Extended investigations in the physics of oligonucleotide microarrays

K. Nguyen

Mathematical Sciences Institute, Building 27, The Australian National University, Acton ACT 0200, Canberra, Australia

Gene expression technology is commonly used in molecular biology and medicine to evaluate parallel expression levels of a large number of genes. The technology relies on the hybridisation of fluorescently labelled DNA/RNA fragments to oligonucleotide probes. Existing algorithms available to biologists for converting fluorescence intensities to target concentrations are statistically sophisticated, but fail to take proper account of the physical chemistry involved in the experiment. Recently Burden has developed a physico-chemical model which has been tested on the U95 and U133 Affymetrix Latin Square spike-in data sets. The model fits well to the U95 data set but fails to fit accurately to the U133 data set at low spike-in concentrations. In order to remedy this, the current study considers two new physical effects not included in the original model, namely incomplete probe synthesis and target depletion. In general, we find that including incomplete probe synthesis provides little improvement. On the other hand, statistical test shows that inclusion of target depletion improves the fit, though some problems remain.

1. INTRODUCTION

DNA is a polymeric molecule which contains genetic instructions for the development and functioning of all known living cells. A gene is a region on the DNA and encodes for a specific protein. The information on the gene is transcribed to an RNA molecule *via* a process called *transcription*. The RNA molecule is subsequently involved in the synthesis of the corresponding protein [1].

A microarray consists of tens of thousands of microscopic spots of oligonucleotides arranged in an arrayed series. Each spot is called a *feature* and contains a specific short sequence of DNA chosen to match part of a human gene. Each expressed gene is present in a cell as its complementary RNA sequence, called *target*, in concentrations of the order of picomoles per litre [2]. The experiment consists of two main steps: *hybridization* between the targets and the probes and the *detection* of the specific sequences by some form of labeling techniques. The probes are supported on a solid surface made of glass or silicon chip by covalent interactions with a chemical matrix. Microarrays originated from Southern blotting, where DNA fragments are attached to a substrate, and then duplexed with a known gene [3, 4]. The processes can also be used for the detection of RNA as cDNA obtained after reverse transcription.

They have many applications in biological sciences. For instance, they are used in comparative genomic hybridization where the gene contents from different cells or organisms are compared to examine the evolutionary relationship [5] or in the identification of single nucleotide polymorphisms that may cause genetic defects including potentially cancer-causing mutations, known as *SNP microarrays* [6]. They are also useful in the detection of the protein binding site (active site) in protein binding genes, known as *Chromatin immunoprecipitation* (ChIP) on Chip [7]. The oligonucleotide microarrays considered in this study are involved in one of the most significant applications, *gene expression profiling* where they allow the simultaneous expression of a large number of genes in prepared target RNA samples to be evaluated [8]. This is particularly useful in studies of diseases by the comparison of gene expression levels between abnormal and normal cells.

The oligonucleotide microarray experiment involves the hybridization between the target RNA and the oligonucleotide sequences on the microarrays *via* complementary base pairing. The aim is to find the gene expression level from the concentration of the target RNA in the loaded solution, which is related to the density of hybridized probe-target duplexes formed. Different labeling techniques applied to the target RNA such as fluorescent dye, radionucleotide markers and quantum dots allow the detection of the duplexes

to be made *via* some intensity measurements of the observed signals.

There are a number of designs for DNA microarrays. *Affymetrix GeneChip* arrays are the subject of this study [10]. An array of this type contains a probeset of 11-16 pairs of features, whose lengths are all 25 bases. Each pair contains an element with a perfect match sequence (PM) and another element with a mismatch sequence (MM), which is the same as the PM but the middle 13th base being replaced by its complementary base. The PM sequences are selected so that they are non-overlapping subsequences of the target gene and based on their ability to specifically hybridize with the gene. Fluorescent dye is used to label the target cRNA molecules.

