ESTIMATING ABSOLUTE TRANSCRIPT CONCENTRATION FOR MICROARRAYS USING LANGMUIR ADSORPTION THEORY

by

MIN MO

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2008

ACKNOWLEDGEMENTS

I sincerely thank my advisor Dr. D. L Hawkins for his constant dedication, generosity over the past six years. Without his wisdom, knowledge, insight, and encouragement, this paper would not have been possible. I give my best wish to Dr. D. L Hawkins speedy recovery well!

My sincere gratitude goes to Dr. Chien-Pai Han, the chair of my committee after Dr. Hawkins was in the hospital, for his support and invaluable advice.

I would like to thank other members of my committee, Dr. Danny Dyer, Dr. Andrzej Korzeniowski and Dr. Pawel Michalak for their constructive comments and valuable suggestion for this research.

I thank my father Zucai Mo and my husband Dr. Shutian Deng for their support, patience and unconditional love. Last but not least, my lovely son Dawson, so full of live and joy, always gives me the most beautiful smile and gives me the encouragement I need. My studies in UTA would not have been finished without their understanding and effort.

April 16, 2008

ABSTRACT

ESTIMATING ABSOLUTE TRANSCRIPT CONCENTRATION FOR MICROARRAYS USING LANGMUIR ADSORPTION THEORY

MIN MO, PhD.

The University of Texas at Arlington, 2008

Supervising Professor: Doyle L. Hawkins

This paper estimates the Langmuir parameters for probe on microarray then improves estimation of absolute transcript concentration using Langmuir adsorption model. We use the spike-in probes found on commercial microarrays, along with Langmuir adsorption model to estimate Langmuir parameters for spike-in probes, then combine with an assumed log-linear model for those Langmuir parameters in terms of the spike-in probe sequence features, to estimate the assumed-invariant model coefficients. These estimated coefficients are then used, along with the probe sequence features of the target probes, to estimate the Langmuir parameters for each target probe. Finally, these estimated Langmuir parameters are combined with the expression measurements to produce estimates of the absolute transcript concentrations. The performance of this method, which amounts to extrapolation of a model fit over the space of the spike-in probe features to the space of the target probe features, will depend on the extent of this extrapolation. Simulation results will be presented to describe the performance of the method. The optimal choice of spike-in probes is given to the chip design.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Microarray technology has been widely used in determine thousands of genes expression pattern in few hours; examining mRNA from different tissues in normal and abnormal; determining which genes and environmental conditions can be lead to disease and identifying protein binding site, etc. DNA microarray, such as cDNA spotted microarrays [Duggan, et. al, 1999] and in-situ oligonucleotide arrays (e.g., Affymetrix chips) [Lipshutz, et. al, 1999], have orderly arrangements of nucleic acid spots at high density, provided high-throughput measurements in molecular biology, yielded information for the reconstruction of complex gene control networks [Lee, 2004]. Researcher can monitor expression level for thousands of genes simultaneously. In this chapter, the problem, biological background and theory are introduced.

1.1 Biological Background

1.1.1 DNA and Central Dogma

A cell is the minimal unit of life. Deoxyribonucleic acid (DNA) carried the information necessary for the functioning of cell. DNA is composed of four nucleotides, each nucleotide is made up of three elements: a phosphate group, a deoxyribose sugar and one of four different nitrogen bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Watson and Crick discovered the structure of DNA is a double helix, which is a chain of nucleotides, in 1953 [Watson, 1953 and 1997]. The pair principle is that G pairs only with C, and A pairs only with T. DNA can be copied and pass out nucleus, the genetic information can also be copied as ribonucleic acid (RNA) molecules, which is single-stranded and complementary to one of the two DNA strands. This process is called transcription.

RNA has a pyrimidine base uracil (U) instead of T, with U always pairing with A. There are two main classes of RNA: messenger RNA (mRNA) and functional RNA which including transfer

RNA (tRNA) and ribosomal RNA (rRNA). The RNA is transferred to machinery that synthesizes protein molecules based on the information carried by the RNA, this process is called translation.

Gene is the segment of the DNA sequence that controls the identifiable hereditary traits of an organism. The central dogma of molecular biology states that DNA is transcribed into mRNA molecule in nucleus, which is then translated into a protein during synthesis (Figure 1.1 [Primer on Molecular Genetics, 1992]). The process of reading the mRNA sequence and converting it into an amino acid sequence is called translation. The A, G, C and T is translated into 20-amino-acid alphabet of proteins in ribosome which is big complex of several proteins and ribosomal RNA. A gene determines when, what amount and what kind of protein will be generated in the cell. The protein and its interaction with the environment then determine the phenotypes of the cells and the organism.

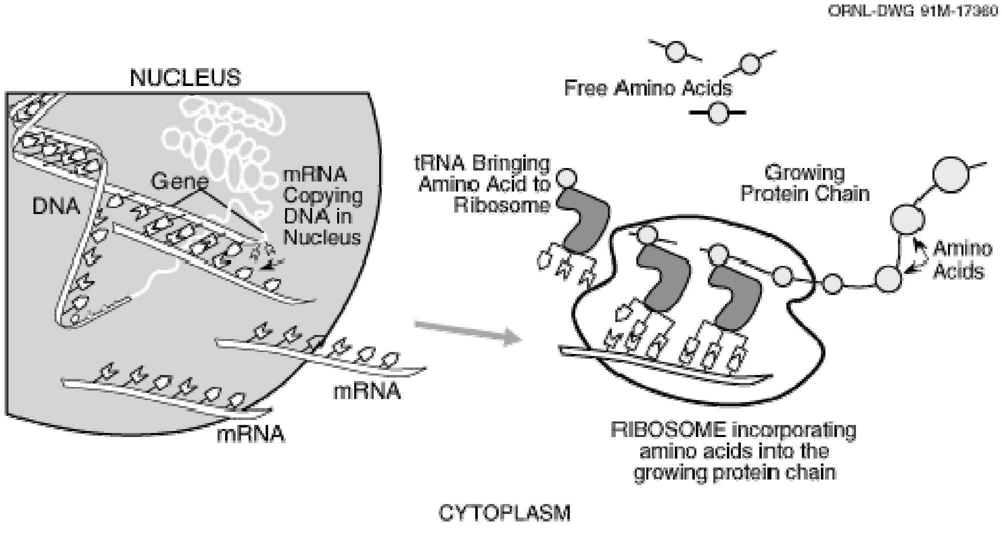
ORNL-DWG 91M-17380
NUCLEUS
DNA
Gene
mRNA Copying DNA in Nucleus
mRNA
mRNA
Free Amino Acids
tRNA Bringing Amino Acid to Ribosome
Growing Protein Chain
Amino Acids
RIBOSOME incorporating amino acids into the growing protein chain
CYTOPLASM

Figure 1.1 Gene Expression. DNA is transcribed into mRNA molecule in nucleus, translation is processed in ribosome.

1.1.2 Measuring Gene Expression

To understand the function of a gene, it is necessary to know which protein it encodes, the condition which leads to its activation and the level of activity which these conditions induce. Gene expression is the process by which mRNA and protein are synthesized from the DNA template of each gene. The first stage of this process is called transcription, when one strand of DNA is copied as RNA; the second stage of gene expression is translation of mRNA into protein. The gene expression can be measured at two levels: mRNA (what is transcribed) and protein level (how much is made). Respite recent advances in the field of proteomics [Lewin, 1997], it is difficult to measure a gene expression at protein level, an alternative definition of gene expression can be obtained by mRNA level. Under the assumption that the presence of mRNA indicate gene expression and be used to control the protein for which gene encodes, a gene is referred to as expressed if its DNA has been transcribed to RNA, then measure of gene expression is the abundance of mRNA (mRNA concentration). The DNA microarray measures gene expression at mRNA level.

1.1.3 Microarray Technology

Microarray offers an efficient method of gathering data that can be used to determine the expression patterns of tens of thousands of genes in only a few hours. Microarray methods allow researchers to examine the mRNA from different tissues in normal and disease and determine which genes and environmental conditions can be lead to disease. Similarly, microarray can be used to determine which genes are expressed in which tissues and at which time during embryonic development.

The first complementary DNA microarray was invented in 1995 at Stanford University, it contained only 48 cDNAs, but today, there are tens of thousands of genes and even whole genomes on an array.

The basic concept behind all microarrays is the precise positioning of DNA fragments at high density on a solid support, and the natural affinity of single stranded DNA to bind with its complementary sequence. There are two main microarray technologies: spotted microarray (cDNA spotted microarray and oligonucleotide spotted microarray) and in-situ oligonucleotide microarray (e.g. Affermetrix) [Lee, 2004]. We only discuss oligonucleotide microarray in this dissertation.

1.1.4 Oligonucleotide Microarray

1.1.4.1 Construction of the Microarrays

In oligonucleotide arrays, each target gene is represented by a probe set containing 14 carefully selected perfect match probes (PM) and 14 mismatch probes. Each PM probe is a 25-mer long (base sites) segment of the target gene. The set of PM probes is chosen to uniquely identify the target gene. Each mismatch probe is same as one corresponding PM probe, except the middle base (13th base) (Figure 1.2). The purpose of the MM probe design is to measure non-specific binding (mRNA transcript not hybridizing to its complementary counterpart) and background noise (unexpected noise, e.g. optical noise).

AATCCCAGTCTT**C**CTGAGGATACGC Perfect Match probe

AATCCCAGTCTT**G**CTGAGGATACGC MisMatch probe

Figure 1.2 Perfect Match and Mismatch Probes Construction

Affymetrix Genechip arrays, the focus of this dissertation, consist of a substrate onto which short single strand DNA oligonucleotide probes have been synthesized using a photolithographic process. A chip surface is divided into hundreds of thousands of regions typically tens of microns in size (Figure 1.3 [Affymetrix.com]), each region for one probe.

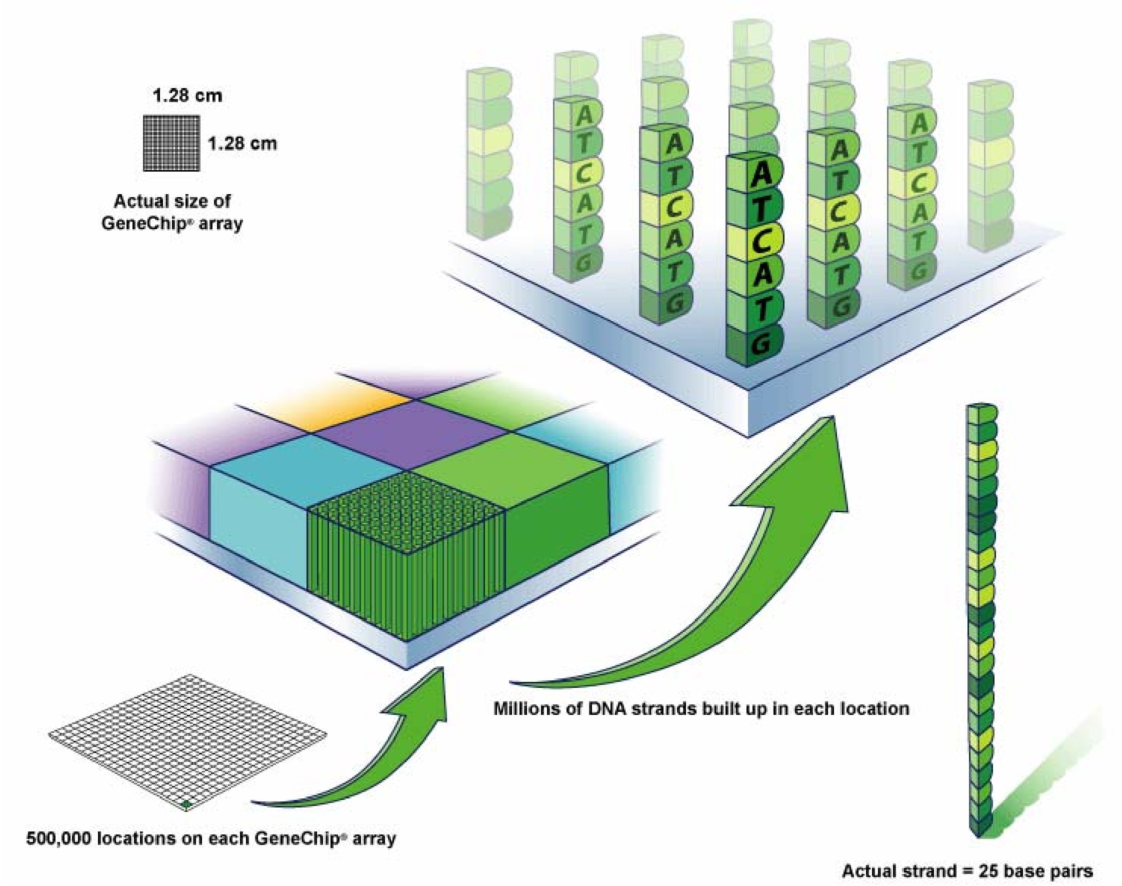
1.28 cm
1.28 cm
Actual size of GeneChip® array
A
T
C
A
T
G
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

Figure 1.3 Microarray Surface. A chip surface is divided into hundreds of thousands of regions typically tens of microns in size.

1.1.4.2 Target versus Spike-in Probes

The probe sets on an array are of two types:

(1) So called ‘target’ probe sets, which are designed to detect the presence of the mRNA of the target genes in the study sample.

(2) So called ‘spike-in’ probe sets, which are designed to detect the presence of ‘spike-in’ mRNA in the study sample.

Spike-in mRNA is artificial (to the study organism) mRNA which has been mixed, at known concentration, into the study RNA sample for the purpose of monitoring the validity of the array expression measures.

1.1.4.3 How does a Microarray Work

The target mRNA is collected from the study organism under the desired experimental condition, and mixed with the spike-in mRNA to form the study sample. The individual stands of mRNA are called transcripts. This mixture is labeled with fluorescent dye. Using a complex process, the study sample is hybridized onto the array. If mRNA transcript in the study sample finds its complementary counterpart among the probes on the array, it will hybridize (stick) to that probe. If it does not find its counterpart, then hopefully it does not stick to any probe (Figure 1.4 [Affymetrix.com]). After hybridization, the array is exposed to a laser light, which causes the dye to fluoresce. The fluorescence intensity is obtained by using a laser scanner. The more mRNA is stuck on the probe, the higher is a probe's fluorescence intensity (Figure 1.5 [Affymetrix.com]).

