. J Phys Condens Matter, 18:S537-S566.

61. Heim, T., Wolterink, J.K., Carlon, E. and Barkema, G.T. 2006. Effective affinities in microarray data. J Phys Condens Matter, 18:S525-S536.

62. Hekstra, D., Taussig, A.R., Magnasco, M. and Naef, F. 2003. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. Nucleic Acids Res, 31:1962-1968.

63. Halperin, A., Buhot, A. and Zhulina, E.B. 2004. Sensitivity, specificity, and the hybridization isotherms of DNA chips. Biophys J, 86:718-730.

64. Halperin, A., Buhot, A. and Zhulina, E.B. 2006. On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions. J Phys Condens Matter, 18:S463-S490.

65. Atkins, P.W. and De Paula, J. 2006. Atkins' Physical chemistry, 8th edn. New York: W.H. Freeman.

66. Held, G.A., Grinstein, G. and Tu, Y. 2003. Modeling of DNA microarray data by using physical properties of hybridization. Proc Natl Acad Sci U S A, 100:7575-7580.

67. Abdueva, D., Skvortsov, D. and Tavare, S. 2006. Non-linear analysis of GeneChip arrays. Nucleic Acids Res, 34:e105.

68. Wu, Z., Irizarry, R., Gentleman, R., Murillo, F.M. and Spencer, F. 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. J Am Stat Assoc, 99:909-917.

69. Zhang, L., Miles, M.F. and Aldape, K.D. 2003. A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol, 21:818-821.

70. Gharaibeh, R.Z., Fodor, A.A. and Gibas, C.J. 2008. Background correction using dinucleotide affinities improves the performance of GCRMA. BMC Bioinformatics, 9:452.

71. Li, S., Pozhitkov, A. and Brouwer, M. 2008. A competitive hybridization model predicts probe signal intensity on high density DNA microarrays. Nucleic Acids Res, 36:6585-6591.

72. Langdon, W.B., Upton, G.J.G. and Harrison, A.P. 2009. Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips. Brief Bioinform, 10:259-277.

73. Glazer, M., Fidanza, J.A., McGall, G.H., Trulson, M.O., Forman, J.E., Suseno, A. and Frank, C.W. 2006. Kinetics of oligonucleotide hybridization to photolithographically patterned DNA arrays. Anal Biochem, 358:225-238.

74. Burden, C.J. 2008. Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed. Phys Biol, 5:16004.

75. Zuker, M. 2000. Calculating nucleic acid secondary structure. Curr Opin Struct Biol, 10:303-310.

76. Southern, E., Mir, K. and Shchepinov, M. 1999. Molecular interactions on microarrays. Nat Genet, 21:5-9.

77. Naef, F. and Magnasco, M.O. 2003. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Phys Rev E Stat Nonlin Soft Matter Phys, 68:011906.

78. Gharaibeh, R.Z., Fodor, A.A. and Gibas, C.J. 2007. Software Note: Using probe secondary structure information to enhance Affymetrix GeneChip background estimates. Comput Biol Chem, 31:92-98.

79. Binder, H. 2006. Thermodynamics of competitive surface adsorption on DNA microarrays. J Phys Condens Matter, 18:S491-S523.

80. Hooyberghs, J., Van Hummelen, P. and Carlon, E. 2009. The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. Nucleic Acids Res, 37:e53.

81. Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc Natl Acad Sci U S A, 98:31-36.

82. Wu, Z. and Irizarry, R.A. 2005. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol, 12:882-893.

83. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol, 14:1675-1680.

84. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat, 4:249-264.

85. Forman, J.E., Walton, I.D., Stern, D., Rava, R.P. and Trulson, M.O. 1998. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesized oligonucleotide arrays. ACS Symp Ser, 682:206-228.

86. Chudin, E., Walker, R., Kosaka, A., Wu, S.X., Rabert, D., Chang, T.K. and Kreder, D.E. 2002. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. Genome Biol, 3:R5.1.