Given the intensity measurements from the experiment, an efficient functional relationship based on logical considerations of physical chemistry principles relating them and the absolute target RNA concentration is necessary instead of the empirical approach involving the use of expression indices obtained by the subtraction of the PM intensity by the background data estimated by the MM intensity. The *Langmuir* adsorption theory was initially proved to be a reasonable model for the description of surface hybridization processes [11, 12]. The probe-target hybridization is, however, influenced by a number of other factors such as thermodynamic stability of the duplexes, surface chemical adsorption, side reactions in bulk solution, electrostatics of the fragmentation of the DNA, washing step etc [13, 14, 15]. Consequently, the initial *Langmuir* adsorption model has to be modified by taking these factors into considerations.

2. STATEMENT OF THE PROBLEM

A physico-chemical model of the functioning of oligonucleotide microarrays was previously developed by considering a set of hybridisation reactions at the microarray surface, in the supernatant solution (Eqs. 1-5) and partial dissociation of hybrid duplexes during post-hybridisation washing phase (Eq. 6) [12, 14, 16].

Fig. 1: Typical fits to observed fluorescence intensities in spike-in U95 experiments.

According to the model, the duplex coverage fraction q for each feature can be described in terms of the specific target concentration x at washing time t_w as

$$q(x, t_w) = a(t_w) + b(t_w) \frac{Kx}{1 + Kx}, \quad (7)$$

where the parameters a , b and K are dependent on binding energies of probe-target duplexes; target-target duplexes; folded probes and targets; and non-specific target concentrations. The model predicts a hyperbolic response function for the measured fluorescence intensities.

$$I(x) = a + bq(x, t_w) = A + B \frac{Kx}{1 + Kx}, \quad (8)$$

where a and b are constant across the microarray, and $A = a + ba$, $B = bb$.

The Affymetrix U95 and U133 spike-in experiments were used to examine the model, which allowed A , B and K to be estimated for each feature of each spiked-in gene. The model fits particularly well in the datasets of U95 with and without the complex human pancreas background (Fig. 1) (Table 1 from ref. [16]). On the other hand, many examples from the U133 dataset displayed a sigmoidal rather than hyperbolic shape (Fig. 2). The aim of this study is to improve the current chemical adsorption model for U133 by

considering two possible physical effects: incomplete synthesis of the oligonucleotide probes [13] and target depletion [16].

Fig. 2: An example of the sigmoidal behavior for the

fluorescence intensities in spike-in U133 experiments.

Sections 3 and 4 describe how each of these effects will change the current model and evaluate the validity of the two resulting physico-chemical models by considering their behaviors asymptotically and their fits to the dataset.

3. INCOMPLETE SYNTHESIS OF OLIGONUCLEOTIDE PROBES

Oligonucleotide microarrays are produced *via* photolithography, which involves the use of light and light masking agent adding one nucleotide at a time on the whole array in a selective manner [9]. The “mask” is only added to the end of the growing probes at certain locations on the array each time. The chemically unreactive protecting groups on these applicable probes are removed prior to exposure to a solution of a specific nucleoside (A, T, G or C). This process is called “unmasking”. The deprotected end of the probe can undergo reaction to incorporate this next nucleotide to the chain. The masking reaction then takes place again, the next set of selected probes are unmasked and coupled to the next nucleotide. The entire process is repeated until the probes reach the desired length with specific sequences.

The model described in Section 2 assumes that all these reactions are 100% efficient, which is, in

reality, not the case [18]. The yield of the deprotection process was reported previously to be $p_{synt} = 82 - 98\%$ [13]. Those probes at which the “unmasking” process fails can no longer be extended. Consequently, it is expected that the microarrays contain probes with various lengths.

Assuming that L be the ideal length of the probe and p_{synt} be the synthetic yield of each nucleotide on the probe. L for the Spike-in experiments is 25. The probability distribution of the length of the probe, a , is

$$f_{synt}(a, L) = \begin{cases} p_{synt}^{a-1}(1 - p_{synt}) & \text{for } a = 1, \dots, L - 1 \\ p_{synt}^{a-1} & \text{for } a = L \end{cases} \quad (9)$$

where $0 \leq p_{synt} \leq 1$. When $p_{synt} = 1$, all the probes are L in length.