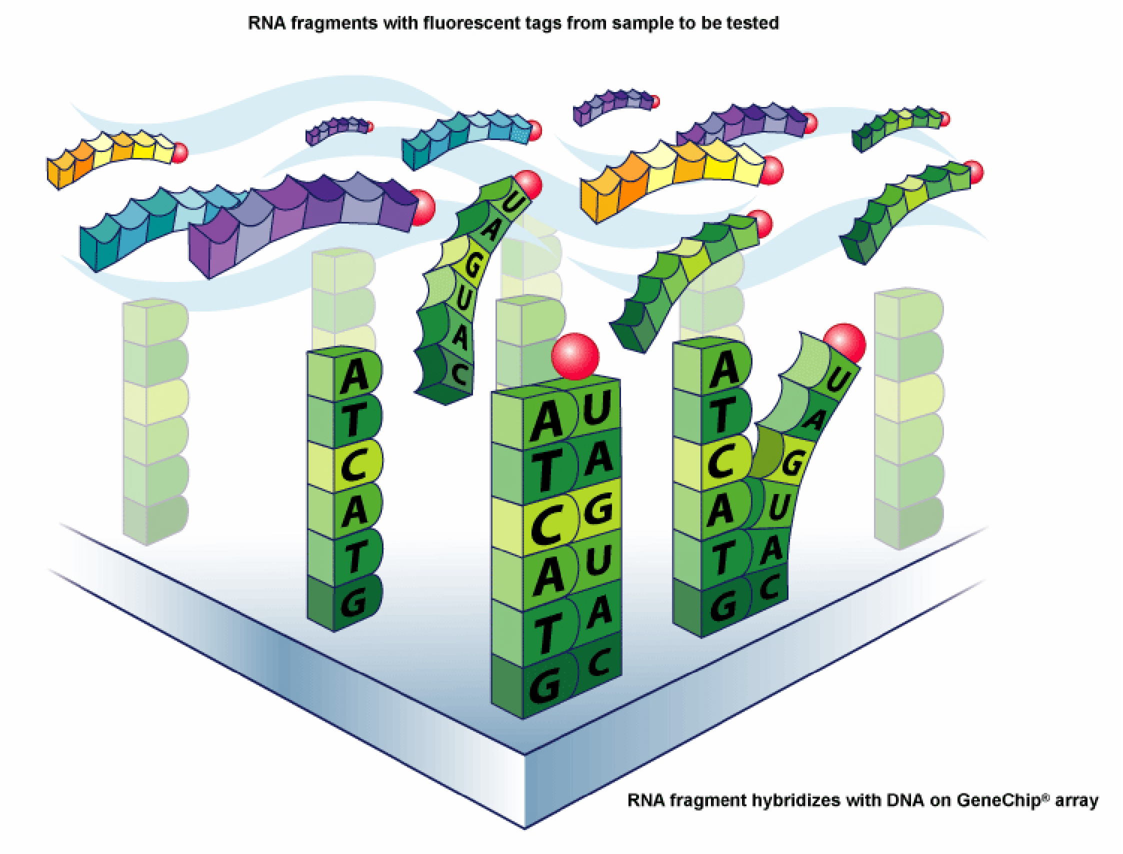
RNA fragments with fluorescent tags from sample to be tested
U
A
G
U
A
C
A
T
C
A
T
G
A U
T A
C G
A U
T A
G C
A
T
C
A
T
G
U
A
G
U
A
C
RNA fragment hybridizes with DNA on GeneChip® array

Figure 1.4 RNA Fragment Hybridizes with DNA on GeneChip Array

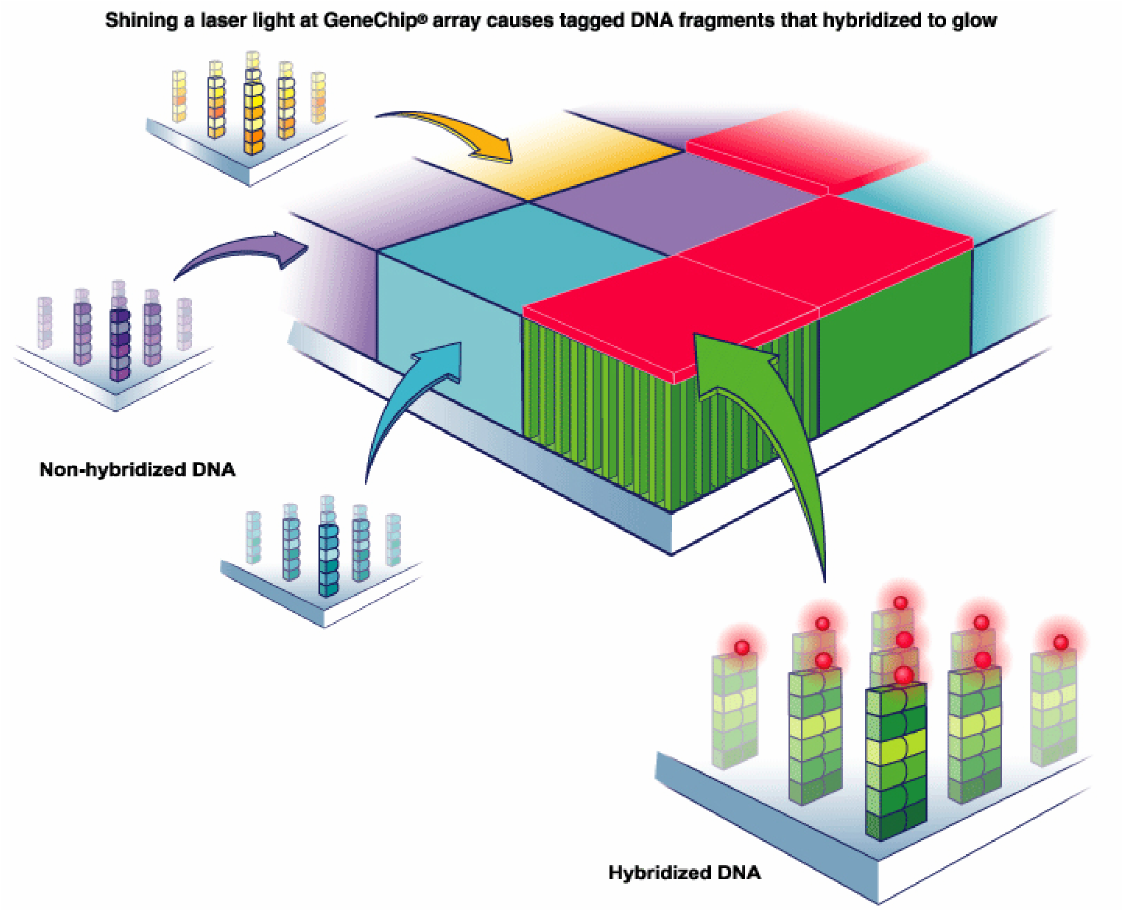
Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

Figure 1.5 Shinning a Laser Light at GeneChip Array Causes Tagged DNA Fragments that Hybridized to Glow

1.2 Absolute Concentration VS Fluorescent Intensity

While gene expression is defined in term of absolute concentration of mRNA (i.e. number of corresponding mRNA transcripts per unit volume), absolute mRNA concentration cannot, at present, be obtained directly. Thus, technological barriers limit researchers to using fluorescence intensity as an indirect measure of gene expression (See e.g. Li et al, 2001, Gauiter et al, 2004, Wu et al, 2004, Iriazrry et al, 2003 and Zhang et al, 2003). Early on, the justification for this indirect measure was the belief (See e. g. www.Affymetri.com) that absolute mRNA concentration is roughly linearly related to probe fluorescence intensity, so that measuring the latter suffices for the former.

However, it was eventually realized (Hekstra et al. [2003], Abdueva et al [2006], Burden et al. [2004] and Zhang et al [2006]) that this assumed linearity does not hold. Specifically, at high levels of absolute concentration, the fluorescence intensity tends to reach an upper limit and becomes insensitive to further increase in absolute concentration.

Recent technical advances hold promise for direct measurement of absolute concentration, but at present are not practical. Hence, recent research, including the present, has attempted estimation of absolute concentration from fluorescence intensity.

1.3 Attempting to Determine Absolute Concentration from Fluorescent Intensity: A Literature review

1.3.1 Hekstra's Discovery

Heskstra [2003] demonstrated, using spike-in experiments that the relationship between fluorescence intensity and absolute mRNA concentration is not linear. Further, he demonstrated that Genechip fluorescence intensity data follows Langmuir adsorption isotherms.

1.3.1.1 The Langmuir Adsorption Model in General

The Langmuir adsorption isotherm is a theory of physical chemistry, described by Atkins as the most elementary model of surface adsorption [Atkins, 1994]. The theory was developed by Irving Langmuir in 1916 to describe the dependence of the surface coverage of an adsorbed gas on the pressure of the gas above the surface at a fixed temperature. It is assumed that gas molecules striking the surface have a given probability of adsorbing. Molecules already adsorbed similarly have a given probability of desorbing. At equilibrium, equal numbers of molecules desorbs and adsorb at any time. The probabilities are related to the strength of the interaction between the adsorbent surface and the adsorb gas.

The Langmuir model is usually expressed as:

$$\frac{V}{V_m} = \frac{Cx}{1+Cx}$$

where V= volume of gas adsorbed at pressure *P*; V_m is volume of gas which could cover the entire adsorbing surface with a monomolecular layer; V_0 is saturation pressure of the gas, *i.e.,* the pressure of the gas in an equilibrium with bulk liquid at the temperature of the measurement; $x = P / P_0$ is relative pressure ($0 \le x \le 1$); C is constant for the gas/solid combination.

1.3.1.2 The Langmuir Adsorption Model Applied to Microarray: Hekstra's First Idea

Since microarray measurement involve adherence of particles (mRNA) to substrates (probes), the Langmuir adsorption model can be applied to microarray data analysis.

Assuming the measured fluorescence intensity of a probe is proportional to the number of mRNA transcripts stuck to the probe surface, the Langmuir model for the fluorescence intensity, I, in terms of the absolute concentration x, is:

$$I = a \frac{x}{x+b} + d \, , \qquad [1.1]$$

where a, b and d are probe specific parameters. Specifically, a is proportionality constant, b is the concentration at which the complementary RNA saturates half of the probe surface if there is no non-specific hybridization, and d presents the contribution from non-specific hybridization (i.e. material stuck to the probe which the probe is not intended to hybridize) .

Hekstra used probe-level fluorescence intensity measures from spike-in experiments (i.e. in which the absolute mRNA concentration were known) to estimate a, b and d for probe p , by weighted least-squares fits of [1.1]. I.e. they minimized the sum of weighted square errors:

$$S_p = \sum_{i=1}^{n} \frac{1}{I_{ip}} [I_{ip} - (\frac{a_p x_{ip}}{b_p + x_{ip}} + d_p)]^2 \, , \qquad [1.2]$$

where i indexes arrays, I_{ip} is the fluorescence intensity measurement for probe p on array i ,and x_{ip} is the known absolute concentration corresponding to probe p on array i .They then produced the plots of Y_{ip} versus X_{ip} in Figure 1.6 (Heskstra [2003]], where $X_{ip} = \frac{x_{ip}}{\hat{b}_p}$ and $Y_{ip} = \frac{I_{ip} - \hat{d}_p}{\hat{a}_p}$. The clear adherence of these plots to the functional form $Y = \frac{X}{1+X}$ (which is equivalent to [1.1]) shows the conformity of the fluorescence intensity and absolute concentration relationship to the Langmuir model.

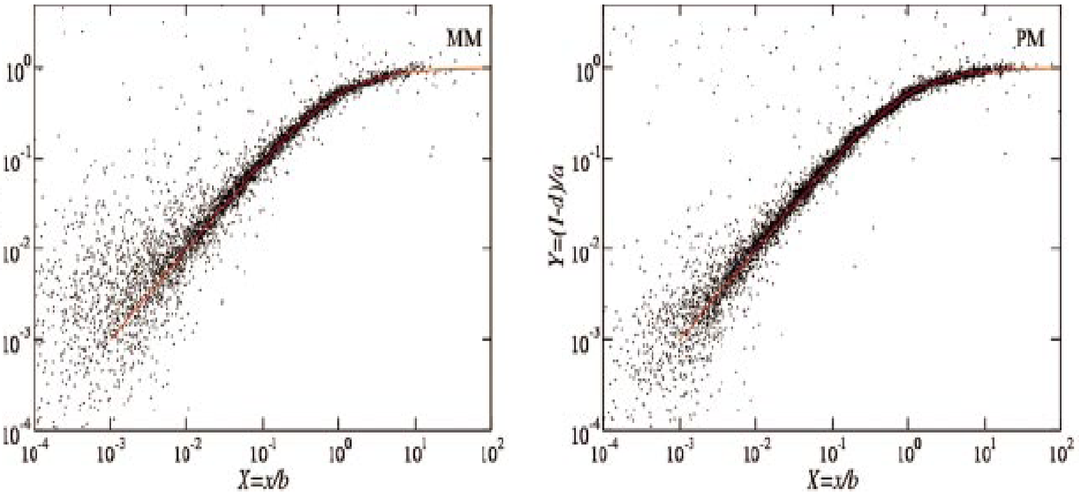

Figure 1.6. Langmuir Isotherm Provide Accurate Description of GeneChip hybridization

1.3.1.3 Hekstra's Second Idea: the Probe Parameters Depend on the Probe Structure

Researchers are interested in absolute concentration, but only obtain fluorescence intensity from the array. By Hekstra's Langmuir model [1.1], if a, b and d could be estimated, then one could estimate absolute concentration from fluorescence intensity. Since the probe structures are known, Hekstra proposed a statistical method for estimating the probe parameters in term of probe features. Specifically, he proposed the linear model [1.3].

$$\begin{pmatrix} \ln \hat{a}_p \\ \ln \hat{b}_p \\ \ln \hat{d}_p \end{pmatrix} = \begin{pmatrix} \gamma_A^a & \gamma_C^a & \gamma_G^a \\ \gamma_A^b & \gamma_C^b & \gamma_G^b \\ \gamma_A^d & \gamma_C^d & \gamma_G^d \end{pmatrix} * \begin{bmatrix} n_{A,p} \\ n_{C,p} \\ n_{G,p} \end{bmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \qquad [1.3]$$

where $\hat{a}_p, \hat{b}_p$ and $\hat{d}_p$ are probe parameter estimates for probe p , $n_{A,p}, n_{C,p}$ and $n_{G,p}$ are the number of A, C and G bases in probe p , the $\gamma's$ and $\beta's$ are unknown constants assumed to be the same for all probes, and $\varepsilon's$ are error terms. Hekstra obtained R^2 about 50% for each of the three components, in model [1.3], suggesting some merit for the idea.

It is important to note that in his development of model [1.3], Hekstra used probe parameter estimates obtained via least square, using intensity data from spike-in experiments (i.e. with known absolute concentration). He did not, however, show how to apply his ideas to the practical setting in which there is no spike-in data, so that [1.2] cannot be used to determine $\hat{a}, \hat{b}$ and $\hat{d}$. We remark that, using only the fluorescence intensity data, the least squares criterion cannot simultaneously identify all of $\hat{a}, \hat{b}$, $\hat{d}$ and x_{pi} in Equation [1.3].

But see section 1.3.3 below about the methods of Abdueva et al. [Abdueva, 2006]. Hekstra also proposed (assuming $\hat{a}, \hat{b}$ and $\hat{d}$ could somehow be obtained) estimation of absolute mRNA concentration of the target gene from probe p , via

$$\hat{x}_p = \hat{b}_p \frac{I_p - \hat{d}_p}{\hat{a}_p + \hat{d}_p - I_p} \qquad [1.4]$$

Since the absolute mRNA concentration is by definition, non-negative, there are two necessary constraints: $I > \hat{d}$ and $\hat{a} + \hat{d} > I$. Hence he proposed excluding any probes with $I < \hat{d}$ or $I > \hat{a} + \hat{d}$. He proposed averaging the by-probe estimates over all probes p in the probe set to obtain a single estimate of the target mRNA concentration.

1.3.2 Other Proposals for Estimating Absolute Concentration from Fluorescence Intensity

Recently studies have begun to address these issues by appealing to models based on principles of physical chemistry, such as Langmuir adsorption model, offer the possibility of predicting absolute concentration.

There are some methods which estimate concentration by using Langmuir adsorption:

(1) Held et al. [Held, 2003] demonstrate a correlation between hybridization intensity and calculated free energy of hybridization. Then combine hybridization rate quation, calculated free energy of hybridization, and base on Langmuir adsorption model to compute absolute transcript concentration for target gene.

(2) Burden et al. [Burden, 2004] develop several dynamic adsorption models, which based on the Langmuir adsorption model, relating fluorescent intensity to target RNA concentration, using an appropriately defined median over probes within probe set rather than the mean to improve estimators of absolute concentration by reducing bias, and enable to estimate confidence interval. They also mention the challenging problem of establishing an algorithm for extracting Langmuir parameters from a given probe sequence, a problem which Hekstra proposed to solve statistically ala model [1.3].

(3) Binder et al. [Binder, 2006] predicted the parameters of the Langmuir adsorption model in a sequence-specific fashion using a sum of positional-dependent and based-specific nearest-neighbor free energy terms, they used both PM and MM probes information, estimate absolute mRNA concentration from the PM-MM probe intensities difference using the Langmuir model.

(4) Abdueva et al. aimed at absolute concentration by using the Langmuir model, they fitted Langmuir parameters within a single global fitting routine instead of estimating the background before obtaining gene expression measure, and described a logarithm in linear model of Langmuir parameters to estimate concentration [Abdueva, 2006]. Abdueva used Hesktra's first idea, gave an initial estimation of concentration, plus

$$\log(PM_{pjl}) = concentration_j + probe\,affinity_p + \varepsilon_{pjl} \qquad [1.5]$$

Where p is probe index, j a condition index, l a replicate, PM is the fluorescence intensity of perfect match probe, into

$$I = a\frac{c}{c+b} + d \qquad [1.2]$$

To estimate $\hat{a}, \hat{b}$, $\hat{d}$, then use [1.2] again to estimate initial concentration.The result depends on the starting value they chose, so it would not be possible if the starting value is randomly chosen. They can not estimate all $a, b, d\ and\ c$ simultaneously, since all parameters are not identifiable.

We remark that none of those papers considers using spike-in probes, whether already installed on arrays or perhaps to be designed estimate the absolute mRNA concentration. There are the ideas we study in this dissertation.