87. Nielsen, H.B., Gautier, L. and Knudsen, S. 2005. Implementation of a gene expression index calculation method based on the PDNN model. Bioinformatics, 21:687-688.

88. Irizarry, R.A., Wu, Z. and Jaffee, H.A. 2006. Comparison of Affymetrix GeneChip expression measures. Bioinformatics, 22:789-794.

89. Shchepinov, M.S., Case-Green, S.C. and Southern, E.M. 1997. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. Nucleic Acids Res, 25:1155-1161.

90. Lima, W.F., Monia, B.P., Ecker, D.J. and Freier, S.M. 1992. Implication of RNA structure on antisense oligonucleotide hybridization kinetics. Biochemistry, 31:12055-12061.

91. Anthony, R.M., Schuitema, A.R., Chan, A.B., Boender, P.J., Klatser, P.R. and Oskam, L. 2003. Effect of secondary structure on single nucleotide polymorphism detection with a porous microarray matrix; implications for probe selection. Biotechniques, 34:1082-1089.

92. Markham, N.R. and Zuker, M. 2005. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res, 33:W577-581.

93. Koehler, R.T. and Peyret, N. 2005. Effects of DNA secondary structure on oligonucleotide probe binding efficiency. Comput Biol Chem, 29:393-397.

94. The human genome U133 Latin Square dataset [http://www.affymetrix.com/support/technical/sample_data/datasets.affx]

95. Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M. and Halfon, M.S. 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol, 6:R16.

96. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. and Korsmeyer, S.J. 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet, 30:41-47.

97. Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J. and Winzeler, E.A. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science, 301:1503-1508.

98. Fodor, A.A., Tickle, T.L. and Richardson, C. 2007. Towards the uniform distribution of null P values on Affymetrix microarrays. Genome Biol, 8:R69.

99. Pyott, S.J., Meredith, A.L., Fodor, A.A., Vazquez, A.E., Yamoah, E.N. and Aldrich, R.W. 2007. Cochlear function in mice lacking the BK channel alpha, beta1, or beta4 subunits. J Biol Chem, 282:3312-3324.

100. Meredith, A.L., Wiler, S.W., Miller, B.H., Takahashi, J.S., Fodor, A.A., Ruby, N.F. and Aldrich, R.W. 2006. BK calcium-activated potassium channels regulate circadian behavioral rhythms and pacemaker output. Nat Neurosci, 9:1041-1049.

101. Sugnet, C.W., Srinivasan, K., Clark, T.A., Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D. and Ares, M. 2006. Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. PLoS Comput Biol, 2:e4.

102. R Development Core Team. 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

103. Ding, Y., Chan, C.Y. and Lawrence, C.E. 2004. Sfold web server for statistical folding and rational design of nucleic acids. Nucl Acids Res, 32:W135-141.

104. Ratushna, V.G., Weller, J.W. and Gibas, C.J. 2005. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. BMC Genomics, 6:31.

105. Dalma-Weiszhausz, D.D., Warrington, J., Tanimoto, E.Y. and Miyada, C.G. 2006. The affymetrix GeneChip platform: an overview. Methods Enzymol, 410:3-28.

106. Allison, D.B., Cui, X., Page, G.P. and Sabripour, M. 2006. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet, 7:55-65.

107. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res, 31:e15.

108. Qin, L.X., Beyer, R.P., Hudson, F.N., Linford, N.J., Morris, D.E. and Kerr, K.F. 2006. Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. BMC Bioinformatics, 7:23.

109. Vardhanabhuti, S., Blakemore, S.J., Clark, S.M., Ghosh, S., Stephens, R.J. and Rajagopalan, D. 2006. A comparison of statistical tests for detecting differential expression using Affymetrix oligonucleotide microarrays. Omics, 10:555-566.

110. Wu, Z. and Irizarry, R.A. 2004. Preprocessing of oligonucleotide array data. Nat Biotechnol, 22:656-658.

111. Schuster, E., Blanc, E., Partridge, L. and Thornton, J. 2007. Estimation and correction of non-specific binding in a large-scale spike-in experiment. Genome Biology, 8:R126.

