Modeling and Estimation for Real-Time Microarrays *

HARIS VIKALO[a] BABAK HASSIBI[b] ARJANG HASSIBI[a]

[a]ECE Department, The University of Texas, Austin, TX 78701
[b]EE Department, California Institute of Technology, Pasadena, CA 91125
e-mail: hvikalo,arjang@mail.utexas.edu,hassibi@caltech.edu

Abstract

Microarrays are used for collecting information about a large number of different genomic particles simultaneously. Conventional fluorescent-based microarrays acquire data after the hybridization phase. During this phase, the targets analytes (e.g., DNA fragments) bind to the capturing probes on the array and, by the end of it, supposedly reach a steady state. Therefore, conventional microarrays attempt to detect and quantify the targets with a single data point taken in the steady-state. On the other hand, a novel technique, the so-called real-time microarray, capable of recording the kinetics of hybridization in fluorescent-based microarrays has recently been proposed. The richness of the information obtained therein promises higher signal-to-noise ratio, smaller estimation error, and broader assay detection dynamic range compared to conventional microarrays. In the current paper, we study the signal processing aspects of the real-time microarray system design. In particular, we develop a probabilistic model for real-time microarrays and describe a procedure for the estimation of target amounts therein. Moreover, leveraging on system identification ideas, we propose a novel technique for the elimination of cross-hybridization. These are important steps toward developing optimal detection algorithms for real-time microarrays, and to understanding their fundamental limitations.

Index Terms:

DNA microarrays, real-time, cross-hybridization, statistical modeling

*This work was supported in part by a Grubstake Award from California Institute of Technology, a grant from the David and Lucille Packard Foundation, and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech.

1 Introduction

DNA microarrays [1]-[8] are affinity-based biosensors capable of testing tens of thousands of different genes simultaneously. Sensing in DNA microarrays is based on hybridization, a chemical processes in which single DNA strands bind to each other creating structures in lower energy states. Typically, the surface of a DNA microarray comprises an array of spots, each spot containing a large number of identical single-stranded DNA sequences (*probes*) designed to capture DNA molecules (*targets*) of interest. Microarrays are often used to measure gene expression levels, i.e., to quantify the process of transcription of DNA information into messenger RNA molecules (mRNA). The information transcribed into mRNA is further translated to proteins, the molecules that perform most of the functions in cells. Therefore, by measuring gene expression levels, we may be able to infer critical information about the functionality of cells or whole organisms [9]-[11], study diseases and the effects of drugs on them [12]-[18], etc.

Today, the sensitivity, dynamic range, and resolution of the conventional DNA microarrays is limited by shot-noise, cross-hybridization, saturation, probe density variations, sample preparation, as well as several other sources of noise and systematic errors in the detection procedure [7], [19], [20]. For instance, during a hybridization phase, including the steady-state, the number of formed target-probe pairs varies due to the probabilistic nature of hybridization. It has been observed that these variations are very similar to shot-noise (Poisson noise) at high expression levels, yet more complex at low expression levels where interference becomes the dominating limiting factor of the signal strength [19], [21]. The interference is due to cross-hybridization, a process in which targets may bind not only to their specific probes but to others as well. On the other hand, saturation may limit dynamic range if the number of targets is much larger than the number of available probes. Additionally, the measurements are also corrupted by the noise due to imperfect instrumentation and other biochemistry independent noise sources. The sources of noise in conventional

DNA microarrays are illustrated in Figure 1.