1.4 Motivation of Our Method

The motivation of this dissertation is, in brief, that Hekstra's ideas to develop a practical method for estimating absolute concentration from fluorescence intensity. Assuming that spike-in probes-- either already installed on the arrays or perhaps specially designed - -are available on the arrays and that the corresponding spike-in material is mixed into the target sample at known concentration, then one should to be able to:

(1) Estimate $\hat{a}, \hat{b}$ and $\hat{d}$ for each spike-in probe in model [1.2] using known florescence intensity, concentration and the Langmuir model

$$I = (\frac{a * c}{b + c} + d) * \varepsilon \qquad [1.6]$$

(2) Estimate universal $\gamma's$ and $\beta's$ in model [1.3] from such $\hat{a}, \hat{b}$, $\hat{d}$.

(3) Estimate $\hat{a}, \hat{b}$ and $\hat{d}$ for each target probe by using model [1.3] and universal $\gamma's$ and $\beta's$.

(4) Estimate absolute concentration for each target probe using $\hat{a}, \hat{b}$ and $\hat{d}$ and model [1.6].

Abdueva's method does not use any spike-in probe information, but her result depends on a carefully chosen initial absolute concentration value. By comparing with Abdueva's method, we use spike-in probe information, which is on array, to estimate absolute concentration of target gene without depending on any other starting value selecting.

CHAPTER 2

OUR PROPOSED METHOD FOR ESTIMATING ABSOLUTE CONCENTRATION WHEN SPIKE-IN PROBES ARE GIVEN

In this chapter we assume that we have available identical arrays with spike-in probes already installed. We make no particular assumption about the spike-in probe, except that it is possible to mix the spike-in material into the target samples at known concentrations, which vary across the arrays for a give experimental condition. In chapter 3, we take up the matter of optional design of the spike-in probes, should this be possible.

2.1 Our Assumption

2.1.1 Practical Assumptions

(1) Given spike-in probes already on the arrays, with corresponding spike-in material included in the target samples;

(2) The spike-in probe sequence and concentrations are known;

(3) For each experimental condition, we have multiple arrays with varying spike-in concentrations across the arrays.

2.1.2 Theoretical Assumptions

(1) Hekstra's model [1.3] holds for each probe.

(2) Hekstra's empirical model [1.6] holds with normal error.

(3) $\gamma's$ & $\beta's$ in [1.3] are the same for all probes.

2.2 Proposed Method

For the spike-in probes, since the corresponding absolute concentrations are known, we can estimate the Langmuir parameters a, b and d for each spike-in probe by extending model [1] to the statistical model:

$$I_{S,p,i} = (\hat{a}_{S,p} \frac{c_{s,i}}{c_{s,i} + \hat{b}_{s,p}} + \hat{d}_{s,p}) * \varepsilon_{s,p,i} \qquad [2.1]$$

where s indicates spike-in probe, $p = 1,2....28$. indexes probes, $i = 1,2,...N$ indexes arrays for the same experimental condition. $\log \varepsilon_{S,p,i}$ is assumed $N(0,\sigma^2)$; $I_{s,p,i}$ is the florescence intensity measure for spike-in probe p on array i . $c_{s,i}$ denotes the known absolute concentration of spike-in transcripts corresponding to probe p on array i , which is assumed to vary across the arrays i . $a_{s,p}, b_{s,p}$ and $d_{s,p}$ are the unknown Langmuir parameters of spike-in probe p .

1. For each spike-in probe, p , we can obtain $\{\hat{a}_{s,p}, \hat{b}_{s,p}, \hat{d}_{s,p}, s \in SI, p = 1,2,...28\}$ from model [1.4] by using nonlinear regression, minimizing:

$$S_{s,p} = \sum_{i=1}^{N} [\log I_{s,p,i} - \log(\frac{\hat{a}_{s,p} c_{s,i}}{\hat{b}_{s,p} + c_{s,i}} + \hat{d}_{s,p})]^2 \qquad [2.2]$$

with respect to $\hat{a}_{s,p}, \hat{b}_{s,p}, \hat{d}_{s,p}$. Let SI denotes the set of spike-in probes, T denotes the set of probes corresponding to a particular target gene. Since there are 3 parameters ($a, b\, and\, d$), so $N \geq 3$ is required.

2. Then use $\{\hat{a}_{s,p}, \hat{b}_{s,p}, \hat{d}_{s,p}, s \in SI, p = 1,2,...28\}$, to obtain the assumed universal $\gamma' s\ \&\ \beta' s(\underline{\beta})$ by applying model [1.3].

$$\left(\ln\hat{a}_{s.p} \quad \ln\hat{b}_{s,p} \quad \ln\hat{d}_{s,p}\right) = \underline{X} * \underline{\beta} + \underline{\varepsilon}$$

$$= (n_{s,p,A} \quad n_{s,p,C} \quad n_{s,p,G} \quad 1) * \begin{pmatrix} \gamma_A^a & \gamma_A^b & \gamma_A^d \\ \gamma_C^a & \gamma_C^b & \gamma_C^d \\ \gamma_G^a & \gamma_G^b & \gamma_G^d \\ \beta_1 & \beta_2 & \beta_3 \end{pmatrix} + \left(\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3\right) \tag{2.3}$$

In model [2.2], $\hat{a}_{g,p}, \hat{b}_{g,p}, \hat{d}_{g,p}$ are known, $n_{s,p,A}, n_{s,p,C}$ and $n_{s,p,G}$ are the known nucleotide counts for spike-in probe p . We use OLS component wise in [2.3] to estimate $\underline{\beta}$, one column at a time.

3. After obtaining the estimates of $\underline{\beta}$, then use model [2.3] and the known nucleotide counts for the target probes to estimate $\hat{a}_{T,p}, \hat{b}_{T,p}, \hat{d}_{T,p}$ for target probes p .

$$\left(\ln\hat{a}_{T,p} \quad \ln\hat{b}_{T,p} \quad \ln\hat{d}_{T,p}\right) = \underline{X} * \underline{\hat{\beta}} = (n_{T,p,A} \quad n_{T,p,C} \quad n_{T,p,G} \quad 1) * \begin{pmatrix} \hat{\gamma}_A^a & \hat{\gamma}_A^b & \hat{\gamma}_A^d \\ \hat{\gamma}_C^a & \hat{\gamma}_C^b & \hat{\gamma}_C^d \\ \hat{\gamma}_G^a & \hat{\gamma}_G^b & \hat{\gamma}_G^d \\ \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 \end{pmatrix} \tag{2.4}$$

here $n_{T,p,A}, n_{T,p,C}$ & $n_{T,p,G}$ are the known nucleotide counts for target probe p , so we predict $\hat{a}_{T,p}, \hat{b}_{T,p}, \hat{d}_{T,p}$ for target genes.

4. Finally, we plug $\hat{a}_{T,p}, \hat{b}_{T,p}, \hat{d}_{T,p}$ of target probes into:

$$I_{T,p,i} = (\hat{a}_{T,p} \frac{c_T}{c_T + \hat{b}_{T,p}} + \hat{d}_{T,p}) * \varepsilon_{T,p,i} \tag{2.5}$$

to estimate $\hat{C}_T$ (concentration of target gene), by minimizing the sum of square error with respect to $\hat{C}_T$:

$$S_T = \sum_{p=1}^{28} \sum_{i=1}^{N} [\log I_{T,p,i} - \log(\frac{\hat{a}_{T,p} c_T}{\hat{b}_{T,p} + c_T} + \hat{d}_{T,p})]^2 \tag{2.6}$$

estimate one gene at a time.

2.3 Simulation

In order to check whether our method works under the stated assumption and its sensitivity to spike-in/target probes spacing, noise, etc. we did a simulation study.

2.3.1 Simulation Process

We use SAS program to simulate and analyze the data.

(1) Data set

We simulate 100 replicates, in each hypothetical experimental condition, there are:

a) Spike-in probes and target genes

There are 3*28 spike-in probes (3*14 PM and 3*14 MM probes) and 10 target genes (10*14 PM and 10*14 MM probes), the difference between perfect match probe and Mismatch probe is the 13th bite. Each probe is randomly selected, and the numbers of nucleotides on the probe are independent identically distributed as Multivariate Normal distribution.

Let n_{gp}^{A} , n_{gp}^{T} , n_{gp}^{C} and n_{gp}^{G} are the number of A, T, C and G on the gene g , probe p ,

$\underline{X}_{gp} = [n_{gp}^{A} \quad n_{gp}^{T} \quad n_{gp}^{C} \quad n_{gp}^{G}]$, $\underline{\pi}_{gp} = \lfloor \pi_{gp1} \quad \pi_{gp2} \quad \pi_{gp3} \quad \pi_{gp4} \rfloor$,

we assume:

$$\underline{X}_{gp} \sim MN(25 * \underline{\pi}_{gp}, \Sigma_{gp}^{2}) .$$

The following logit model has been use to generate the probabilities:

$$\log(\frac{\pi_{gpk}}{\pi_{gp4}}) = \delta_k + \delta_g + \delta_p \qquad [2.7]$$

Where $k = 1,2,3$, g is gene index and p is the probe index, π_{gp1} indicates the probabilities of nucleotide A on the gene g probe p , π_{gp2} indicates the probabilities of nucleotide T on the gene g probe p , π_{gp3} indicates the probabilities of nucleotide C on the gene g probe p.

Let $T_{gpk} = \delta_k + \delta_g + \delta_p$, $\frac{\pi_{gpk}}{\pi_{gp4}} = e^{T_{gpk}}$, then

$$\pi_{gp1} = \frac{T_{gp1}}{1 + T_{gp1} + T_{gp2} + T_{gp3}}$$

$$\pi_{gp2} = \frac{T_{gp2}}{1 + T_{gp1} + T_{gp2} + T_{gp3}}$$

$$\pi_{gp3} = \frac{T_{gp3}}{1 + T_{gp1} + T_{gp2} + T_{gp3}}$$

$$\pi_{gp4} = \frac{1}{1 + T_{gp1} + T_{gp2} + T_{gp3}}$$

They satisfy

$$\sum_{k=1}^{4} \pi_{gpk} = 1.$$

The separation of spike-in probes and target genes depend on δ_g and δ_p, the more difference on δ_g and δ_p, the father away between spike-in and target genes.

Examples 1: The probabilities $\pi_{gp1}, \pi_{gp2}, \pi_{gp3}\ and\ \pi_{gp4}$ spike-in probes and target probes are equal. δ_k, δ_g and δ_p are same for spike-in and target genes.

Table 2.1 Separation between Spike-in and target probes with same δ_g

Probability	Spike-in	Target
π_{gp1}	0.1	0.1
π_{gp2}	0.2	0.2
π_{gp3}	0.4	0.4
π_{gp4}	0.3	0.3

We control spike-in and target probes separation by changing δ_g .

Example 2: The probabilities $\pi_{gp1}, \pi_{gp2}, \pi_{gp3}\ and\ \pi_{gp4}$ spike-in probes and target probes are separate, $\delta_k = 0.1, \delta_p = 0.2$ for both spike-in and target probes.

Table 2.2 Separation between Spike-in and target probes with different δ_g

	Spike-in	Target
Probability	$\delta_g = 0.1$	$\delta_g = 1$
π_{gp1}	0.24124	0.27327
π_{gp2}	0.26661	0.30201
π_{gp3}	0.29465	0.33377
π_{gp4}	0.19751	0.090962

b) Arrays

5 arrays was generated in each replicate, the transcript concentration on spike-in probes were set vary across the arrays while the concentration of target genes depend on gene, the transcript concentration belong to (2, 4, …, 1024) PM.

c) Universal $\underline{\beta}^{'}s$

Under our assumption, $\underline{\beta}^{'}s$ in model [2.3] are the same for all probes, so we assign value to $\underline{\beta}^{'}s$. Then compute the Langmuir parameters ($\hat{a}, \hat{b}$ and $\hat{d}$) by using model [1.3] for each probes with the noise.

(2) Varied factors

a) The separation between spike-in probes and target probes ($\delta_k, \delta_g\ and\ \delta_p$ in [2.7]).

b) The fluorescence intensity is computed by using model [1.4]

$$\log(I_{g,p,i}) = \log(\hat{a}_{g,p} \frac{c_{g,i}}{c_{g,i} + \hat{b}_{g,p}} + \hat{d}_{g,p}) + \log(\varepsilon_{g,p,i}) \qquad [2.8]$$

The noise $\log \varepsilon_{g,p,i}$ changes across array, gene and probe.

(3) Program

We use SAS software for the whole simulation, PROC IML is used to generate the data set, and PROC NLIN is used for the non-linear regression model.

2.3.2 Estimate Absolute mRNA Concentration

By using our proposed method, vary standard deviation of the noise ($\varepsilon_{s,p,i}$) in [2.8] and separation between spike-in probes and target probes, which δ_g is different between spike-in probes and target probes, we estimate the absolute mRNA concentration of target genes, the result is very good in term of relative bias, average of square standard error and variance of estimate absolute mRNA concentration in each scenarios. Where

$bias = \frac{1}{R}\sum_{r=1}^{R} \hat{c}_r - c_{true}$, r indicate the number of replication.

$relative\,bias = \frac{bias}{c_{true}}$;

$\text{var}(\hat{c})$ is the unconditional variance of estimate absolute concentration of target gene;

$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$ is variance of estimate absolute concentration of target gene under given target probe information;

δ_{sg} is δ_g for spike-in probes, δ_{Tg} is δ_g for target probes.

Table 2.3: Scenario 1-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = \delta_{Tg} = 2$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.2039163	0.3362011	0.1236815	0.3362011

Table 2.4: Scenario 1-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = \delta_{Tg} = 2$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.05	0.02	1.03	0.10
2	4	4.07	0.02	2.05	0.21
3	8	8.04	0.01	3.58	0.61
4	16	15.85	-0.01	9.43	2.24
5	32	31.74	-0.01	57.65	8.23
6	64	62.39	-0.03	309.60	27.26
7	128	131.99	0.03	1608.52	160.77
8	256	261.97	0.02	6525.38	967.70
9	512	516.52	0.01	38161.93	5300.00
10	1024	1126.91	0.10	866320.93	120532.91

Table 2.5: Scenario 2-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.28749	0.4739908	0.1743715	0.0641477

Table 2.6: Scenario 2-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.04	0.02	1.03	0.10
2	4	4.08	0.02	2.01	0.29
3	8	8.02	0.01	2.65	0.70
4	16	15.82	-0.01	5.15	2.18
5	32	31.72	-0.01	19.22	7.27
6	64	62.64	-0.02	101.76	23.17
7	128	130.86	0.02	875.55	96.46
8	256	261.83	0.02	3645.89	567.67
9	512	512.16	0.01	10650.56	2087.48
10	1024	1057.09	0.03	84910.60	25927.06

Table 2.7: Scenario 3-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.2996401	0.494023	0.1817409	0.024596

Table 2.8: Scenario 3-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

Target gene	True C	$\hat{C}$	Relative bias	$\mathrm{var}(\hat{c})$	$\mathrm{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.05	0.02	4.39	0.07
2	4	4.08	0.02	9.68	0.31
3	8	8.01	0.01	12.37	0.71
4	16	15.82	-0.01	26.70	2.20
5	32	31.73	-0.01	86.09	7.28
6	64	62.69	-0.02	402.12	23.19
7	128	130.66	0.02	3021.57	90.19
8	256	261.54	0.02	12078.55	505.77
9	512	511.44	-0.01	39296.65	1876.18
10	1024	1050.95	0.03	227273.44	22374.79

Table 2.9: Scenario 4-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.3071437	0.5063944	0.1862921	0.0001699

Table 2.10: Scenario 4-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

Target gene	True C	$\hat{C}$	Relative bias	$\operatorname{var}(\hat{c})$	$\operatorname{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.05	0.02	4.39	0.07
2	4	4.08	0.02	10.10	0.31
3	8	8.01	0.001	12.29	0.73
4	16	15.82	-0.01	25.98	2.22
5	32	31.73	-0.01	85.69	7.32
6	64	62.72	-0.02	372.73	22.88
7	128	130.68	0.02	2849.16	87.61
8	256	261.44	0.02	12043.91	493.15
9	512	511.23	-0.01	37583.94	1809.78
10	1024	1050.29	0.03	2072203.47	20752.00

Table 2.11: Scenario 5-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = \delta_{Tg} = 2$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.2039163	0.3362011	0.1236815	0.3362011

Table 2.12: Scenario 5-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = \delta_{Tg} = 2$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.12	0.06	0.44	0.36
2	4	4.18	0.05	0.91	0.58
3	8	8.14	0.02	2.57	1.14
4	16	15.82	-0.01	9.18	3.19
5	32	31.73	-0.01	34.01	17.86
6	64	61.18	-0.04	106.35	87.24
7	128	137.96	0.08	793.65	641.98
8	256	273.47	0.07	4601.23	2341.89
9	512	539.12	0.05	28777.39	15134.99
10	1024	1302.24	0.27	615064.02	919306.05

Table 2.13: Scenario 6-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.28749	0.4739908	0.1743715	0.0641477

Table 2.14: Scenario 6-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.11	0.06	0.44	0.36
2	4	4.21	0.05	1.25	0.56
3	8	8.10	0.01	2.92	0.80
4	16	15.75	-0.02	8.91	1.75
5	32	31.64	-0.01	30.26	5.28
6	64	61.62	-0.04	91.70	27.07
7	128	134.73	0.05	428.93	311.68
8	256	270.84	0.06	2512.33	1136.04
9	512	518.42	0.01	9844.24	3791.40
10	1024	1138.54	0.11	154617.32	29456.11

Table 2.15: Scenario 7-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.2996401	0.494023	0.1817409	0.024596

Table 2.16: Scenario 7-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.11	0.06	2.14	0.30
2	4	4.25	0.05	4.91	1.30
3	8	8.09	0.01	5.69	2.98
4	16	15.76	-0.02	11.74	8.99
5	32	31.66	-0.01	35.60	30.29
6	64	61.72	-0.04	142.48	92.06
7	128	134.22	0.05	1068.80	391.77
8	256	269.86	0.05	4841.14	2215.72
9	512	516.27	0.01	15049.46	8875.46
10	1024	1117.67	0.09	130783.97	124893.34

Table 2.17: Scenario 8-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.3071437	0.5063944	0.1862921	0.0001699

Table 2.18: Scenario 8-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 0.67$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.12	0.06	2.14	0.30
2	4	4.22	0.05	5.27	1.35
3	8	8.08	0.01	5.82	3.03
4	16	15.75	-0.02	11.45	9.10
5	32	31.65	-0.01	34.26	30.47
6	64	61.76	-0.03	135.09	91.09
7	128	134.23	0.05	1036.58	378.96
8	256	269.59	0.05	4780.34	2157.83
9	512	515.52	0.01	14029.73	8399.37
10	1024	1113.19	0.09	113620.56	116191.12

Table 2.19: Scenario 9-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = \delta_{Tg} = 2$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.2039163	0.3362011	0.1236815	0.3362011

Table 2.20: Scenario 9-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = \delta_{Tg} = 2$.