112. Irizarry, R.A., Cope, L.M. and Wu, Z. 2006. Feature-level exploration of a published Affymetrix GeneChip control dataset. Genome Biol, 7:404.

113. Dabney, A.R. and Storey, J.D. 2006. A reanalysis of a published Affymetrix GeneChip control dataset. Genome Biol, 7:401.

114. Gaile, D.P. and Miecznikowski, J.C. 2007. Putative null distributions corresponding to tests of differential expression in the Golden Spike dataset are intensity dependent. BMC genomics [electronic resource], 8:105.

115. Pearson, R.D. 2008. A comprehensive re-analysis of the Golden Spike data: towards a benchmark for differential expression methods. BMC Bioinformatics, 9:164.

116. NNFIT. [http://webpages.uncc.edu/~rgharaib/nnfit]

117. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y. and Zhang, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol, 5:R80.

118. Affycomp. [http://affycomp.biostat.jhsph.edu/AFFY2/TABLES.hgu/0.html#table133]

119. Cope, L., Irizarry, R., Jafee, H.W. and Speed, T.P. 2003. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics, 1:1-13.

120. Baldi, P. and Long, A.D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 17:509-519.

121. Bloomfield, V.A., Crothers, D.M. and Tinoco, I. 2000. Nucleic acids : structures, properties, and functions. Sausalito, Calif.: University Science Books.

122. Zhang, L., Wu, C., Carta, R., Baggerly, K. and Coombes, K.R. 2004. Response to Preprocessing of oligonucleotide array data. Nat Biotechnol, 22:658.

123. Relogio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R.B. and Valcarcel, J. 2005. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. J Biol Chem, 280:4779-4784.

124. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537.

125. Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., Wu, L.F., Altschuler, S.J., Edwards, S., King, J., Tsang, J.S., Schimmack, G., Schelter, J.M., Koch, J., Ziman, M., Marton, M.J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M.R., Mao, M., Burchard, J., Kidd, M.J., Dai, H., Phillips,

J.W., Linsley, P.S., Stoughton, R., Scherer, S. and Boguski, M.S. 2001. Experimental annotation of the human genome using microarray technology. Nature, 409:922-927.

126. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. 2000. Functional discovery via a compendium of expression profiles. Cell, 102:109-126.

127. Rabbee, N. and Speed, T.P. 2006. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics, 22:7-12.

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

129. Dorris, D.R., Nguyen, A., Gieser, L., Lockner, R., Lublinsky, A., Patterson, M., Touma, E., Sendera, T.J., Elgahanian, R. and Mazumder, A. 2003. Oligonucleotide probe accessibility on a three-dimensional DNA microarray surface and the effect of hybridization time on the accuracy of expression ratios. BMC Biotechnology, 3:1472-6750.

130. Held, G.A., Grinstein, G. and Tu, Y. 2006. Relationship between gene expression and observed intensities in DNA microarrays--a modeling study. Nucleic Acids Res, 34:e70.

131. Gong, P. and Levicky, R. 2008. DNA surface hybridization regimes. Proc Natl Acad Sci U S A, 105:5301-5306.

132. He, Z., Wu, L., Li, X., Fields, M.W. and Zhou, J. 2005. Empirical establishment of oligonucleotide probe design criteria. Appl Environ Microbiol, 71:3753-3760.

133. Dong, Y., Glasner, J.D., Blattner, F.R. and Triplett, E.W. 2001. Genomic interspecies microarray hybridization: rapid discovery of three thousand genes in the maize endophyte, Klebsiella pneumoniae 342, by microarray hybridization with Escherichia coli K-12 open reading frames. Appl Environ Microbiol, 67:1911-1921.

134. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S. and Fodor, S.P.A. 1996. Accessing Genetic Information with High-Density DNA Arrays. Science, 274:610-614.

135. Relogio, A., Schwager, C., Richter, A., Ansorge, W. and Valcarcel, J. 2002. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res, 30:e51.

136. Peterson, A.W., Wolf, L.K. and Georgiadis, R.M. 2002. Hybridization of mismatched or partially matched DNA at surfaces. J Am Chem Soc, 124:14601-14607.