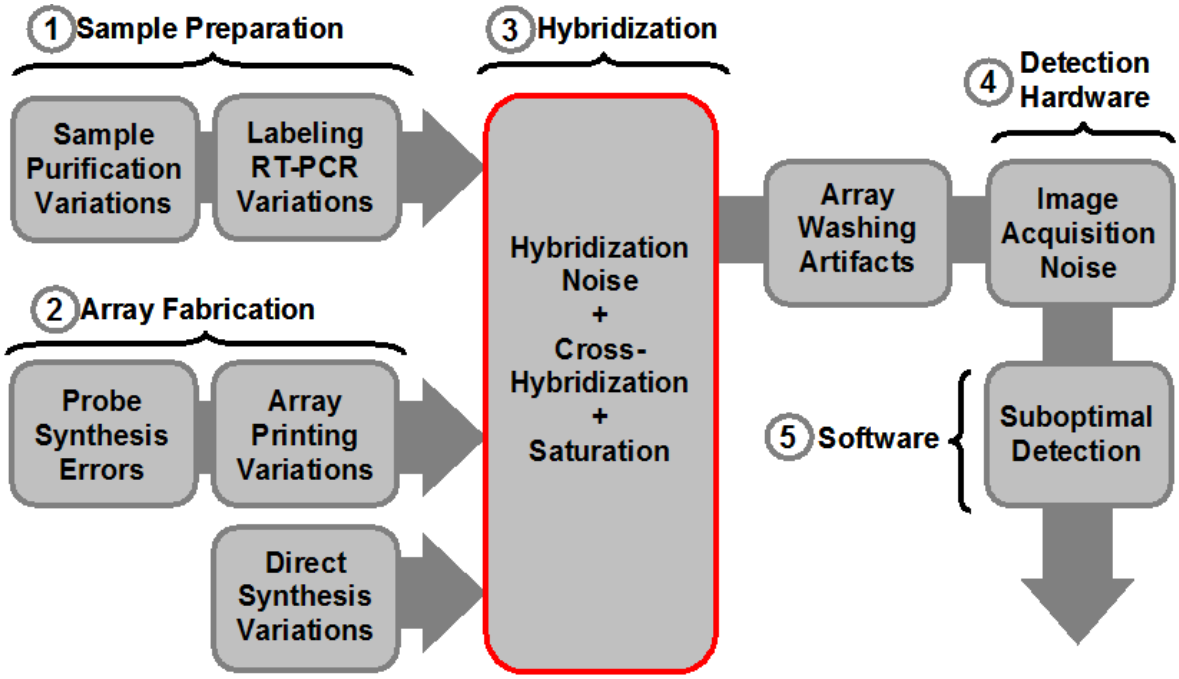

Figure 1: *Sources of noise in conventional microarrays.*

Many of the aforementioned limitations of conventional microarrays stem from the fact that they attempt to characterize hybridization process based on a single measurement of its steady-state. In conventional microarrays, measured signals emanate from the fluorescently labeled target molecules which have hybridized to the probes on the surface of a microarray. Typically, detection of the captured targets is carried out by scanning and/or various other imaging techniques after the hybridization step is completed. The reason for this is simple: a large concentration of floating (i.e., unbounded) labeled targets in the hybridization solution may overwhelm the specific signal emanating from the captured targets. Hence, the conventional microarrays typically do not allow the presence of the solution during the fluorescent and reporter intensity measurements. Therefore, the solution is typically washed away before the measurements are taken.

Intuitively, acquiring larger amount of useful data may improve the signal-to-noise ratio (SNR) and the

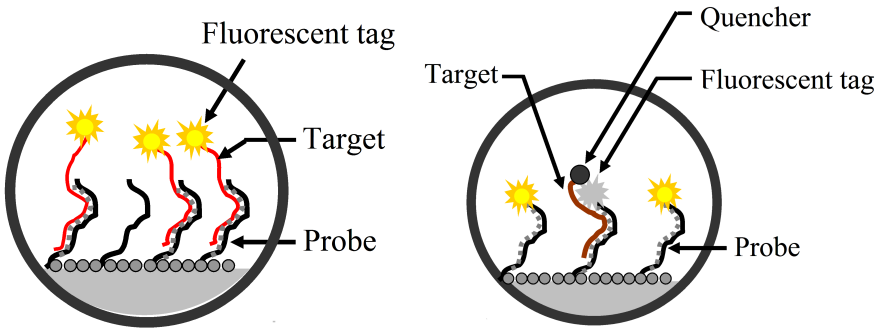

Figure 2: *Illustration how the target-probe binding event is reported in conventional (left) and real-time (right) microarrays.*