From this distribution, we aim to find an expression for the overall surface coverage of the microarray, q . Section 3.1 will show the derivation of a simple model without probe folding on the microarray surface. The model is further extended with the effects of probe folding in Section 3.2. Sections 3.3 and 3.4 investigate in the behavior of the model.

3.1 Initial simplified model

For an initial simplistic model, the two main chemical reactions at the microarray surface are considered for probes with a definite length a . Specific hybridization:

$$P_a + S \rightleftharpoons P_a \cdot S \quad \text{with } K_a^S = \frac{[P_a \cdot S]}{[S] \cdot [P_a]}. \quad (10)$$

Non-specific hybridization:

$$P_a + NS_i \rightleftharpoons P_a \cdot NS_i \quad \text{with } K_a^{NS_i} = \frac{[P_a \cdot NS_i]}{[NS_i] \cdot [P_a]}, \quad (11)$$

where the species NS_i is any target RNA molecule with a subsequence complementary to the i th subsequence of the specific target sequence. Let the coverage for probes with a length a be q_a . This is the sum of the coverage of these probes by both specific and non-specific target mRNA, expressed as:

$$q_a = q_a^S + \sum_i q_a^{NS_i}. \quad (12)$$

The total concentration of probes of length a is

$$[P_a]_{total} = [P_a] + [P_a \cdot S] + [P_a \cdot NS_i]. \quad (13)$$

Therefore,

$$\begin{aligned} \frac{[P_a]}{[P_a]_{total}} &= 1 - q_a, \\ \frac{[P_a \cdot S]}{[P_a]_{total}} &= q_a^S, \\ \frac{[P_a \cdot NS_i]}{[P_a]_{total}} &= q_a^{NS_i}. \end{aligned} \quad (14)$$

From (10), (11) and (14) one obtains after some working

$$\begin{aligned} q_a^S &= K_a^S \cdot [S] \cdot (1 - q_a), \\ q_a^{NS_i} &= K_a^{NS_i} \cdot [NS_i] \cdot (1 - q_a). \end{aligned} \quad (15)$$

Substituting (15) into (12) to obtain an expression for q_a in terms of $[S]$ and $\mathring{a}_i [NS_i]$.

$$q_a = \frac{X_a}{1 + X_a}, \quad (16)$$

where

$$\begin{aligned} X_a &= X_a^S + X_a^{NS} \\ \text{with } X_a^S &= K_a^S [S] \\ \text{and } X_a^{NS} &= \mathring{a}_i K_a^{NS_i} [NS_i]. \end{aligned} \quad (17)$$

The overall hybridization isotherm can be expressed as

$$q = \mathring{a}_{a=1}^L f_{synt}(a, L) \cdot q_a. \quad (18)$$

Probe folding, target-target hybridisation in bulk solution and the post-hybridisation washing step have not been taken into account. Hybridisation in bulk solution has the effect of reducing spike-in concentration x to a free target concentration, $[S] = x / (1 + K^{S \cdot fold} + \mathring{a}_i K_i^{bulk} [S_i^{bulk}])$, which is common for all probes on the microarrays. The washing step will scale each q_a by a factor that reduces the asymptote of each q_a to less than 1, but does not otherwise change the shape of the function.