Target gene	True C	$\hat{C}$	Relative bias	$\mathrm{var}(\hat{c})$	$\mathrm{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.20	0.10	1.08	0.28
2	4	4.33	0.08	2.10	0.66
3	8	8.30	0.04	5.92	1.09
4	16	15.93	-0.001	20.74	2.65
5	32	31.95	-0.001	77.29	15.29
6	64	60.42	-0.06	227.83	73.15
7	128	146.45	0.14	2498.21	500.23
8	256	291.38	0.14	13048.81	3001.76
9	512	593.47	0.16	122128.82	16041.43
10	1024	1230.03	0.20	738286.86	274758.29

Table 2.21: Scenario 10-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.28749	0.4739908	0.1743715	0.0641477

Table 2.22: Scenario 10-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.21	0.10	1.07	0.28
2	4	4.38	0.09	2.93	0.53
3	8	8.24	0.03	6.67	0.74
4	16	15.79	-0.01	20.01	1.33
5	32	31.77	-0.01	68.96	4.70
6	64	60.96	-0.05	198.44	25.77
7	128	139.60	0.09	1110.49	253.50
8	256	282.84	0.10	6188.39	1033.67
9	512	531.80	0.04	26880.12	2650.46
10	1024	1309.99	0.28	682290.18	75993.81

Table 2.23: Scenario 11-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.2996401	0.494023	0.1817409	0.024596

Table 2.24 Scenario 11-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.21	0.10	1.61	0.72
2	4	4.39	0.09	4.04	3.04
3	8	8.22	0.03	4.49	6.80
4	16	15.80	-0.01	9.62	20.19
5	32	31.80	-0.01	27.59	68.98
6	64	61.12	0.05	96.30	199.71
7	128	138.62	0.08	729.14	973.81
8	256	280.70	0.10	3487.17	5379.65
9	512	527.48	0.03	9517.49	24186.42
10	1024	1248.83	0.22	120064.07	474241.73

Table 2.25: Scenario 12-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.3071437	0.5063944	0.1862921	0.0001699

Table 2.26: Scenario 12-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.00$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.12	0.06	2.14	0.30
2	4	4.22	0.05	5.27	1.35
3	8	8.08	0.01	5.82	3.03
4	16	15.75	-0.02	11.45	9.10
5	32	31.65	-0.01	34.26	30.47
6	64	61.76	-0.03	135.09	91.09
7	128	134.23	0.05	1036.58	378.96
8	256	269.59	0.05	4780.34	2157.83
9	512	515.52	0.01	14029.73	8399.37
10	1024	1113.19	0.09	113620.56	116191.12

Table 2.27: Scenario 13-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = \delta_{Tg} = 2$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.2039163	0.3362011	0.1236815	0.3362011

Table 2.28: Scenario 13-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = \delta_{Tg} = 2$.

Target gene	True C	$\hat{C}$	Relative bias	$\mathrm{var}(\hat{c})$	$\mathrm{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.32	0.16	1.55	0.22
2	4	4.51	0.13	3.90	0.43
3	8	8.51	0.06	10.98	0.81
4	16	16.16	0.01	35.10	2.29
5	32	32.41	0.1	141.72	15.84
6	64	60.04	-0.06	392.70	52.20
7	128	159.84	0.25	2176.21	324.30
8	256	319.39	0.25	34028.99	1785.79
9	512	636.04	0.25	64720.63	6197.86
10	1024	1309.88	0.28	778270.11	225127.48

Table 2.29: Scenario 14-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.2039163	0.3362011	0.1236815	0.3362011
	Target probe	0.28749	0.4739908	0.1743715	0.0641477

Table 2.30: Scenario 14-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = 2$, $\delta_{Tg} = 4$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.33	0.16	2.04	0.22
2	4	4.60	0.15	5.51	0.43
3	8	8.43	0.05	12.24	0.75
4	16	15.95	-0.01	39.20	1.34
5	32	32.10	0.01	126.48	4.86
6	64	60.63	-0.05	344.32	18.60
7	128	145.98	0.14	2561.09	161.68
8	256	295.59	0.17	12503.20	644.65
9	512	555.24	0.08	63734.99	2058.49
10	1024	1477.71	0.53	7299997.89	687173.40

Table 2.31: Scenario 15-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.2996401	0.494023	0.1817409	0.024596

Table 2.32: Scenario 15 Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 5$.

Target gene	True C	$\hat{C}$	Relative bias	$\operatorname{var}(\hat{c})$	$\operatorname{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.32	0.16	1.40	0.81
2	4	4.62	0.15	5.95	2.46
3	8	8.39	0.05	12.64	2.63
4	16	15.94	-0.01	36.84	5.22
5	32	32.11	0.01	127.05	17.27
6	64	144.07	-0.05	344.56	50.46
7	128	144.07	0.13	1979.18	353.20
8	256	293.72	0.15	10329.43	2245.39
9	512	543.29	0.06	49803.55	8211.02
10	1024	1528.28	0.49	2332952.80	260826.03

Table 2.33: Scenario 16-- Standard Deviation of the Noise and Separation between Spike-in Probe and Target Probes with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

$\varepsilon_{s,p,i}$	Separation	π_{gp1}	π_{gp2}	π_{gp3}	π_{gp4}
0.67	Spike-in probe	0.12925	0.2130973	0.0783941	0.5792585
	Target probe	0.3071437	0.5063944	0.1862921	0.0001699

Table 2.34: Scenario 16-- Relative Bias, Average of Variance and Variance of Estimate Absolute mRNA with $\varepsilon_{s,p,i} = 1.33$, $\delta_{sg} = 1$, $\delta_{Tg} = 10$.

Target gene	True C	$\hat{C}$	Relative bias	$\text{var}(\hat{c})$	$\text{var}(\hat{c} \mid \underline{\hat{\theta}})$
1	2	2.32	0.16	1.39	0.81
2	4	4.62	0.15	5.96	2.46
3	8	8.39	0.05	12.64	2.63
4	16	15.94	-0.01	36.84	5.22
5	32	32.10	0.01	127.05	17.27
6	64	60.89	-0.05	344.56	50.46
7	128	144.07	0.13	1979.18	353.20
8	256	293.72	0.15	10329.43	2245.39
9	512	543.29	0.06	49803.55	8211.02
10	1024	1528.27	0.49	2321952.80	260826.03

From above tables, we can see that our method works very well based on the estimates, relative bias and variance. The value of vary standard deviation of the noise ($\varepsilon_{s,p,i}$) in [2.8] is smaller and the value of separation between spike-in probes and target probes is smaller, the result is better!

CHAPTER 3

OPTIMAL CHOICE FOR SPIKE-IN PROBES

3.1 Motivation

The question considered here is, given a set of known target probes and the opportunity to design spike-in probes, how to choose spike-in probes to minimize the variance of our absolute concentration estimator. This topic would be of interest in chip design.

Our simulations suggest, and theoretical results to be given here confirm, that under our working assumptions, our absolute concentration estimates are approximately unbiased, and approximately normal, with a sampling variance which depends, among other things, on the spike-in or target probes reparative.

Assuming that target probes are given, we proceed by deriving the variance of our absolute concentration estimator in terms of the spike-in probe feature. This is possible, using standard variance probative results (delta method), since our procedure consists of consecutive applications of well studied tools (non-linear least squares, linear least square):

1. Distribution of Langmuir parameters for spike-in probes;
2. OLS estimator of universal $\gamma' s \,\&\, C' s$;
3. Distribution of Langmuir parameter for target probes;
4. Distribution of estimator of target absolute concentration.

Then, we minimize the variance of estimator of target absolute concentration, to get the optimal choice of the probability of bite (probability of number of A, T, C and G on the spike-in probe), we minimize the variance in two scenarios:

1. One gene a time;
2. More than one gene a time.

3.2 Variance of $\hat{c}_{TA}$ in Term of Spike-in Probe Features

3.2.1 Distribution of Langmuir Parameter Estimates for Spike-in Probes

The model we use for spike-in probe is Log Langmuir adsorption model

$$I_{spi} = (a_{sp} \frac{c_{si}}{b_{sp} + c_{si}} + d_{sp}) * \varepsilon_{spi} \qquad [3.1]$$

where i is array index. We do it one gene a time.

Rewrite the model [3.1] as:

$$\log I_{spi} = \log(a_{sp} \frac{c_{si}}{b_{sp} + c_{si}} + d_{sp}) + \log \varepsilon_{spi} \qquad [3.2]$$

Assume $\log \varepsilon_{spi} \sim N(0, \sigma_{sp}^2)$, where σ_{sp}^2 is assumed unknown.

Let $\underline{\hat{\theta}}_{sp} = \begin{bmatrix} \hat{a}_{sp} \\ \hat{b}_{sp} \\ \hat{d}_{sp} \end{bmatrix}$

Where s denotes spike-in, p is probe index, and $\hat{a}, \hat{b}$ and $\hat{d}$ are Langmuir parameter estimates.

Those estimates are obtained, one probe, p , at a time, by minimizing

$$\sum_{i=1}^{n} [\log I_{spi} - \log(a_{sp} \frac{c_{si}}{b_{sp} + c_{si}} + d_{sp})]^2 ,$$

where the spike-in concentrations $c_{si}, i = 1,...n$ are known for all the arrays. By well known properties of non-linear least square estimators (Seber and Wild [1989]), we have

$$\underline{\hat{\theta}}_{sp(3\times1)} = \begin{bmatrix} \hat{a}_{sp} \\ \hat{b}_{sp} \\ \hat{d}_{sp} \end{bmatrix} \sim N(\theta_{sp(3\times1)}, \underline{S}_{sp(3\times3)}) ,$$

Where $\underline{S}_{sp(3\times3)} = \sigma_{sp}^2 * [\underline{D}(\underline{\theta}_{sp})^T * \underline{D}(\underline{\theta}_{sp})]^{-1}$

While

$\mu_{spi} = \log(a_{sp} \frac{c_{si}}{b_{sp} + c_{si}} + d_{sp})$ and

$$\underline{D}(\underline{\theta}_{sp})_{(n\times3)} = \begin{bmatrix} \frac{\partial\mu_{sp1}}{\partial a_{sp}} & \frac{\partial\mu_{sp1}}{\partial b_{sp}} & \frac{\partial\mu_{sp1}}{\partial d_{sp}} \\ \vdots & \vdots & \vdots \\ \frac{\partial\mu_{spn}}{\partial a_{sp}} & \frac{\partial\mu_{spn}}{\partial b_{sp}} & \frac{\partial\mu_{spn}}{\partial d_{sp}} \end{bmatrix}.$$

Where n indexes the number of array.

3.2.2 OLS Estimator of Universal $\gamma's$ & $C's$

For each spike-in probe, let

$\hat{\underline{\theta}}_{sp} = \begin{bmatrix} \hat{a}_{sp} \\ \hat{b}_{sp} \\ \hat{d}_{sp} \end{bmatrix}$, and

$\underline{\beta}_{i(4\times1)} = \begin{bmatrix} \gamma_i^A \\ \gamma_i^C \\ \gamma_i^G \\ C_i \end{bmatrix}$ $i = a, b \;\&\; d$, the universal parameters

$\underline{\beta}_{(12\times1)} = \begin{bmatrix} \gamma_a^A \\ \gamma_a^C \\ \gamma_a^G \\ C_a \\ \vdots \\ \vdots \\ \gamma_d^A \\ \gamma_d^C \\ \gamma_d^G \\ C_d \end{bmatrix} = \begin{bmatrix} \underline{\beta}_a \\ \underline{\beta}_b \\ \underline{\beta}_d \end{bmatrix}$, is assumed same for all probes on the array, and $\underline{X}^{*}_{sp(1\times 4)} = [n^{A}_{sp} \quad n^{C}_{sp} \quad n^{G}_{sp} \quad 1]$, sp is the index of spike-in probe, n^{A}_{sp}, n^{C}_{sp} and n^{G}_{sp} are the number of A, C and G on the probe;

$$\underline{X}_{sp(3\times 12)} = \begin{bmatrix} \underline{X}^{*}_{sp} & 0 & 0 \\ 0 & \underline{X}^{*}_{sp} & 0 \\ 0 & 0 & \underline{X}^{*}_{sp} \end{bmatrix}$$, then for one gene

$$\underline{X}_{s(84\times 12)} = \begin{bmatrix} \underline{X}_{s1} \\ \underline{X}_{s2} \\ \vdots \\ \vdots \\ \underline{X}_{s28} \end{bmatrix},$$

Model [3.2] can be applied to:

$$\begin{bmatrix} \ln \hat{\underline{\theta}}_{s1(3\times 1)} \\ \ln \hat{\underline{\theta}}_{s2(3\times 1)} \\ \vdots \\ \ln \hat{\underline{\theta}}_{s28(3\times 1)} \end{bmatrix} = \begin{bmatrix} \underline{X}_{s1} \\ \underline{X}_{s2} \\ \vdots \\ \underline{X}_{s28} \end{bmatrix} \cdot \underline{\beta} + \underline{\varepsilon}_{(84\times 1)} \qquad [3.3]$$

where $\underline{\varepsilon}_{(84\times 84)} \sim N(0, \underline{\Gamma}_{S}(\underline{\theta}_{S}))$.

By using $\delta - method$, let

$$\underline{T}_{sp(3\times 3)} = \begin{bmatrix} \frac{\partial \ln a_{sp}}{\partial a_{sp}} & \frac{\partial \ln a_{sp}}{\partial b_{sp}} & \frac{\partial \ln a_{sp}}{\partial d_{sp}} \\ \frac{\partial \ln b_{sp}}{\partial a_{sp}} & \frac{\partial \ln b_{sp}}{\partial b_{sp}} & \frac{\partial \ln b_{sp}}{\partial d_{sp}} \\ \frac{\partial \ln d_{sp}}{\partial a_{sp}} & \frac{\partial \ln d_{sp}}{\partial b_{sp}} & \frac{\partial \ln d_{sp}}{\partial d_{sp}} \end{bmatrix} = \begin{bmatrix} \frac{1}{a_{sp}} & 0 & 0 \\ 0 & \frac{1}{b_{sp}} & 0 \\ 0 & 0 & \frac{1}{d_{sp}} \end{bmatrix},$$

then

$$\underline{\Gamma}_{sp(3\times 3)} = (\underline{T}_{sp}(\ln \hat{\underline{\theta}}_{sp})^{T} \underline{S}_{sp} \underline{T}_{sp}(\ln \hat{\underline{\theta}}_{sp})), \qquad [3.4]$$

where $\underline{S}_{sp(3\times3)} = \sigma_{sp}^2 * [\underline{D}(\underline{\theta}_{sp})^T * \underline{D}(\underline{\theta}_{sp})]^{-1}$.