performance of microarrays. However, the conventional fluorescent-based DNA microarray are incapable of providing such additional data. This is the motivation behind *real-time microarrays* which are capable of evaluating the abundance of multiple targets in a sample by performing a *real-time* detection of the target-probe binding events [22], [23]. Real-time microarrays comprise probes that are labeled with fluorescent molecules and are used to evaluate the abundance of targets that are labeled with quenchers, entities that deactivate (quench) excited states of fluorescent molecules (by, say, energy transfer). In particular, in the event of a target-probe binding, the quencher attached to the target sequence gets in close proximity of the fluorescent molecule located at the end of the probe sequence. The fluorescence resonance energy transfer (FRET) interaction between the fluorescent molecule and the quencher results in quenching, which in turn indicates the amount of targets captured. Since in real-time microarrays the floating targets are not fluorescently-labeled, it is possible to image the array as the hybridization reaction is unfolding. This allows one to measure the kinetics of the reaction in real-time by observing the rate at which the light intensity of the interacting probes decrease (due to the quenching). Moreover, real-time microarrays may employ various time averaging schemes to suppress the Poisson noise and fluctuation of the target bindings. Due to all these advantages, the real-time microarray systems achieve higher SNR, potentially significantly smaller

estimation error, and broader detection dynamic range compared to the conventional microarrays. Figure 2 illustrates how the target-probe binding event is reported in conventional (on the left) and real-time (on the right) microarrays.

Figure 3 indicates which of the problems that affect conventional microarrays (shown in Figure 1) are circumvented in real-time microarrays. In particular, since we can scan a real-time microarray before adding any of the targets, we can acquire information about the probe spots prior to an actual experiment and thus correct for variations due to the array fabrication process. Moreover, the wealth of data provided by real-time microarrays enables us to deal with the hybridization noise and saturation (see the discussion in Section 2 and Section 3); ultimately, it improves the detection and quantification of targets. We should also note that due to the real-time data acquisition, real-time microarrays do not require the washing step and are thus not affected by array washing artifacts (as implied by comparing Figure 3 with Figure 1).