3.2 A more complex model with probe folding

In addition to the two reactions described by (10) and (11), probe folding is considered.

$$P_a \rightleftharpoons P_a^{\phi} \quad \text{with} \quad K_a^{Pfold} = \frac{[P_a^{\phi}]}{[P_a]}. \quad (19)$$

This gives

$$\begin{aligned} q_a^S &= \frac{X_a^S \cdot (1 - q_a)}{1 + K_a^{Pfold}}, \\ q_a^{NS} &= \mathring{a}_i q_a^{NS_i} = \frac{X_a^{NS} \cdot (1 - q_a)}{1 + K_a^{Pfold}}. \end{aligned} \quad (20)$$

Let m be the constant $1/(1 + K_a^{Pfold})$. The hybridization isotherm for probes with length a is

$$q_a = \frac{m X_a}{1 + m X_a}. \quad (21)$$

The overall q is still expressed in terms of these q_a 's as in Eq. (18). The consideration of probe folding in this model does not introduce any extra parameter other than a reducing factor m to the binding strength in the expression for q_a . As a result, the simplified model built in section 3.1 is used for assessing whether the case of incomplete synthesis will give the sigmoidal behavior of the total coverage at low specific target concentrations as observed in the plots of Fig. 2. This is covered in section 3.3.

3.3 Prediction of the behavior of the model

The model for q described in Eq. (18) is a superposition of the individual isotherms of probes of length a . As suggested in Eq. (16), each contributing isotherm behaves in a *Langmuir* manner. It is expected that the binding strength increases with probe lengths, suggesting that at low target concentrations, longer probes hybridize more efficiently than shorter ones [13]. Therefore, q_a values for larger a 's contribute a much more significant portion to the mean value q than those of smaller a at low $[S]$. The resultant shape of the total coverage is predicted to be similar to that of the coverage of longer probes. This indicates that q will be likely to still show a *Langmuir*-type behavior and hence will not give the desired sigmoidal shape at low target mRNA concentrations.

3.4 Qualitative analysis of the model

The model described in Eq. (18) can be fully written as

$$q = \prod_{a=1}^L f_{\text{synt}}(a, L) \cdot \frac{X_a}{1 + X_a}, \quad (22)$$

where $p_{\text{synt}}(a, L)$ is defined in Eq. (9) and only dependent on the synthetic yield p_{synt} for each nucleotide on the probe and the ideal length of the probe. For this study, $L = 25$.

We aim to investigate how the relationship between q and the target concentration will vary upon changing the extra parameter, p_{synt} . It is necessary to firstly express X_a in terms of $[S]$. X_a is the sum of X_a^S and X_a^{NS} . However, as the model involving target depletion does not take non-specific hybridization into account, X_a^{NS} is not considered for this model for consistency. On the other hand, X_a^S was defined as $K_a^S[S]$. The specific equilibrium constant K_a^S needs to be evaluated. From ref. [16], the constant can be expressed as

$$K_a^S = e^{I_S(m_S - \frac{DG(a)}{RT})}, \quad (23)$$

where the parameters I_S and m_S have been estimated as 0.00677 and -688.9 respectively in that study.

The main difference between our model and that in the previous study is that the effective binding free energy DG is no longer a constant, but depends on the length of the probe, a . Based on physical study done by Sugimoto *et al.* [20] together with a few steps of manipulations, the dimensionless relationship between DG and a is approximated as

$$\frac{DG(a)}{RT} = 5.478 + (-2.274) \cdot a. \quad (24)$$

Having estimated all the involved parameters, q can then be plotted in terms of $[S]$ for different p_{synt} values as shown in Fig. 3. The plot of q does not deviate strongly from the hyperbolic shape of the *Langmuir*-isotherm model. The sigmoid behavior at low target concentration can not be obtained. In addition, the curve for q shifts downwards by a small amount with decreasing p_{synt} , showing that the behavior of q does not change significantly upon changing p_{synt} .

Fig. 3: Plot of the overall coverage of the microarrays against the mRNA target concentration $[S]$ for various values of p_{synt} range from 0.5 (bottom curve) to 1.0 (top curve).