By assuming each probe is independent, so for one spike-in gene (28 probes), we have

$$\underline{\Gamma}_{s(84x84)} = \begin{bmatrix} \underline{\Gamma}_{s1} & 0 & 0 & 0 \\ 0 & \underline{\Gamma}_{s2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \underline{\Gamma}_{s28} \end{bmatrix}. \qquad [3.5]$$

Since $\hat{\underline{\beta}} = (\underline{X}_S^T \underline{X}_S)^{-1} \cdot \underline{X}_S^T \cdot \hat{\underline{\theta}}_S$,

by using OLS one gene a time

$$Var(\hat{\underline{\beta}}) = (\underline{X}_s^T \underline{X}_s)^{-1} \underline{X}_s^T \underline{\Gamma}_s \underline{X}_s (\underline{X}_s^T \underline{X}_s)^{-1} = \underline{V} \ . \qquad [3.6]$$

3.2.3 Distribution of Langmuir Parameters for Target Probes

Since $\underline{\beta}$ is assumed universal for all probes on array,

Let

$$\underset{84x1}{\hat{\underline{\theta}}_T} = \begin{bmatrix} \hat{\underline{\theta}}_{T1} \\ \vdots \\ \vdots \\ \hat{\underline{\theta}}_{T28} \end{bmatrix} = \begin{bmatrix} \hat{a}_{T1} \\ \hat{b}_{T1} \\ \hat{d}_{T1} \\ \vdots \\ \vdots \\ \vdots \\ \hat{a}_{T28} \\ \hat{b}_{T28} \\ \hat{d}_{T28} \end{bmatrix}$$ is Langmuir parameters for TA probes, and let

$$\underline{X}^*{}_{TP} = \begin{bmatrix} n_{TP}^A & n_{TP}^B & n_{TP}^C & 1 \end{bmatrix},$$

$$\underset{3x12}{\underline{X}_{TP}} = \begin{bmatrix} \underline{X}^*_{Tp} & 0 & 0 \\ 0 & \underline{X}^*_{Tp} & 0 \\ 0 & 0 & \underline{X}^*_{Tp} \end{bmatrix} = \begin{bmatrix} \underline{X}_{TP1} \\ \underline{X}_{TP2} \\ \underline{X}_{TP3} \end{bmatrix}$$, while

$$\underset{84x12}{\underline{X}_{T}} = \begin{bmatrix} \underline{X}_{T1} \\ \underline{X}_{T2} \\ \underline{X}_{T3} \\ \vdots \\ \vdots \\ \vdots \\ \underline{X}_{T26} \\ \underline{X}_{T27} \\ \underline{X}_{T28} \end{bmatrix},$$

$$\hat{\underline{\theta}}_T = \exp(\underline{X}_T \cdot \hat{\underline{\beta}}) = \begin{bmatrix} \exp(\underline{X}_{TP1} \cdot \underline{\beta}) \\ \exp(\underline{X}_{TP2} \cdot \underline{\beta}) \\ \exp(\underline{X}_{TP3} \cdot \underline{\beta}) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} g_{11} \\ g_{12} \\ g_{13} \\ \vdots \\ \vdots \\ \vdots \\ g_{281} \\ g_{282} \\ g_{283} \end{bmatrix},$$

then we have

$$\ln \hat{\underline{\theta}}_T = \underline{X}_T \cdot \hat{\underline{\beta}} + \underline{\varepsilon}. \qquad [3.7]$$

The variance of $\hat{\underline{\theta}}_T$ is wanted, let

$\underline{X}_{Ti}\exp(\underline{X}_{Ti}\underline{\beta}) = \begin{bmatrix} \underline{X}_{Ti1}\exp(\underline{X}_{Ti1}\underline{\beta}) \\ \underline{X}_{Ti2}\exp(\underline{X}_{Ti2}\underline{\beta}) \\ \underline{X}_{Ti3}\exp(\underline{X}_{Ti3}\underline{\beta}) \end{bmatrix}$ and $\underline{m}_i = \underline{X}_{Ti}\exp(\underline{X}_{Ti}\underline{\beta})$, the Jacobin Matrix of the map from $\underline{\beta}$ to $\underline{m}_i$ is:

$$\underset{84x12}{\underline{J_g(\beta)}} = \begin{bmatrix} \frac{\partial g_{11}}{\partial \underline{\beta}}_{(1x12)} \\ \frac{\partial g_{12}}{\partial \underline{\beta}}_{(1x12)} \\ \frac{\partial g_{13}}{\partial \underline{\beta}}_{(1x12)} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} \underline{X}_{T11}\exp(\underline{X}_{T11}\underline{\beta}) \\ \underline{X}_{T12}\exp(\underline{X}_{T12}\underline{\beta}) \\ \underline{X}_{T13}\exp(\underline{X}_{T13}\underline{\beta}) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} \underline{X}_{T1}\exp(\underline{X}_{T1}\underline{\beta}) \\ \underline{X}_{T2}\exp(\underline{X}_{T2}\underline{\beta}) \\ \underline{X}_{T3}\exp(\underline{X}_{T3}\underline{\beta}) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \underline{X}_{T28}\exp(\underline{X}_{T28}\underline{\beta}) \end{bmatrix}_{\substack{3x12 \\ \\ 3x12}} = \begin{bmatrix} \underline{m}_1 \\ \\ \\ \\ \underline{m}_{28} \end{bmatrix}_{\substack{3x12 \\ \\ 3x12}}$$

Let $C_{ij(3\times 3)} = \underline{m}_{i(3\times 12)} \cdot \underline{V}_{(12\times 12)} \cdot \underline{m}^{T}_{i(12\times 3)}$, then the variance of $\hat{\underline{\theta}}_T$ can be expressed as

$$Var(\hat{\underline{\theta}}_T) \approx \underline{J}_g(\hat{\underline{\beta}}) \cdot Var(\hat{\underline{\beta}}) \cdot \underline{J}^{T}_g(\hat{\underline{\beta}})$$

$$= \begin{bmatrix} \underline{m}_1 \\ \vdots \\ \underline{m}_{28} \end{bmatrix}_{84x12} \cdot \underbrace{\underline{V}}_{12x12} \cdot \begin{bmatrix} \underline{m}_1 \\ \vdots \\ \underline{m}_{28} \end{bmatrix}^{T}_{12x84}$$

$$= \begin{bmatrix} \underline{m}_1 \cdot \underline{V} \cdot \underline{m}_1^T & \underline{m}_1 \cdot \underline{V} \cdot \underline{m}_2^T & \cdots & \cdots & \underline{m}_1 \cdot \underline{V} \cdot \underline{m}_{28}^T \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{m}_{28} \cdot \underline{V} \cdot \underline{m}_1^T & \underline{m}_{28} \cdot \underline{V} \cdot \underline{m}_2^T & \cdots & \cdots & \underline{m}_{28} \cdot \underline{V} \cdot \underline{m}_{28}^T \end{bmatrix} \qquad [3.8]$$

$$= \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & \cdots & C_{1,28} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{28,1} & C_{28,2} & \cdots & \cdots & C_{28,28} \end{bmatrix}.$$

3.2.4 Distribution of $\hat{C}_T$

Log Langmuir model for Target probe

$$LogI_{TPi} = \log(\frac{\hat{a}_{TP}C_T}{\hat{b}_{TP} + C_T} + \hat{d}_{TP}) + \log \varepsilon_{TPi}$$

$$= f(\underline{\hat{\theta}}_{Tp}; C_T) + \log \varepsilon_{TPi}$$

Since $Var(\log \varepsilon_{TPi} \mid \underline{\hat{\theta}}_T) = \sigma_T^2$ by assumption

$$Var(\hat{C}_T \mid \underline{\hat{\theta}}_T) = \sigma_T^2 * (\dot{f}^T \cdot \dot{f})^{-1} \qquad [3.9]$$

where

$$\dot{f}(\underline{\hat{\theta}}_T)_{(28\times1)} = \begin{bmatrix} \frac{\partial f}{\partial C_T}(\hat{C}_T, \underline{\hat{\theta}}_{T1}) \\ \vdots \\ \vdots \\ \vdots \\ \frac{\partial f}{\partial C_T}(\hat{C}_T, \underline{\hat{\theta}}_{T28}) \end{bmatrix} = \begin{bmatrix} \dot{f}_1(\underline{\hat{\theta}}_{T1}) \\ \vdots \\ \vdots \\ \vdots \\ \dot{f}_{28}(\underline{\hat{\theta}}_{T28}) \end{bmatrix} = \begin{bmatrix} \frac{a_1 b_1}{a_1\hat{C}_T(b_1 + \hat{C}_T) + d_1(b_1 + \hat{C}_T)^2} \\ \vdots \\ \vdots \\ \vdots \\ \frac{a_{28}b_{28}}{a_{28}\hat{C}_T(b_{28} + \hat{C}_T) + d_{28}(b_{28} + \hat{C}_T)^2} \end{bmatrix}$$

$$\Rightarrow \qquad Var(\hat{C}_T) = Var(E(\hat{C}_T \mid \underline{\hat{\theta}}_T)) + E_{\underline{\hat{\theta}}_T}(Var(\hat{C}_T \mid \underline{\hat{\theta}}_T))$$

Since $E(\hat{C}_T \mid \underline{\hat{\theta}}_T)$ is a constant, so $Var(E(\hat{C}_T \mid \underline{\hat{\theta}}_T)) = 0$, and combine with equation [3.9], we rewrite the variance as

$$Var(\hat{C}_T) = 0 + E_{\hat{\underline{\theta}}_T}(Var(\hat{C}_T \mid \hat{\underline{\theta}}_T))$$
$$= 0 + E[\sigma_T^2 \cdot (\dot{f}^T(\hat{\underline{\theta}}_T) \cdot \dot{f}\ \ (\hat{\underline{\theta}}_T))^{-1}]$$

Since σ_T^2 is a constant, so

$$Var(\hat{C}_T) = \ \sigma_T^2 * E(\dot{f}^T(\hat{\underline{\theta}}_T) \cdot \dot{f}\ \ (\hat{\underline{\theta}}_T))^{-1}.$$

By using the $\delta - method$

$$E(\dot{f}^T(\hat{\underline{\theta}}_T) \cdot \dot{f}\ \ (\hat{\underline{\theta}}_T))^{-1} \approx \frac{1}{E(\dot{f}^T(\hat{\underline{\theta}}_T) \cdot \dot{f}\ \ (\hat{\underline{\theta}}_T))}, \qquad [3.10]$$

and

$$\begin{aligned} E[\dot{f}^T(\hat{\underline{\theta}}_T) \cdot \dot{f}(\hat{\underline{\theta}}_T)] &= [E(\dot{f}^T(\hat{\underline{\theta}}_T))] * [E(\dot{f}(\hat{\underline{\theta}}_T))] + tr(\mathrm{var}(\dot{f}^T(\hat{\underline{\theta}}_T)) \\ &\approx \dot{f}^T(\hat{\underline{\theta}}_T) * \dot{f}(\hat{\underline{\theta}}_T) + tr(\mathrm{var}(\dot{f}^T(\hat{\underline{\theta}}_T)) \end{aligned} \qquad [3.11]$$

Combine [3.10] and [3.11], the variance can be rewritten as

$$Var(\hat{C}_T) = \frac{\sigma_T^2}{\dot{f}^T(\hat{\underline{\theta}}_T) \cdot \dot{f}\ \ (\hat{\underline{\theta}}_T) + tr[Var(\dot{f}^T(\hat{\underline{\theta}}_T))]} \qquad [3.12]$$

To evaluate the denominator of [3.12], we let $\underbrace{B_i}_{1x3} = \frac{\partial \dot{f}_i}{\partial \underline{\theta}_i} = \begin{bmatrix} \frac{\partial \dot{f}_i}{\partial a_i} & \frac{\partial \dot{f}_i}{\partial b_i} & \frac{\partial \dot{f}_i}{\partial d_i} \end{bmatrix}$, then the Jacobin Matrix of the map from $\hat{\underline{\theta}}_T$ to $\underline{B}_i$ is

$$J_{\dot{f}}(\hat{\underline{\theta}}_T)_{(28\times 84)} = \begin{bmatrix} \frac{\partial f_1}{\partial \underline{\theta}_1} & 0 & 0 & 0 \\ 0 & \frac{\partial f_2}{\partial \underline{\theta}_2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{\partial f_{28}}{\partial \underline{\theta}_{28}} \end{bmatrix} = \begin{bmatrix} B_1 & 0 & 0 & 0 \\ 0 & B_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & B_{28} \end{bmatrix}.$$

By δ - Method, and substitute equation [3.8] into:

$$Var(\dot{f}\ \ (\underline{\hat{\theta}}_T)) = J_{\dot{f}}(\underline{\hat{\theta}}_T) \cdot Var(\underline{\hat{\theta}}_T) \cdot J^T{}_{\dot{f}}(\underline{\hat{\theta}}_T)$$

$$= \begin{bmatrix} B_1 & 0 & 0 & 0 \\ 0 & B_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & B_{28} \end{bmatrix} \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,28} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ c_{28,1} & c_{28,2} & \cdots & c_{28,28} \end{bmatrix} \begin{bmatrix} B_1^T & 0 & 0 & 0 \\ 0 & B_2^T & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & B_{28}^T \end{bmatrix}$$

$$= \begin{bmatrix} B_1 c_{1,1} B_1^T & B_1 c_{1,2} B_2^T & \cdots & B_1 c_{1,28} B_{28}^T \\ B_2 c_{2,2} B_2^T & B_2 c_{2,2} B_2^T & \cdots & B_2 c_{2,28} B_{28}^T \\ \vdots & \vdots & \ddots & \vdots \\ B_{28} c_{28,1} B_1^T & \cdots & \cdots & B_{28} c_{28,28} B_{28}^T \end{bmatrix}$$

$$\therefore \quad tr(\text{var}(\dot{f}(\underline{\theta}_T))) = \sum_{i=1}^{28} B_i C_{ii} B_i^T$$

$$= \sum_{i=1}^{28} B_i \underline{m}_i \underline{V} \underline{m}_i^T B_i^T$$

Where $B_i, \underline{m}_i$ depend on Target probes, only $\underline{V}$ depends on spike-in probes.

$$\therefore \qquad Var(\hat{c}_T) = \frac{\sigma_T^2}{\sum_{i=1}^{28} [f_i^2(\underline{\theta}_T) + B_i \underline{m}_i \underline{V} \underline{m}_i^T B_i^T]} \qquad [3.13]$$

where $\dot{f}_i(\underline{\hat{\theta}}_T) = \dfrac{a_i b_i}{a_i \hat{C}_T (b_i + \hat{C}_T) + d_i (b_i + \hat{C}_T)^2}$ is a scalar.

Only $\underline{V}$ depends on spike-in probes.

3.3 Minimize $Var(\hat{c}_T)$ Respect to Spike-in Probe Features

Since only $\underline{V}$ depends on spike-in probes, so we only care about the denominator.

$$\underline{V} = Var(\underline{\hat{\beta}}) = (\underline{X}_s^T \underline{X}_s)^{-1} \underline{X}_s^T \underline{\Gamma}_s \underline{X}_s (\underline{X}_s^T \underline{X}_s)^{-1} . \qquad [3.6]$$

3.3.1 Assumption

Assume $\underline{X}^*_{Sp(1x4)}$ iid over $p = 1,2....28$.