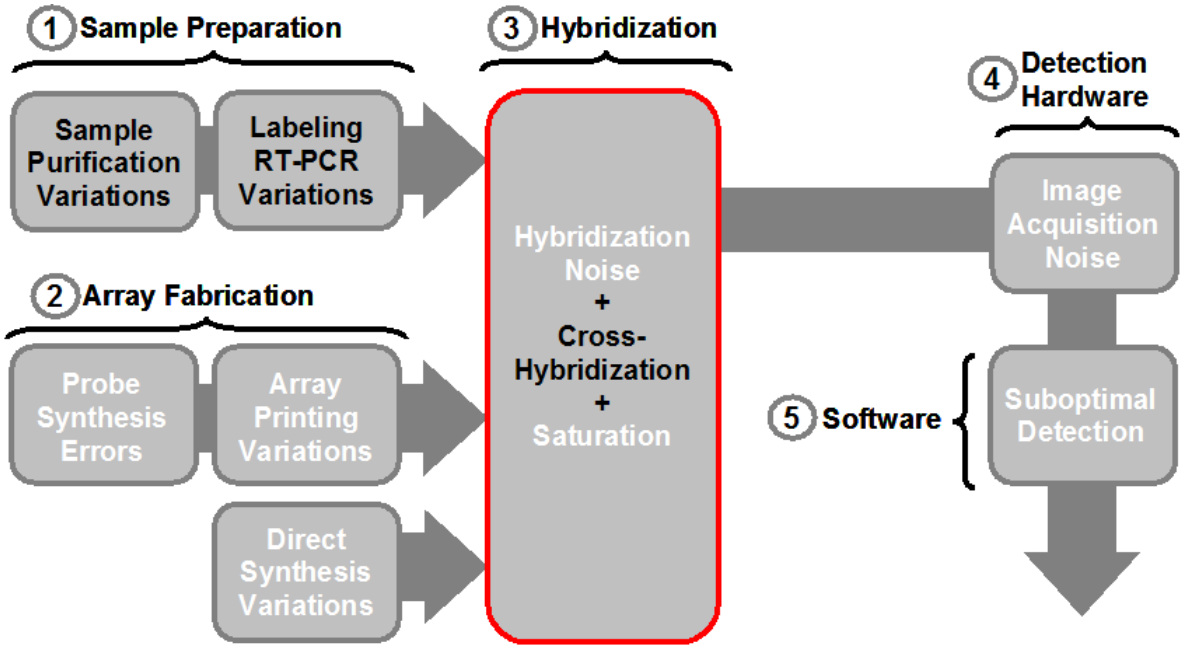

Figure 3: *Sources of noise in real-time microarrays. The effects of array fabrication variations, hybridization noise, and saturation are lessened, the array washing artifacts are eliminated, and the quality of detection is improved, as compared to conventional microarrays (see Figure 1).*

As a preview of the more detailed experimental results which will follow later, the process of data acquisition in real-time microarrays is illustrated in Figure 4. There, a few of the images acquired at different stages of the hybridization process in a custom-designed array are shown. It can be seen how the light in the probe spots which capture targets vanishes over time.