4. TARGET DEPLETION

Another physical effect that may cause the deviation of the hybridization isotherms of microarrays from a *Langmuir*-type behavior is target depletion. In the previous derivation of q [16], when considering chemical reactions at the microarray surface, it was assumed that the spike-in concentration of specific target mRNA was sufficiently large compared to that of bound target mRNA in forms of $P.S$ duplexes. As a consequence, the concentration of free targets, $[S]$ was treated as if it were a constant and the same as the total spike-in concentration. Nonetheless, at low spike-in concentrations, target depletion by $P.S$ duplex formation is likely to cause the actual concentration of free target, $[S]$ to deviate from that of the spike-in concentration by a greater proportion. The previously made assumption no longer holds and $[S]$ needs to be considered as a function of both the spike-in target concentration and duplex concentration $[P.S]$ [21]. This section aims to construct a model resulting from the effects of target depletion.

4.1 The initial simplified model

In this model, only the chemical reaction for probe-target binding is considered.

$$P + S \rightleftharpoons P.S, \quad \text{with } K^S = \frac{[P.S]}{[S].[P]}. \quad (25)$$

Let P_{total} and x be the total probe concentration and the target mRNA spike-in concentration respectively. In this case, P_{total} is the sum of the concentrations of free probes, $[P]$ and duplexed probes, $[P.S]$. On the other hand, x is the sum of the free-target concentration, $[S]$ and duplexed target, $[P.S]$. Explicitly,

$$P_{total} = [P] + [P.S], \quad (26)$$

$$x = [S] + [P.S]. \quad (27)$$

From (26), the relationship between q and the probe concentrations can be obtained.

$$q = \frac{[P.S]}{P_{total}}, \quad 1 - q = \frac{[P]}{P_{total}}. \quad (28)$$

Hence, one obtains the isotherm

$$q = K^S (1 - q)(x - qI), \quad \text{where } I = P_{total}. \quad (29)$$

An explicit expression of q can then be obtained by solving the quadratic equation (29).

$$q(x) = \frac{x + I + \frac{1}{K^S} - \sqrt{(x + I + \frac{1}{K^S})^2 - 4Ix}}{2I}. \quad (30)$$

There are, in fact, two solutions to the obtained solution. The asymptotic behaviour of q eliminates the other solution. This is the first simplest model of q with target depletion. It involves only one extra parameter I compared to the *Langmuir*-type model. The simple model of Eq. (30) will be used for computational purposes to enable comparisons between fits to the U33 data with and without the assumption of target depletion in Section 4.4. A more complex version of the model including reactions in bulk solution and non-specific hybridization is given in the Appendix.

4.2 Prediction of the behaviour of the model

The simple model described in Eq. (30) will be firstly used for the purpose of prediction and further computational work. There are three basic tests that can be applied to the model.

- When there is no target mRNA in the reaction mixture, $q(x = 0) = 0$.

- As I approaches 0, the model will be expected to return the hyperbolic function with *Langmuir*-type behaviour as before the effects of target depletion were introduced.

$$\begin{aligned} q(x) &= \frac{x + I + \frac{1}{K^S} - \sqrt{(x + I + \frac{1}{K^S})^2 - 4Ix}}{2I} \\ &= \frac{x + I + \frac{1}{K^S} - \frac{2Ix + I + \frac{1}{K^S}}{2I}}{2I} \\ &= \frac{x}{x + \frac{1}{K^S}} + O(I^2) \\ &\approx \frac{K^S x}{1 + K^S x} \end{aligned} \quad (31)$$

- The slope of q at $x = 0$,

$$\begin{aligned} \left. \frac{dq}{dx} \right|_{x=0} &= \frac{1}{2I} \left[1 - \frac{2(x + I + \frac{1}{K^S}) - 4I}{2[(x + I + \frac{1}{K^S})^2 - 4Ix]^{\frac{1}{2}}} \right] \Big|_{x=0} \\ &= \frac{K^S}{1 + K^S I}. \end{aligned} \quad (32)$$

The slope at $x = 0$ is smaller than K when $I > 0$, which may give a behaviour closer to the sigmoidal behaviour observed for the U133 dataset.

Initial investigations carried out on the model suggest that the model has some possibility of fitting in the U133 intensity data. The next section will examine the model qualitatively.