$\Rightarrow \underline{X}^*_{Sp(1x4)} \sim Mult(n, \underline{\pi})$,

Where $\underline{\pi} = [\pi_1\ \pi_2\ \pi_3\ \pi_4]$, n=25. let

$$\underline{X}_{Sp(3x12)} = \begin{bmatrix} \underline{X}^*_{Sp(1x4)} & 0 & 0 \\ 0 & \underline{X}^*_{Sp(1x4)} & 0 \\ 0 & 0 & \underline{X}^*_{Sp(1x4)} \end{bmatrix}$$, then for one gene

$$\underline{X}_{s(84\times 12)} = \begin{bmatrix} \underline{X}_{s1} \\ \underline{X}_{s2} \\ \vdots \\ \vdots \\ \underline{X}_{s28} \end{bmatrix}.$$

3.3.2 One Gene a Time

Let

$$\underline{D}(\underline{\pi}) = \begin{bmatrix} \pi_1 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 \\ 0 & 0 & \pi_3 & 0 \\ 0 & 0 & 0 & \pi_4 \end{bmatrix}$$

$$\because \quad E(\underline{X}^{*T}_{Sp} \cdot \underline{X}^*_{Sp}) = Var(\underline{X}^*_{Sp}) + E(\underline{X}^*_{Sp})^T \cdot E(\underline{X}^*_{Sp})$$
$$= 25(D(\underline{\pi}) + (25-1)\underline{\pi}\underline{\pi}^T) = \underline{E}$$

Where

$$\underline{X}_{Sp(3x12)} = \begin{bmatrix} \underline{X}^*_{Sp(1x4)} & 0 & 0 \\ 0 & \underline{X}^*_{Sp(1x4)} & 0 \\ 0 & 0 & \underline{X}^*_{Sp(1x4)} \end{bmatrix}.$$

$$\therefore \quad E(\underline{X}^T_{sp} \cdot \underline{X}_{sp}) = \begin{bmatrix} \underline{E} & 0 & 0 \\ 0 & \underline{E} & 0 \\ 0 & 0 & \underline{E} \end{bmatrix} = I_3 \otimes \underline{E}$$

$$\frac{1}{28}\underline{X}^T_S \underline{X}_S = \frac{1}{28}\sum_{p=1}^{28} \underline{X}^T_{Sp}\underline{X}_{Sp} \xrightarrow{p} E(\underline{X}^T_{sp}\underline{X}_{sp}) = I_3 \otimes \underline{E}$$

$$\therefore \ \underline{X}^T_S \underline{X}_S = 28 * I_3 \otimes \underline{E}$$

Assume the variance matrix $\underline{\Gamma}$ is same for all probes, so

$$\Gamma_S = \begin{bmatrix} \underline{\Gamma} & & & \\ & \underline{\Gamma} & & \\ & & \ddots & \\ & & & \underline{\Gamma} \end{bmatrix}$$, where $\underline{\Gamma} = \begin{bmatrix} \Gamma_{11} & \Gamma_{21} & \Gamma_{31} \\ \Gamma_{21} & \Gamma_{22} & \Gamma_{23} \\ \Gamma_{31} & \Gamma_{32} & \Gamma_{33} \end{bmatrix}$, since $\underline{X}^*_{Sp(1x4)}$ iid

over $p = 1,2....28.$, so $\underline{X}_{Sp(1x4)}$ also iid over $p = 1,2....28.$, then

$$\begin{aligned} \underline{X}^T_S \Gamma_S \underline{X}_S &= \frac{1}{28}\sum_{p=1}^{28} \underline{X}^T_{Sp}\underline{\Gamma}\underline{X}_{Sp} \\ &\approx E(\underline{X}^T_{Sp}\underline{\Gamma}\underline{X}_{Sp}) \\ &= E\begin{bmatrix} \Gamma_{11}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} & \Gamma_{21}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} & \Gamma_{31}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} \\ \Gamma_{21}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} & \Gamma_{22}\cdot\underline{X}^T_{Sp}\cdot\underline{X}_{Sp} & \Gamma_{23}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} \\ \Gamma_{31}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} & \Gamma_{32}\cdot\underline{X}^T_{Sp}\cdot\underline{X}^*_{Sp} & \Gamma_{33}\cdot\underline{X}^T_{Sp}\cdot\underline{X}_{Sp} \end{bmatrix} \\ &= \underline{\Gamma} \otimes \underline{E} \end{aligned}$$

Therefore

$$\begin{aligned}
\underline{V} &= (\underline{X}_S^T \underline{X}_S)^{-1} \cdot (\underline{X}_S^T \Gamma_S \underline{X}_S) \cdot (\underline{X}_S^T \underline{X}_S)^{-1} \\
&= \frac{1}{28} \cdot (\frac{1}{28} \cdot \underline{X}_S^T \underline{X}_S)^{-1} \cdot (\frac{1}{28} \cdot \underline{X}_S^T \underline{\Gamma}_S \underline{X}_S) \cdot (\frac{1}{28} \cdot \underline{X}_S^T \underline{X}_S)^{-1} \\
&= \frac{1}{28} \cdot (I_3 \otimes \underline{E}^{-1}) \cdot (\underline{\Gamma} \otimes \underline{E}) \cdot (I_3 \otimes \underline{E}^{-1}) \qquad [3.14] \\
&= \frac{1}{28} \cdot (\underline{\Gamma} \otimes \underline{E}^{-1})
\end{aligned}$$

$$\because \qquad \underline{E} = 25(D(\underline{\pi}) + (25-1)\underline{\pi}\underline{\pi}^T)$$

$$\begin{aligned}
\therefore \qquad \underline{E}^{-1} &= \frac{1}{25}[D^{-1}(\underline{\pi}) - \frac{(\sqrt{25-1} \cdot D^{-1}(\underline{\pi}) \cdot \underline{\pi})(\sqrt{25-1} \cdot \underline{\pi}^T \cdot D^{-1}(\underline{\pi}))}{1+(25-1)\underline{\pi}^T \cdot D^{-1}(\underline{\pi}) \cdot \underline{\pi}}] \\
&= \frac{1}{25}[D^{-1}(\underline{\pi}) - \frac{(25-1) \cdot (\underline{I} \cdot \underline{I}^T)}{1+(25-1)\sum_{j=1}^{4} \pi_j}] \\
&= \begin{bmatrix} \frac{1}{25\pi_1} & 0 & 0 & 0 \\ 0 & \frac{1}{25\pi_2} & 0 & 0 \\ 0 & 0 & \frac{1}{25\pi_3} & 0 \\ 0 & 0 & 0 & \frac{1}{25\pi_4} \end{bmatrix} - \frac{25-1}{25 * 25} \underline{1}_{(4)} \cdot \underline{1}_{(4)}^T \qquad [3.15]
\end{aligned}$$

Form equation [3.13]:

$$Var(\hat{c}_T) = \frac{\sigma_T^2}{\sum_{i=1}^{28}[f_i^2(\underline{\theta}_T) + \underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T]} = \frac{\sigma_T^2}{\sum_{i=1}^{28} f_i^2(\underline{\theta}_T) + \sum_{i=1}^{28} \underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T},$$

where

$$\underline{m}_{i(3x12)} = \underline{X}_{Ti} \bullet \exp(\underline{X}_{Ti} \bullet \underline{\beta})$$

$$= \begin{bmatrix} \underline{X}_{Ti}^{*} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} & 0 & 0 \\ 0 & \underline{X}_{Ti}^{*} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} & 0 \\ 0 & 0 & \underline{X}_{Ti}^{*} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \end{bmatrix}$$

Then combine with equation [3.14], we have

$$\underline{m}_i \underline{V} \underline{m}_i^T = \frac{1}{28} \cdot \underline{\Gamma} \otimes (\underline{X}_{Ti}^{*} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \cdot \underline{E}^{-1} \cdot \underline{X}_{Ti}^{*T} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}})$$

Let $Q_i = \underline{X}_{Ti}^{*} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \cdot \underline{E}^{-1} \cdot \underline{X}_{Ti}^{*T} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}}$

Where $\underline{X}_{Ti}^{*} = [n_{Ti}^{A} \quad n_{Ti}^{C} \quad n_{Ti}^{G} \quad 1]$, n_{Ti}^{A}, n_{Ti}^{C} and n_{Ti}^{G} are the number of A, C and G on the probe i;

$$\underset{3x12}{\underline{X}_{TP}} = \begin{bmatrix} \underline{X}_{Tp}^{*} & 0 & 0 \\ 0 & \underline{X}_{Tp}^{*} & 0 \\ 0 & 0 & \underline{X}_{Tp}^{*} \end{bmatrix} = \begin{bmatrix} \underline{X}_{TP1} \\ \underline{X}_{TP2} \\ \underline{X}_{TP3} \end{bmatrix}.$$

Therefore

$$\underline{m}_i \underline{V} \underline{m}_i^T = \frac{1}{28} \cdot \underline{\Gamma} \otimes Q_i = \frac{1}{28} \cdot \begin{bmatrix} \Gamma_{11} \cdot Q_i & 0 & 0 \\ 0 & \Gamma_{11} \cdot Q_i & 0 \\ 0 & 0 & \Gamma_{11} \cdot Q_i \end{bmatrix}$$

And $\underset{1x3}{\underbrace{B_i}} = \frac{\partial f_i}{\partial \underline{\theta}_i} = \begin{bmatrix} \frac{\partial f_i}{\partial a_i} & \frac{\partial f_i}{\partial b_i} & \frac{\partial f_i}{\partial d_i} \end{bmatrix} = \begin{bmatrix} B_{i1} & B_{i2} & B_{i3} \end{bmatrix}$, so

$$\underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T = \frac{Q_i}{28} \cdot \sum_{i=1}^{3} B_{ij}^2 \cdot \Gamma_{jj} = \frac{Q_i}{28} \cdot s_i$$

where $s_i = \sum_{i=1}^{3} B_{ij}^2 \cdot \Gamma_{jj} > 0$.

So the variance is

$$\begin{aligned} Var(\hat{c}_T) &= \frac{\sigma_T^2}{\sum_{i=1}^{28} f_i^2(\underline{\theta}_T) + \sum_{i=1}^{28} \underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T} \\ &= \frac{\sigma_T^2}{\sum_{i=1}^{28} f_i^2(\underline{\theta}_T) + \sum_{i=1}^{28} \frac{Q_i}{28} \cdot s_i} \end{aligned} \qquad [3.16]$$

Since $\dot{f}_i(\hat{\underline{\theta}}_T) = \frac{a_i b_i}{a_i \hat{C}_T (b_i + \hat{C}_T) + d_i (b_i + \hat{C}_T)^2}$ does not respect to spike-in probes ($\underline{\pi}$), only Q_i depends on $\underline{\pi}$, so we can maximize $\sum_{i=1}^{28} \frac{Q_i}{28} \cdot s_i$ to minimize $Var(\hat{c}_T)$.

To evaluate $Q_i = \underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \cdot \underline{E}^{-1} \cdot \underline{X}_{Ti}^{*T} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}}$,

where $\underline{X}_{Ti}^* = [n_{Ti}^A \quad n_{Ti}^C \quad n_{Ti}^G \quad 1] = \begin{bmatrix} X_{Ti1} & X_{Ti2} & X_{Ti3} & X_{Ti4} \end{bmatrix}$,

substitute [3.15]

$$\underline{B}^{-1} = \begin{bmatrix} \frac{1}{25\pi_1} & 0 & 0 & 0 \\ 0 & \frac{1}{25\pi_2} & 0 & 0 \\ 0 & 0 & \frac{1}{25\pi_3} & 0 \\ 0 & 0 & 0 & \frac{1}{25\pi_4} \end{bmatrix} - \frac{25-1}{25*25} I_{(4x4)}$$

into Q_i, then we have

$$Q_i = \frac{1}{25} \cdot e^{2\underline{X}_{Ti}\underline{\beta}} \left(\sum_{k=1}^{4} \frac{X_{Tik}^2}{\pi_k} - \frac{24}{25} \cdot \sum_{k=1}^{4} X_{Tik}^2 \right)$$

Let $R_i = \frac{24}{25*25} \cdot e^{2\underline{X}_{Ti}\underline{\beta}} \sum_{k=1}^{4} X_{Tik}^2$ and $t_{ki}^2 = \frac{1}{25} \cdot (e^{2\underline{X}_{Ti}\underline{\beta}} X_{Tik}^2)$, then

$$Q_i = \sum_{k=1}^{4} \frac{t_{ti}^2}{\pi_k} - R_i$$

Go back to

$$\begin{aligned} \sum_{i=1}^{28} \frac{Q_i}{28} \cdot s_i &= \frac{1}{28} \sum_{i=1}^{28} s_i (\sum_{k=1}^{4} \frac{t_{ti}^2}{\pi_k} - R_i) \\ &= \frac{1}{28} \sum_{i=1}^{28} \sum_{k=1}^{4} \frac{s_i \cdot t_{ki}^2}{\pi_k} - \frac{1}{28} \sum_{i=1}^{28} s_i \cdot R_i \end{aligned} \qquad [3.17]$$

Only the first part depends on to spike-in probes ($\underline{\pi}$), so we just consider it.

Let $D_{ki}^2 = s_i \cdot t_{ki}^2$, then

$$U = \frac{1}{28} \sum_{i=1}^{28} \sum_{k=1}^{4} \frac{s_i \cdot t_{ki}^2}{\pi_k} = \frac{1}{28} \sum_{i=1}^{28} \sum_{k=1}^{4} \frac{D_{ki}^2}{\pi_k} = \frac{1}{28} \sum_{i=1}^{28} [\frac{D_{i1}^2}{\pi_1} + \frac{D_{i2}^2}{\pi_2} + \frac{D_{i3}^2}{\pi_3} + \frac{D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)}]$$

In order to maximize U respect to $\underline{\pi}$, we take the partial differentiate of U and let them equal to 0:

$$\frac{\partial U}{\partial \pi_1} = -\frac{\sum_{i=1}^{28} D_{i1}^2}{\pi_1^2} + \frac{\sum_{i=1}^{28} D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (1)$$

$$\frac{\partial U}{\partial \pi_2} = -\frac{\sum_{i=1}^{28} D_{i2}^2}{\pi_2^2} + \frac{\sum_{i=1}^{28} D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (2)$$

$$\frac{\partial U}{\partial \pi_3} = -\frac{\sum_{i=1}^{28} D_{i3}^2}{\pi_3^2} + \frac{\sum_{i=1}^{28} D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (3)$$

Let $G_j^2 = \sum_{i=1}^{28} D_{ij}^2$, by (1), (2) and (3)

$$\frac{1}{\pi_1^2}G_1^2 = \frac{1}{\pi_2^2}G_2^2 = \frac{1}{\pi_3^2}G_3^2 \Rightarrow \frac{G_1}{\pi_1} = \frac{G_2}{\pi_2} = \frac{G_3}{\pi_3}$$

By (1)$\Rightarrow$

$$\Rightarrow \pi_1 = \frac{G_1}{G_1 + G_2 + G_3 + G_4}$$

By (2), (3)$\Rightarrow$

$$\pi_2 = \frac{G_2}{G_1 + G_2 + G_3 + G_4}$$

$$\pi_3 = \frac{G_3}{G_1 + G_2 + G_3 + G_4}$$

$$\because \quad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

$$\therefore \quad \pi_4 = \frac{G_4}{G_1 + G_2 + G_3 + G_4}$$

3.3.3 More then One Gene a Time

The variance of absolute concentration of G target genes, with the weight w_g for gene g is

$Var(c) = \sum_{g=1}^{G} w_g \operatorname{var}(\hat{c}_g)$, using equation [3.16], we have

$$\begin{aligned} Var(c) &= \sum_{g=1}^{G} w_g \operatorname{var}(\hat{c}_g) \\ &= \sum_{g=1}^{G} \frac{w_g \cdot \sigma_g^2}{\sum_{i=1}^{28} f_i^2(\underline{\theta}_g) + \sum_{i=1}^{28} \frac{Q_{gi}}{28} \cdot s_{gi}} \end{aligned}$$

Form equation [3.17]

$$\sum_{i=1}^{28}\frac{Q_{gi}}{28}\cdot s_{gi}=\frac{1}{28}\sum_{i=1}^{28}\sum_{k=1}^{4}\frac{s_{gi}\cdot t_{gki}^{2}}{\pi_k}-\frac{1}{28}\sum_{i=1}^{28}s_{gi}\cdot R_{gi}$$

$$=U_g--\frac{1}{28}\sum_{i=1}^{28}s_{gi}\cdot R_{gi}$$

Where

$$U=\frac{1}{28}\sum_{i=1}^{28}\sum_{k=1}^{4}\frac{s_{gi}\cdot t_{gki}^{2}}{\pi_k}=\frac{1}{28}\sum_{i=1}^{28}\sum_{k=1}^{4}\frac{D_{gki}^{2}}{\pi_k}=\frac{1}{28}\sum_{i=1}^{28}[\frac{D_{gi1}^{2}}{\pi_1}+\frac{D_{gi2}^{2}}{\pi_2}+\frac{D_{gi3}^{2}}{\pi_3}+\frac{D_{gi4}^{2}}{(1-\pi_1-\pi_2-\pi_3)}]$$