137. Gibbs, A., Armstrong, J., Mackenzie, A.M. and Weiller, G.F. 1998. The GPRIME package: computer programs for identifying the best regions of aligned genes to target in nucleic acid hybridisation-based diagnostic tests, and their use with plant viruses. J Virol Methods, 74:67-76.

138. Gao, Y., Wolf, L.K. and Georgiadis, R.M. 2006. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. Nucleic Acids Res, 34:3370-3377.

139. Weckx, S., Carlon, E., DeVuyst, L. and Van Hummelen, P. 2007. Thermodynamic behavior of short oligonucleotides in microarray hybridizations can be described using Gibbs free energy in a nearest-neighbor model. J Phys Chem B, 111:13583-13590.

140. Fish, D.J., Horne, M.T., Brewood, G.P., Goodarzi, J.P., Alemayehu, S., Bhandiwad, A., Searles, R.P. and Benight, A.S. 2007. DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. Nucleic Acids Res, 35:7197-7208.

141. Kibbe, W.A. 2007. OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Res, 35:W43-46.

142. SantaLucia, J., Jr. 2007. Physical principles and visual-OMP software for optimal PCR design. Methods Mol Biol, 402:3-34.

143. Markham, N.R. and Zuker, M. 2008. UNAFold. In: Bioinformatics. 3-31.

144. Smyth, G. 2005. limma: Linear Models for Microarray Data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. 397-420.

145. R Development Core Team. 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

146. Burden, C.J., Pittelkow, Y.E. and Wilson, S.R. 2004. Statistical analysis of adsorption models for oligonucleotide microarrays. Stat Appl Genet Mol Biol, 3:Article35.

147. unccMM. [http://bioinfo.uncc.edu/rgharaib/unccMM]

148. 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. and Linsley, P.S. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol, 19:342-347.

149. Deng, Y., He, Z., Van Nostrand, J.D. and Zhou, J. 2008. Design and analysis of mismatch probes for long oligonucleotide microarrays. BMC Genomics, 9:491.

150. Mulders, G.C., Barkema, G.T. and Carlon, E. 2009. Inverse Langmuir method for oligonucleotide microarray analysis. BMC Bioinformatics, 10:64.

151. R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

152. McCall, M.N. and Irizarry, R.A. 2008. Consolidated strategy for the analysis of microarray spike-in data. Nucleic Acids Res, 36:e108.

153. Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., Foy, C., Fuscoe, J., Gao, X., Gerhold, D.L., Gilles, P., Goodsaid, F., Guo, X., Hackett, J., Hockett, R.D., Ikonomi, P., Irizarry, R.A., Kawasaki, E.S., Kaysser-Kranich, T., Kerr, K., Kiser, G., Koch, W.H., Lee, K.Y., Liu, C., Liu, Z.L., Lucas, A., Manohar, C.F., Miyada, G., Modrusan, Z., Parkes, H., Puri, R.K., Reid, L., Ryder, T.B., Salit, M., Samaha, R.R., Scherf, U., Sendera, T.J., Setterquist, R.A., Shi, L., Shippy, R., Soriano, J.V., Wagar, E.A., Warrington, J.A., Williams, M., Wilmer, F., Wilson, M., Wolber, P.K., Wu, X. and Zadro, R. 2005. The External RNA Controls Consortium: a progress report. Nat Methods, 2:731-734.

154. Gharaibeh, R.Z., Newton, J.M., Weller, J.W. and Gibas, C.J. 2009. Application of equilibrium models of solution hybridization to microarray design and analysis. in review.

155. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19:185-193.

156. Shi, L., Tong, W., Su, Z., Han, T., Han, J., Puri, R., Fang, H., Frueh, F., Goodsaid, F., Guo, L., Branham, W., Chen, J., Xu, Z.A., Harris, S., Hong, H., Xie, Q., Perkins, R. and Fuscoe, J. 2005. Microarray scanner calibration curves: characteristics and implications. BMC Bioinformatics, 6:S11.