4.3 Qualitative analysis of the model

To analyse the qualitative behaviour of the family of isotherm solutions Eq. (30), we introduce a dimensionless concentration $X = Kx$ and dimensionless parameter $L = K/I$, giving

$$q = \frac{X + L + 1 - \sqrt{(X + L + 1)^2 - 4LX}}{2L}, \quad (33)$$

where the dimensionless quantities, X and L are Kx and Kl respectively.

Fig. 4: Plot of the overall coverage of the microarrays against X for various values of L range from 0 to 15.

The results of plotting q against X for different L parameter values are shown in Fig. 4. At $L = 0$, q is identical with that of the *Langmuir*-isotherm model (see Eq. (33)). A more significant downward shift of the curve for q with increasing L is observed compared to the previous q plot with changing p_{syn} values. The curve appears more linear rather than sigmoidal as L increases.

4.4 Statistical Analysis for the model with target depletion

To examine the statistical significance of the model described in Eq. (30), we consider the relationship between the coverage q and the fluorescence intensity at a feature for a given concentration, x of a mRNA transcript

$$I(x) = A + Bq(x), \quad (34)$$

where A is the background intensity resulting from sources such as reflection, components of non-specific hybridization etc. and B corresponds to the maximum fluorescence by the hybridized features on the microarray [11].

Our aim is to make a comparison between the original *Langmuir* model (Model 1) and the model with target depletion (Model 2) via a χ^2 -squared test of the scaled change in deviance as described

in ref. [12], where model 1 is nested within model 2. The test will ultimately illustrate whether the extra parameter l improves the fits significantly.

$$I(x) = A + B \cdot \frac{K^s x}{1 + K^s x} \quad (\text{Model 1}), \quad (35)$$

$$I(x) = A + B \cdot \frac{x + l + \frac{1}{K^s} - \sqrt{(x + l + \frac{1}{K^s})^2 - 4lx}}{2l} \quad (\text{Model 2}). \quad (36)$$

Let r_1 and r_2 be the residual degrees of freedom, D_1 and D_2 be the deviances of models 1 and 2 respectively. The tested object is given in ref. [12] as

$$DD_{\text{scaled}} = (D_1 - D_2) \frac{r_2}{D_2}. \quad (37)$$

Throughout the process of fitting the two generalized linear models to the U133 dataset of Affymetrix Latin Square experiment, it is assumed that the measured fluorescence intensity I is a gamma random variable with mean m and shape parameter n . The probability density and unscaled deviance are given by McCullagh and Nelder [19] respectively as follows

$$\text{Probability density} = \frac{1}{\Gamma(n)} \left(\frac{n}{m} \right)^n e^{-\frac{n}{m} I} I^{n-1} \quad (38)$$

$$D(I; m) = -2 \ln \left(\frac{1}{\Gamma(n)} \left(\frac{n}{m} \right)^n e^{-\frac{n}{m} I} I^{n-1} \right) \quad (39)$$

A number of defective probes and outliers from the U133 dataset are omitted from the analysis (see Table 1). The results of fitting the two generalised linear models to the the U133 data set are shown in Fig. 2.

Gene	Data omitted
203508_at	Probe 2
207777_s	Probe 10
207968_s	Probe 6
207968_s	Probe 7
209606_	Datapoint # 2, MM, Probe 9
206060_s	Datapoint # 8, PM, Probe 31

Table 1: A list of the omitted data from the analysis of both models.

The results of the fitting the two generalized linear models to the U133 dataset can be summarized as followed

	Model 1	Model 2
Total deviance, D	391.3462	377.5971
Degree of freedom	18174	17708

Table 2: The total deviances and degrees of freedom from fitting the two models to the U133 dataset.

The DD_{scaled} is found to be 647.2146, which gives a p -value of $5.0 \cdot 10^{-8}$. This small p -value indicates statistically that the introduction of the extra parameter makes a significantly better fit to the dataset.