We have

$$Var(c)=\sum_{g=1}^{G}\frac{w_g\cdot\sigma_g^2}{\sum_{i=1}^{28}f_i^2(\underline{\theta}_g)+\sum_{i=1}^{28}\frac{Q_{gi}}{28}\cdot s_{gi}}$$

$$=\sum_{g=1}^{G}\frac{\sigma_g^2}{\frac{1}{w_g}\sum_{i=1}^{28}f_i^2(\underline{\theta}_g)+\frac{1}{w_g}(U_g-\frac{1}{28}\sum_{i=1}^{28}s_{gi}\cdot R_{gi})}$$

$$=\sum_{g=1}^{G}\frac{\sigma_g^2}{\frac{1}{w_g}\sum_{i=1}^{28}f_i^2(\underline{\theta}_g)+\frac{U_g}{w_g}-\frac{1}{28\cdot w_g}\sum_{i=1}^{28}s_{gi}\cdot R_{gi}}$$

Only U_g depends on spike-in probes, so if want to minimize $Var(c)$, then try to

maximize $\sum_{g=1}^{G}\frac{U_g}{w_g}$,

$$M=\sum_{g=1}^{G}\frac{U_g}{w_g}$$

$$=\sum_{g=1}^{G}\frac{1}{w_g}\cdot\frac{1}{28}\sum_{i=1}^{28}[\frac{D_{gi1}^{2}}{\pi_1}+\frac{D_{gi2}^{2}}{\pi_2}+\frac{D_{gi3}^{2}}{\pi_3}+\frac{D_{gi4}^{2}}{(1-\pi_1-\pi_2-\pi_3)}]$$

$$=\frac{1}{28}\sum_{g=1}^{G}\sum_{i=1}^{28}(\frac{D_{ig1}^{2}/w_g}{\pi_1}+\frac{D_{ig2}^{2}/w_g}{\pi_2}+\frac{D_{ig3}^{2}/w_g}{\pi_3}+\frac{D_{ig4}^{2}/w_g}{\pi_4})$$

Let $S_k^2=\frac{1}{28}\sum_{g=1}^{G}\sum_{i=1}^{28}(D_{igk}^2/w_g)$, we can minimize $Var(c)$ by picking $\underline{\pi}$,

$$\pi_k = \frac{S_k}{\sum_{k=1}^{4} S_k}$$

Detail:

$$\frac{\partial M}{\partial \pi_1} = -\frac{S_1^2}{\pi_1^2} + \frac{S_4^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (1)$$

$$\frac{\partial M}{\partial \pi_2} = -\frac{S_2^2}{\pi_2^2} + \frac{S_4^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (2)$$

$$\frac{\partial M}{\partial \pi_3} = -\frac{S_3^2}{\pi_3^2} + \frac{S_4^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (3)$$

$$\Rightarrow \frac{S_1^2}{\pi_1^2} = \frac{S_2^2}{\pi_2^2} = \frac{S_3^2}{\pi_3^2}$$

$$\Rightarrow \frac{S_1}{\pi_1} = \frac{S_2}{\pi_2} = \frac{S_3}{\pi_3}$$

Therefore

$$\pi_1 = \frac{S_1}{S_1+S_2+S_3+S_4}$$

$$\pi_2 = \frac{S_2}{S_1+S_2+S_3+S_4}$$

$$\pi_3 = \frac{S_3}{S_1+S_2+S_3+S_4}$$

$$\pi_4 = \frac{S_4}{S_1+S_2+S_3+S_4}$$

To check M is a maximum, we look at the Hessian function:

$$H=\begin{bmatrix}\frac{\partial^2 M}{\partial\pi_1^2} & \frac{\partial^2 M}{\partial\pi_1\partial\pi_2} & \frac{\partial^2 M}{\partial\pi_1\partial\pi_3}\\ \frac{\partial^2 M}{\partial\pi_2\partial\pi_1} & \frac{\partial^2 M}{\partial\pi_2^2} & \frac{\partial^2 M}{\partial\pi_2\partial\pi_3}\\ \frac{\partial^2 M}{\partial\pi_3\partial\pi_1} & \frac{\partial^2 M}{\partial\pi_3\partial\pi_2} & \frac{\partial^2 M}{\partial\pi_3^2}\end{bmatrix}$$

$$=\begin{bmatrix}\frac{2S_1^2}{\pi_1^3}+\frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\\ \frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \frac{2S_2^2}{\pi_2^3}+\frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\\ \frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \frac{2S_3^2}{\pi_3^3}+\frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\end{bmatrix}$$

$$=\frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\cdot I_{(3\times 3)}+\begin{bmatrix}\frac{2S_1^2}{\pi_1^3} & 0 & 0\\ 0 & \frac{2S_2^2}{\pi_2^3} & 0\\ 0 & 0 & \frac{2S_3^2}{\pi_3^3}\end{bmatrix}$$

Since $\frac{2S_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\cdot I_{(3\times 3)}$ is *psd (positive semi-definite matrix)*, and

$\begin{bmatrix}\frac{2S_1^2}{\pi_1^3} & 0 & 0\\ 0 & \frac{2S_2^2}{\pi_2^3} & 0\\ 0 & 0 & \frac{2S_3^2}{\pi_3^3}\end{bmatrix}$ is *pd(positive definite matrix)*, so the result we got are minimum! It means that the variance is unbounded.

3.4 Minimize Bias of $\hat{c}_{TA}$ in Term of Spike-in Probe Features

Since the variance is unbounded, we consider minimizing the bias of $\hat{c}_{TA}$ in term of spike-in probe features.

3.4.1 Conditional Bias of $\hat{c}_{TA}$

The Log Langmuir model for Target probe is

$$LogI_{TPi} = \log(\frac{\hat{a}_{TP}C_T}{\hat{b}_{TP} + C_T} + \hat{d}_{TP}) + \log\varepsilon_{TPi}$$

$$= f(\underline{\hat{\theta}}_{Tp}; C_T) + \log\varepsilon_{TPi}$$

where $f(\underline{\hat{\theta}}_{Tp}; C_T) = \log(\frac{\hat{a}_{TP}C_T}{\hat{b}_{TP} + C_T} + \hat{d}_{TP})$.

Since $Var(\log\varepsilon_{TPi} \mid \underline{\hat{\theta}}_T) = \sigma_T^2$ by assumption, so

$$Var(\hat{C}_T \mid \underline{\hat{\theta}}_T) = \sigma_T^2 * (\dot{f}^T \cdot \dot{f})^{-1}$$

(See [3.9])

where

$$\dot{f}(\underline{\hat{\theta}}_T)_{(28\times1)} = \begin{bmatrix} \frac{\partial f}{\partial C_T}(\hat{C}_T, \underline{\hat{\theta}}_{T1}) \\ \vdots \\ \vdots \\ \vdots \\ \frac{\partial f}{\partial C_T}(\hat{C}_T, \underline{\hat{\theta}}_{T28}) \end{bmatrix} = \begin{bmatrix} \dot{f}_1(\underline{\hat{\theta}}_{T1}) \\ \vdots \\ \vdots \\ \vdots \\ \dot{f}_{28}(\underline{\hat{\theta}}_{T28}) \end{bmatrix} = \begin{bmatrix} \frac{a_1 b_1}{a_1\hat{C}_T(b_1 + \hat{C}_T) + d_1(b_1 + \hat{C}_T)^2} \\ \vdots \\ \vdots \\ \vdots \\ \frac{a_{28} b_{28}}{a_{28}\hat{C}_T(b_{28} + \hat{C}_T) + d_{28}(b_{28} + \hat{C}_T)^2} \end{bmatrix}$$

The Hessian matrix of $f(\underline{\hat{\theta}}_{Tp};C_T)$ is:

$$H(\underline{\hat{\theta}}_T)_{(28\times 1)}=\begin{bmatrix}\frac{\partial^2 f}{\partial C_T^2}(\hat{C}_T,\underline{\hat{\theta}}_{T1})\\ \vdots\\ \vdots\\ \vdots\\ \frac{\partial^2 f}{\partial C_T^2}(\hat{C}_T,\underline{\hat{\theta}}_{T28})\end{bmatrix}=\begin{bmatrix}H_1(\underline{\hat{\theta}}_{T1})\\ \vdots\\ \vdots\\ \vdots\\ H_{28}(\underline{\hat{\theta}}_{T28})\end{bmatrix}$$

By using Box's bias formula [Box, 1971]:

$$\begin{aligned}E[(\hat{c}-c_{true})\,|\,\underline{\hat{\theta}}_T]&=-\frac{1}{2}Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T)\cdot\dot{f}^T\cdot\sigma^{-2}\cdot\left[tr(H_1\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\quad\cdots\quad tr(H_{28}\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\right]\\&=-\frac{1}{2}(\dot{f}^T\cdot\dot{f})^{-1}\cdot\dot{f}^T\cdot\left[tr(H_1\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\quad\cdots\quad tr(H_{28}\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\right]\end{aligned}$$

then we take the logarithm:

$$\begin{aligned}&\log\{E[(\hat{c}-c_{true})\,|\,\underline{\hat{\theta}}_T]\}\\&=\log(\dot{f}^T\cdot\dot{f})^{-1}+\log\{-\frac{1}{2}\cdot\dot{f}^T\cdot\left[tr(H_1\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\quad\cdots\quad tr(H_{28}\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\right]\}\end{aligned}$$

Since

$$\log\{-\frac{1}{2}\cdot\dot{f}^T\cdot\left[tr(H_1\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\quad\cdots\quad tr(H_{28}\cdot Var(\hat{C}_T\,|\,\underline{\hat{\theta}}_T))\right]\}<<\log(\dot{f}^T\cdot\dot{f})^{-1}$$

So

$$\begin{aligned}&\log\{E[(\hat{c}-c_{true})\,|\,\underline{\hat{\theta}}_T]\}\approx\log(\dot{f}^T\cdot\dot{f})^{-1}\\&\Rightarrow E[(\hat{c}-c_{true})\,|\,\underline{\hat{\theta}}_T]\approx(\dot{f}^T\cdot\dot{f})^{-1}\end{aligned}$$

Then the bias of $\hat{c}_{TA}$ **is approximately**

$$bias(\hat{c})=E_{\underline{\theta}}\{\;E[(\hat{c}-c_{true})\,|\,\underline{\hat{\theta}}_T]\}$$

Using the $\delta-method$ and [3.13], we obtain

$$bias(\hat{c}) = E(\dot{f}^T \cdot \dot{f}) \approx \sum_{i=1}^{28}[f_i^2(\underline{\theta}_T) + B_i \underline{m}_i \underline{V} \underline{m}_i^T B_i^T]$$

3.4.2 Minimize Bias Respect to Spike-in Probe

Since only $\underline{V}$ depends on spike-in probes, so we have

$$\begin{aligned}\underline{V} &= Var(\underline{\hat{\beta}}) = (\underline{X}_s^T \underline{X}_s)^{-1} \underline{X}_s^T \underline{\Gamma}_s \underline{X}_s (\underline{X}_s^T \underline{X}_s)^{-1} \\ &= \frac{1}{28} \cdot (\underline{\Gamma} \otimes \underline{E}^{-1})\end{aligned} \quad .$$

(see [3.6])

where $\underline{\Gamma} = \begin{bmatrix} \Gamma_{11} & \Gamma_{21} & \Gamma_{31} \\ \Gamma_{21} & \Gamma_{22} & \Gamma_{23} \\ \Gamma_{31} & \Gamma_{32} & \Gamma_{33} \end{bmatrix}$ and

$$\begin{aligned}\underline{E}^{-1} &= \frac{1}{25}[D^{-1}(\underline{\pi}) - \frac{(\sqrt{25-1} \cdot D^{-1}(\underline{\pi}) \cdot \underline{\pi})(\sqrt{25-1} \cdot \underline{\pi}^T \cdot D^{-1}(\underline{\pi}))}{1 + (25-1)\underline{\pi}^T \cdot D^{-1}(\underline{\pi}) \cdot \underline{\pi}}] \\ &= \frac{1}{25}[D^{-1}(\underline{\pi}) - \frac{(25-1) \cdot (\underline{I} \cdot \underline{I}^T)}{1 + (25-1)\sum_{j=1}^{4} \pi_j}] \\ &= \begin{bmatrix} \frac{1}{25\pi_1} & 0 & 0 & 0 \\ 0 & \frac{1}{25\pi_2} & 0 & 0 \\ 0 & 0 & \frac{1}{25\pi_3} & 0 \\ 0 & 0 & 0 & \frac{1}{25\pi_4} \end{bmatrix} - \frac{25-1}{25 * 25} \underline{1}_{(4)} \cdot \underline{1}_{(4)}^T\end{aligned}$$

Where $\underline{D}(\underline{\pi}) = \begin{bmatrix} \pi_1 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 \\ 0 & 0 & \pi_3 & 0 \\ 0 & 0 & 0 & \pi_4 \end{bmatrix}$

Form equation [3.13]:

$$bias(\hat{c}_T) = \sum_{i=1}^{28} f_i^2(\underline{\theta}_T) + \sum_{i=1}^{28} \underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T \ ,$$

where

$$\underline{m}_{i(3x12)} = \underline{X}_{Ti} \bullet \exp(\underline{X}_{Ti} \bullet \underline{\beta})$$
$$= \begin{bmatrix} \underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} & 0 & 0 \\ 0 & \underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} & 0 \\ 0 & 0 & \underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \end{bmatrix}$$

Then combine with equation [3.14], we have

$$\underline{m}_i \underline{V} \underline{m}_i^T = \frac{1}{28} \cdot \underline{\Gamma} \otimes (\underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \cdot \underline{E}^{-1} \cdot \underline{X}_{Ti}^{*T} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}})$$

Let $Q_i = \underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \cdot \underline{E}^{-1} \cdot \underline{X}_{Ti}^{*T} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}}$

Where $\underline{X}_{Ti}^* = [n_{Ti}^A \quad n_{Ti}^C \quad n_{Ti}^G \quad 1]$, n_{Ti}^A, n_{Ti}^C and n_{Ti}^G are the number of A, C and G on the probe i;

$$\underset{3x12}{\underline{X}_{TP}} = \begin{bmatrix} \underline{X}_{Tp}^* & 0 & 0 \\ 0 & \underline{X}_{Tp}^* & 0 \\ 0 & 0 & \underline{X}_{Tp}^* \end{bmatrix} = \begin{bmatrix} \underline{X}_{TP1} \\ \underline{X}_{TP2} \\ \underline{X}_{TP3} \end{bmatrix}.$$

Therefore

$$\underline{m}_i \underline{V} \underline{m}_i^T = \frac{1}{28} \cdot \underline{\Gamma} \otimes Q_i = \frac{1}{28} \cdot \begin{bmatrix} \Gamma_{11} \cdot Q_i & 0 & 0 \\ 0 & \Gamma_{11} \cdot Q_i & 0 \\ 0 & 0 & \Gamma_{11} \cdot Q_i \end{bmatrix}$$

And $\underbrace{B_i}_{1x3} = \frac{\partial f_i}{\partial \underline{\theta}_i} = \begin{bmatrix} \frac{\partial f_i}{\partial a_i} & \frac{\partial f_i}{\partial b_i} & \frac{\partial f_i}{\partial d_i} \end{bmatrix} = \begin{bmatrix} B_{i1} & B_{i2} & B_{i3} \end{bmatrix}$, so

$$\underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T = \frac{Q_i}{28} \cdot \sum_{i=1}^{3} B_{ij}^2 \cdot \Gamma_{jj} = \frac{Q_i}{28} \cdot s_i$$

where $s_i = \sum_{i=1}^{3} B_{ij}^2 \cdot \Gamma_{jj} > 0$.

So the bias is

$$\begin{aligned} bias(\hat{c}_T) &= \sum_{i=1}^{28} f_i^2(\underline{\theta}_T) + \sum_{i=1}^{28} \underline{B}_i \underline{m}_i \underline{V} \underline{m}_i^T \underline{B}_i^T \\ &= \sum_{i=1}^{28} f_i^2(\underline{\theta}_T) + \sum_{i=1}^{28} \frac{Q_i}{28} \cdot s_i \end{aligned}$$

(see [3.16])

Since $\dot{f}_i(\hat{\underline{\theta}}_T) = \dfrac{a_i b_i}{a_i \hat{C}_T (b_i + \hat{C}_T) + d_i (b_i + \hat{C}_T)^2}$ does not depends on spike-in probes

($\underline{\pi}$), only Q_i depends on $\underline{\pi}$, so we can minimize $\sum_{i=1}^{28} \dfrac{Q_i}{28} \cdot s_i$ to minimize $bias(\hat{c}_T)$. Now

$$Q_i = \underline{X}_{Ti}^* \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}} \cdot \underline{E}^{-1} \cdot \underline{X}_{Ti}^{*T} \cdot e^{\underline{X}_{Ti} \bullet \underline{\beta}},$$

where $\underline{X}_{Ti}^* = [n_{Ti}^A \quad n_{Ti}^C \quad n_{Ti}^G \quad 1] = \left[X_{Ti1} \quad X_{Ti2} \quad X_{Ti3} \quad X_{Ti4}\right]$,

substitute [3.15]

$$\underline{B}^{-1} = \begin{bmatrix} \frac{1}{25\pi_1} & 0 & 0 & 0 \\ 0 & \frac{1}{25\pi_2} & 0 & 0 \\ 0 & 0 & \frac{1}{25\pi_3} & 0 \\ 0 & 0 & 0 & \frac{1}{25\pi_4} \end{bmatrix} - \frac{25-1}{25*25} I_{(4x4)}$$

into Q_i, then we have

$$Q_i = \frac{1}{25} \cdot e^{2\underline{X}_{Ti}\underline{\beta}} \left(\sum_{k=1}^{4} \frac{X_{Tik}^2}{\pi_k} - \frac{24}{25} \cdot \sum_{k=1}^{4} X_{Tik}^2\right)$$

Let $R_i = \frac{24}{25*25}\cdot e^{2\underline{X}_{Ti}\underline{\beta}}\sum_{k=1}^{4} X_{Tik}^2$ and $t_{ki}^2 = \frac{1}{25}\cdot(e^{2\underline{X}_{Ti}\underline{\beta}} X_{Tik}^2)$, then

$$Q_i = \sum_{k=1}^{4}\frac{t_{ti}^2}{\pi_k} - R_i$$

Go back to

$$\begin{aligned}\sum_{i=1}^{28}\frac{Q_i}{28}\cdot s_i &= \frac{1}{28}\sum_{i=1}^{28} s_i\Big(\sum_{k=1}^{4}\frac{t_{ti}^2}{\pi_k} - R_i\Big)\\ &= \frac{1}{28}\sum_{i=1}^{28}\sum_{k=1}^{4}\frac{s_i\cdot t_{ki}^2}{\pi_k} - \frac{1}{28}\sum_{i=1}^{28} s_i\cdot R_i\end{aligned}$$

(see [3.17])

Only the first part depends on to spike-in probes ($\underline{\pi}$), so we just consider it.

Let $D_{ki}^2 = s_i\cdot t_{ki}^2$, then

$$U = \frac{1}{28}\sum_{i=1}^{28}\sum_{k=1}^{4}\frac{s_i\cdot t_{ki}^2}{\pi_k} = \frac{1}{28}\sum_{i=1}^{28}\sum_{k=1}^{4}\frac{D_{ki}^2}{\pi_k} = \frac{1}{28}\sum_{i=1}^{28}\Big[\frac{D_{i1}^2}{\pi_1} + \frac{D_{i2}^2}{\pi_2} + \frac{D_{i3}^2}{\pi_3} + \frac{D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)}\Big]$$

In order to minimize U respect to $\underline{\pi}$, we take the partial differentiate of U and set them equal to 0:

$$\frac{\partial U}{\partial \pi_1} = -\frac{\sum_{i=1}^{28} D_{i1}^2}{\pi_1^2} + \frac{\sum_{i=1}^{28} D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (1)$$

$$\frac{\partial U}{\partial \pi_2} = -\frac{\sum_{i=1}^{28} D_{i2}^2}{\pi_2^2} + \frac{\sum_{i=1}^{28} D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (2)$$

$$\frac{\partial U}{\partial \pi_3} = -\frac{\sum_{i=1}^{28} D_{i3}^2}{\pi_3^2} + \frac{\sum_{i=1}^{28} D_{i4}^2}{(1-\pi_1-\pi_2-\pi_3)^2} = 0 \qquad (3)$$

Let $G_j^2 = \sum_{i=1}^{28} D_{ij}^2$, by (1), (2) and (3)

$$\frac{1}{\pi_1^2}G_1^2 = \frac{1}{\pi_2^2}G_2^2 = \frac{1}{\pi_3^2}G_3^2 \Rightarrow \frac{G_1}{\pi_1} = \frac{G_2}{\pi_2} = \frac{G_3}{\pi_3}$$

By (1)$\Rightarrow$

$$\Rightarrow \pi_1 = \frac{G_1}{G_1 + G_2 + G_3 + G_4}$$

By (2), (3)$\Rightarrow$

$$\pi_2 = \frac{G_2}{G_1 + G_2 + G_3 + G_4}$$

$$\pi_3 = \frac{G_3}{G_1 + G_2 + G_3 + G_4}$$

$$\because \quad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

$$\therefore \quad \pi_4 = \frac{G_4}{G_1 + G_2 + G_3 + G_4}$$

To check U is a minimum, we look at the Hessian function:

$$H=\begin{bmatrix} \dfrac{\partial^2 U}{\partial \pi_1^2} & \dfrac{\partial^2 U}{\partial \pi_1 \partial \pi_2} & \dfrac{\partial^2 U}{\partial \pi_1 \partial \pi_3} \\ \dfrac{\partial^2 U}{\partial \pi_2 \partial \pi_1} & \dfrac{\partial^2 U}{\partial \pi_2^2} & \dfrac{\partial^2 U}{\partial \pi_2 \partial \pi_3} \\ \dfrac{\partial^2 U}{\partial \pi_3 \partial \pi_1} & \dfrac{\partial^2 U}{\partial \pi_3 \partial \pi_2} & \dfrac{\partial^2 U}{\partial \pi_3^2} \end{bmatrix}$$

$$=\begin{bmatrix} \dfrac{2G_1^2}{\pi_1^3}+\dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} \\ \dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \dfrac{2G_2^2}{\pi_2^3}+\dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} \\ \dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} & \dfrac{2G_3^2}{\pi_3^3}+\dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3} \end{bmatrix}$$

$$=\frac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\cdot I_{(3\times 3)}+\begin{bmatrix} \dfrac{2G_1^2}{\pi_1^3} & 0 & 0 \\ 0 & \dfrac{2G_2^2}{\pi_2^3} & 0 \\ 0 & 0 & \dfrac{2G_3^2}{\pi_3^3} \end{bmatrix}$$

Since $\dfrac{2G_4^2}{(1-\pi_1-\pi_2-\pi_3)^3}\cdot I_{(3\times 3)}$ is *psd (positive semi-definite matrix)*, and

$$\begin{bmatrix} \dfrac{2G_1^2}{\pi_1^3} & 0 & 0 \\ 0 & \dfrac{2G_2^2}{\pi_2^3} & 0 \\ 0 & 0 & \dfrac{2G_3^2}{\pi_3^3} \end{bmatrix}$$

is *pd(positive definite matrix)*, so the result we obtained is the minimum bias!

CHAPTER 4

SUMMARY AND FUTURE WORK

Microarray technology has been widely used in biological researche and medical studies since its invention in 1995. It allows one to monitor tens of thousands genes, or over all genes in a genome, simultaneously. The absolute mRNA concentration, which is defined as gene expression, can not be measured directly. The focus of this dissertation is using Langmuir adsorption model to estimate the absolute mRNA concentration while the fluorescence intensity is obtained.

In chapter 1 of this dissertation, the biological background of microarray is given, including: how to measure gene expression, construction of microarray and how does a microarray work. The difference of target probes and spike-in probes are mentioned. Heskstra’s ideas are the main point in this chapter, Heskstra’s first idea is that the Langmuir model, which is a model of physical chemistry, can be applied to microarray data analysis, the relationship between the fluorescence intensity and absolute mRNA concentration can be expressed by the Langmuir model. Hekstra used the real Affymetrix data set: HG-U95A to show that the relationship between the fluorescence intensity and absolute mRNA concentration is not linear and follow the Langmuir model, he estimated three Langmuir parameters for each spike-in probe by minimizing the sum of weighted square errors. Hekstra’s second idea is that the probe parameters depend on the probe structure. He proposed a statistical linear model for estimating the probe parameters in term of probe feature, and obtained R^2 of about 50% for each of the three parameters.

There are some methods which estimate concentration by using Langmuir adsorption, especially, Abdueva et al. [2006] used the same model as ours, but they did not use spike-in information. They gave an initial value of concentration to model [1.3], then estimated probe parameters. The concentration estimates are optimized based on those new probe parameters, the iterative scheme continues until converge obtained. The result depends on the starting value of concentration which they chose.

In chapter 2, we proposed our method for estimating absolute concentration when spike-in probes are given. The proposed method is under practical and theoretical assumptions: we assume that the spike-in probes, which sequence and concentration are known and vary across the array for a given experimental condition, are already installed on the array. Hekstra's model

$$\begin{pmatrix} \ln \hat{a}_p \\ \ln \hat{b}_p \\ \ln \hat{d}_p \end{pmatrix} = \begin{pmatrix} \gamma_A^a & \gamma_C^a & \gamma_G^a \\ \gamma_A^b & \gamma_C^b & \gamma_G^b \\ \gamma_A^d & \gamma_C^d & \gamma_G^d \end{pmatrix} * \begin{bmatrix} n_{A,p} \\ n_{C,p} \\ n_{G,p} \end{bmatrix} + \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \qquad [1.3]$$

holds for each probe, Hekstra's empirical model

$$I_{T,p,i} = (\hat{a}_{T,p} \frac{c_T}{c_T + \hat{b}_{T,p}} + \hat{d}_{T,p}) * \varepsilon_{T,p,i} \qquad [2.8]$$

holds with normal error, $\gamma's \,\&\, C's$ in [1.3] are the same for all probes.

Our method is made in 4 steps:

1. Obtain Langmuir parameters of each spike-in probe from model [2.8] by using nonlinear regression.
2. Use Langmuir parameters of spike-in probes to obtain assumed universal $\gamma's \,\&\, C's(\underline{\beta})$ parameters by applying model [1.3].
3. Estimate Langmuir parameters of each target probe from model [1.3] by using assumed universal $\gamma's \,\&\, C's(\underline{\beta})$ parameters and target probe's feature vector.

4 Estimate absolute concentration of target gene by using target Langmuir parameters and model [2.8].

We did a simulation study to check our proposed method by using SAS program. We simulate 100 replicates, in each hypothetical experimental condition, those are:

a) Spike-in probes and target genes,

b) 5 arrays,

c) Assumed universal $\underline{\beta}^{'}s$,

d) Different value of the standard deviation of the noise ($\varepsilon_{s,p,i}$) in [2.8] and the separation between spike-in probes and target probes are used.

Our method works very well based on the estimates, relative bias and variance. The result is best for small value of standard deviation of the noise ($\varepsilon_{s,p,i}$) in [2.8] and small value separation between spike-in probes and target probes.

We tried to find the optimal choice of spike-in probes by assuming that target probes are given, we proceed by the variance of deriving of our absolute concentration estimator in terms of the spike-in probe feature in chapter 3. We minimize the variance of estimator of target absolute concentration, to get the optimal choice of the probability of bite (probability of number of A, T, C and G on the spike-in probe), we minimize the variance in two scenarios:

1. One gene at a time;

2. More than one gene at a time.

Since the variance is unbounded, we tried to minimize the bias of absolute concentration under the given target Langmuir parameters with respect to spike-in probe feature, the optimal choice of the spike-in probe feature is obtained. It is a very useful for the chip design in practices.

REFERENCES

Abdueva, D., Skvortsov, D. and Tavare, S. Non-linear analysis of GeneChip arrays. Nucleic Acids Research, V34, 15, 2006.

Affymetrix Statistical Algorithms Description Document (2001).

Alwine, J. C., Kemp, D. J., and Stark, G. R., Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethy1 and hybridization with DNA probes. Proceedings of the national academy of sciences of the united states of American 74 (12): 5350-5354 1997.

Atkins, P. W., Physical Chemistry, 5th Edn. Oxford University Press, Oxford, UK (1994).

Bergeron, B. Bioinformatics computing. Prentice Hall, 2002.

Binder, H., and Preibisch, S., GeneChip microarrays-signal intensities, RNA concentration and probe sequences, Journal of Physics: Condensed Matter 18, S537 (2006).

Burden, D., Pittelkow, Y., and Wilson S., Adsorption models of hybridization behavior on oligonucleotide microarrays, q-bio. BM/0411005 v2 (2005).

Burden, D., Pittelkow, Y., and Wilson S., Adsorption models of pose-hybridization behavior on oligonucleotide microarrays, q-bio. BM/0411005 v3 (2006).

Burden, D., Pittelkow, Y., and Wilson S., Statistical analysis of adsorption models for oligonucleotide microarrays, Statistical Applications in Genetics and Molecular Biology, v3, 35 (2004).

Dai, H., Meyer, M., Stepaniants, S., Ziman, M., and Stoughton, R., Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays, Nucleic Acid Research 30 e86 (2002).

Duan, F. Analysis of microarray data. PHD thesis, Yale University ,2005.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M., Expression profiling using cDNA microarray. Nature Genetics Supplement, 21:10-14, 1999.

Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D., Light-directed, spatially addressable parallel chemical synthesis. Science 251 (4995): 767-773 1991.

Gautier, L., Cope, L., Bolstad, B. M. and Irizarry, R. A. Affy-analysis of Affymetrix Genechip data at the probe level. Bioinformatics, 20, 307-315 (2004).

Hekstra, D., Taussig, A. R., Magnasco, M and Naef, F., Absolute mRNA concentrations from sequence specific calibration of oligonucleotide arrays, Nucleic Acid Research 31, 1962 (2003).

Irizarry, R A., Hobbs, B., Collin F., Speed, T., Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics, 4, 2, 249-264, (2003).

Jiang, H. A two-step procedure for multiple pairwise comparisons in microarray experiments. PHD thesis, Purdue University, 2004.

Lee, Mei-Ling Ting. Analysis of Microarray Gene Expression Data. Kluwer Academic Publishers, 2004.

Lewin, B., Genes VI, Oxford: Oxford University Press, 1997.

Li, C and Wong, W. H. Model based analysis of oligonucleotide arrays: expression index computation and outlier detection, Proc. Natl Acad. Sci, 2001, 98

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

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., Expression of monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14, 1675 (1996).

M. J. Box, Bias in Nonlinear Estimation, Journal of the Royal Statistical Society. Vol. 33, No. 2. (1971), 171-201.

Nelson, B. P., Grimsrud, M. R., Liles, M. R., Goodman, R. M., Corn, R. M., Surface plasmon resonance imaging measurements of DNA and RNA Hybridization-Adsorption onto DNA microarrays, Analytical Chemistry 73, 1 (2001).

Pease, A. C., Solas, D. E., Sullivan, J., Cronin, M. T., Holmes, C. P., Fodor, S. P. A., Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proceedings of the national academy of sciences of the united states of American 91 (11): 5022-5026 1994.

Peterson, A. W., Heaton, R. J., and Georgiadis, R. M., The effect of surface probe density on DNA hybridization, Nucleic Acid Research 29, 5163 (2001).

Peterson, A. W., Wolf, L. K., and Georgiadis, R. M., Hybridization of mismatched or partially matched DNA at surfaces, Journal of the American Chemical Society 124, 14601 (2002).

Rao, C R, Linear Statistical Inference and Its Applications, John Wiley & Sons, 1965.

Sambrook, J. D., Russell, W., Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory , 3 edition, 2001.

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

Seber, G.A.F, Wild, C.J, Nonlinear Regression, Wiley Interscience, 1989.

Southern, Detection of specific sequence among DNA fragments separated by gel eletrophoresis. Journal of molecular biology 98 (3):503 & 1975.

Watson, J. D., and Crick, F. H. C., Molecular structure of nucleic acids. Nature, 171:737-738, 1953.

Watson, J. D., The Double Helix: A Personal Account of the Discovery of the Structure of DNA. Weidenfeld & Nicolson, 1997.

Wu Z., Irizarry R.A., Gentleman R., Martinez Murillo F., and Spencer F. 2004. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 1.

Wu, Z., Irizarry, R A., Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays, Journal of Computational Biology. 2005, 12(6): 882-893.

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

Zhang, Y., Ferreira, A., Cheng, C, and Wu, Y., Modeling oligonucleotide microarray signals, Applied Bioinformatics, 5, 153-160, (2006).

BIOGRAPHICAL INFORMATION

Min Mo received her Bachelor in Physics from GuangXi Normal University (China) in 1997. She has been a graduate student at University of Texas at Arlington since 2002, and obtained her Master of Mathematical Statistics in 2004.

Min Mo is interested in Biostatistics, especially in clinical trail analysis, she has been involved in several clinical trail projects. She will contribute to this field in the future.