The plots of the fits of the fluorescence intensity data to the response curve given by the new model are also done for the dataset of 38 genes. There exist some abnormalities in a number of the response curves where at some points in the high target concentration range, a sharp-cornered transition is observed (Fig. 5). This indicates a discontinuity in the gradient of the curve. A physical explanation for this behavior is required. A suggested idea is the limits imposed on the values of the parameters, I and K in the programming process.

5. CONCLUSION

In conclusion, the study has considered two possible physical effects suggested in ref. [16] for the aim of finding a physical explanation for the deviation of the response curve for Affymetrix U133 dataset from the hyperbolic form. Our investigation suggests that the incomplete synthesis of probes does not significantly alter the hyperbolic behavior. The case of target depletion has positive results from the initial statistical test. Further work is required to find a physically meaningful explanation on the abnormalities in the plots obtained.

Fig. 5: Examples of the mentioned abnormal behavior of the fits.

6. ACKNOWLEDGEMENT

Special thanks to my supervisor, Dr Conrad Burden, for all his extremely invaluable help with the project and editing the paper. This work was as part of the Advanced Studies Course series in the Bachelor of Philosophy program.

APPENDIX

More complex model with reactions in bulk solution

The chemical reactions in bulk solutions arise by two main types of interactions:

- Intramolecular folding of the specific target mRNA molecules (S-fold), which disables duplex formation for a portion of the target molecules. (ref. [13])

$$S \rightleftharpoons S\phi \quad \text{with } K^{Sfold} = \frac{[S\phi]}{[S]}. \quad (40)$$

- Non-specific hybridization which involves the duplex formation between between the specific mRNA targets and the non-specific RNA fragments present in the bulk solution.

$$S + NS_i \rightleftharpoons S.NS_i \quad \text{with } K_i^{bulk} = \frac{[S.NS_i]}{[NS_i].[S]}. \quad (41)$$

At the microarray surface, not only specific hybridization (as in (25)) but also non-specific hybridisation are considered. The additional reaction can be expressed as

$$P + NS_i \rightleftharpoons P.NS_i \quad \text{with } K^{NS_i} = \frac{[P.NS_i]}{[NS_i].[P]}. \quad (42)$$

The total spike-in concentration of the target mRNA, x can be rewritten as

$$x = [S] + [P.S] + \sum_i \mathring{a}_i [NS_i] + [S\phi]. \quad (43)$$

From (40), (41) and (43),

$$\begin{aligned} x - [P.S] &= [S] + \sum_i \mathring{a}_i K_i^{bulk} [NS_i] [S] + K^{Sfold} [S] \\ x - [P.S] &= [S] (1 + \sum_i \mathring{a}_i K_i^{bulk} [NS_i] + K^{Sfold}). \end{aligned} \quad (44)$$

Hence,

$$[S] = \frac{x - [P.S]}{(1 + X^{bulk}[NS] + K^{Sfold})}, \quad (45)$$

$$\text{with } X^{bulk}[NS] = \sum_i \mathring{a}_i K_i^{bulk} [NS_i].$$

The total probe coverage consists of two components, specific and non-specific coverages.

$$\begin{aligned} q &= q^S + q^{NS} \\ &= \frac{[P.S]}{P_{total}} + \sum_i \mathring{a}_i \frac{[P.NS_i]}{P_{total}} \\ &= K^S [S] \frac{[P]}{P_{total}} + \sum_i \mathring{a}_i K_i^{NS} [NS_i] \frac{[P]}{P_{total}} \\ &= (1 - q)(K^S [S] + \sum_i \mathring{a}_i K_i^{NS} [NS_i]) \\ &= (1 - q)(K^S [S] + X^{P-NS}). \end{aligned} \quad (46)$$

The aim is to find an expression of $[S]$ in terms of q and other constants.

$$\begin{aligned} [S] &= x - [P.S] \\ &= x - q^S P_{total} \\ &= x - (q - q^{NS}) P_{total} \\ &= x - q P_{total} + P_{total} (1 - q) X^{NS}. \end{aligned} \quad (47)$$

Substituting (47) into (46) to obtain a quadratic equation for q

$$a q^2 + b q + c = 0, \quad (48)$$

with

$$a = \frac{K^S}{g} P_{total} (1 + X^{NS}) \quad (49)$$

$$b = -\left[\frac{K^S}{g} P_{total} (1 + X^{NS} + x + P_{total} X^{NS}) + X^{P-NS} + 1 \right] \quad (50)$$

$$c = \frac{K^S}{g} (x + P_{total} X^{NS}) + X^{P-NS} \quad (51)$$

$$g = 1 + K^{Sfold} + X^{bulk}. \quad (52)$$

q can be written in terms of a , b and c as a solution to the quadratic equation (48). Similar to in the previous case, asymptotic behaviour of q allows one of the solutions to be eliminated

REFERENCES

- [1] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 4th edition, W. H. Freeman, New York, pp. 89-92, 2005
- [2] C. J. Burden, Y.E. Pittelkow and S. R. Wilson, *Statistical analysis and physical modelling of oligonucleotide microarrays*. In I. A. Deutsch, L. Brusch, G. Byrne, de Vries and H.-P. Herzel editors, *Mathematical Modelling of Biological Systems*, vol. I, pages 333-346, Birkhäuser, Boston, MA, USA, 2007
- [3] E. M. Southern, *J. Mol. Biol.*, 98:503-517, 1975
- [4] D. A. Kulesh, D. R. Clive, D. S. Zarlenga and J. J. Greene, *Proc. Natl. Acad. Sci. USA*, 84:8453-8457, 1987
- [5] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov and C. F. Williams, *Nat. Genet.*, 23:41-46
- [6] J. G. Hacia, J. B. Fan, O. Ryder, L. Jin, K. Edgemon, G. Ghandour, R. A. Mayer, B. Sun, L. Hsie, C. M. Robbins, L. C. Brody, D. Wang, E. S. Lander, R. Lipshutz, S. P. Fodor and F. S. Collins, *Nat. Genet.* 22:164-167, 1999
- [7] M. J. Buck and J. D. Lieb, *Genomics*, 83(3):349, 2004
- [8] M. Schena, D. Shalon, R. W. Davis and P. O. Brown, *Science*, 270:467-470, 1995
- [9] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes and S. P. Fodor, *PNAS* 91:5022-5026, 1994
- [10] D. V. Nguyen, A. B. Arpat, N. Wand and R. J. Carroll, *Biometrics*, 58:701-717, 2002
- [11] D. Hekstra, A. R. Taussig, M. Magnasco and F. Naef, *Nucleic Acids Res.*, 31:1962-1968, 2003
- [12] C. J. Burden, Y. E. Pittelkow and S. R. Wilson, *Statistical Applications in Genetics and Molecular Biology*, 3:35, 2004
- [13] H. Binder, *J. Phys. Condens. Matter*, 18:S491-S523, 2006
- [14] C. J. Burden, Y. Pittelkow and S. R. Wilson, *J. Phys. Condens. Matter*, 18:5545-5565, 2006
- [15] T. Heim, L.-C. Tranchevent, E. Carlon and G. T. Barkema, *J. Phys. Chem. B*, 110:22786-22795, 2006
- [16] C. J. Burden, *Phys. Biol.*, 5:016004, 2008
- [17] Online:
<http://www.affymetrix.com/support/technical/sample_data/datasets.affx>
- [18] J. E. Forman, I. D. Walton, D. Stern, R. P. Rava and M. O. Trulson, *Molecular Modeling of Nucleic acids, ACS Symposium Series*, 682:206-228, 1998
- [19] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman and Hal, London, UK, 2nd edition, 1989
- [20] N. Sugimoto, S. Nakano, M. Yoneyama and K. Honda, *Nucleic Acids Res.*, 24:4501-4505, 1996
- [21] N. Ono, S. Suzuki, C. Furusawa, T. Agata, A. Kashiwagi, H. Shimizu and T. Yomo, *Bioinformatics*, 24:1278-1285, 2008