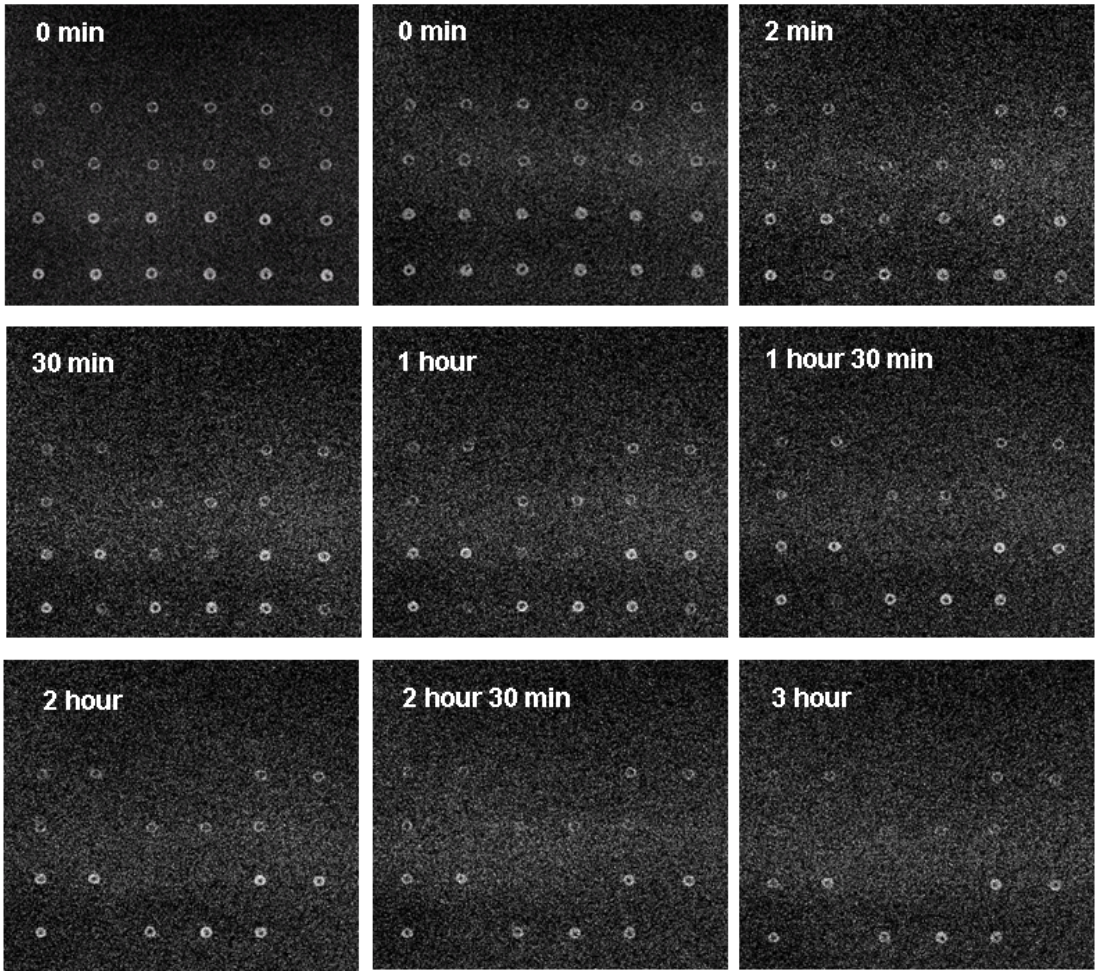

Figure 4: *A series of images acquired during a real-time microarray experiment.*

The paradigm shift in data acquisition, from measuring a single steady-state data point in the conventional microarrays to obtaining full hybridization kinetics in the real-time microarray systems, requires novel detection algorithms. These need to be preceded by the development of probabilistic models of the hy-

bridization process. There has been a significant amount of prior work on modeling hybridization (see, e.g., [24], [25]) and on modeling of hybridization in microarrays (see, e.g., [19], [21], [28], and the references therein). However, there are relatively few attempts on modeling the kinetics of hybridization, and consecutive experimental verification of those models. Examples include the real-time study of hybridization with optical wave guides in [26], and the study of the hybridization process in a fluorescence-based system with a single surface-bound probe and a single target in [27].

In this paper, we study the modeling of and estimation in real-time microarrays [23]. The paper is organized as follows. In Section 2, we develop a probabilistic model of the hybridization process and propose an estimator of the model parameters. The model parameters – in particular, the binding rate – can be used for quantification of the targets that are being tested. This is discussed in Section 3. Motivated by system identification ideas, in Section 4 we develop techniques for canceling cross-hybridization in real-time microarray experiments. Section 5 presents experimental results, while the summary and conclusion are given in Section 6.

2 A probabilistic model of the hybridization process

Before entering the discussion about target estimation, we first need to develop a probabilistic model of the hybridization process. Let the hybridization process start at $t = 0$. Consider the change in the number of target molecules bound to the probes in one of the spots of a real-time microarray during the time interval $(i\Delta t, (i+1)\Delta t)$. We can write

$$n_b(i+1) - n_b(i) = [n_t - n_b(i)]p_b(i)\Delta t - n_b(i)p_r(i)\Delta t,$$

where n_t denotes the total number of the target molecules present, $n_b(i)$ and $n_b(i+1)$ are the numbers of the bound target molecules at $t = i\Delta t$ and $t = (i+1)\Delta t$, respectively. Moreover, $p_b(i)$ denotes the probability that a free target binds to a probe during the i^{th} time interval; we note that $p_b(i)$ consists of two components, the probability that a target molecule is close to a probe and the probability that it binds to the probe. Finally, $p_r(i)$ denotes the probability that a bound target is released from the probe it is bound to during the i^{th} time interval.

Hence, we can write

$$\frac{n_b(i+1) - n_b(i)}{\Delta t} = [n_t - n_b(i)]p_b(i) - n_b(i)p_r(i). \tag{1}$$

The probability of an event where a target binds to a probe depends upon availability of the probes on the surface of an array – depletion of the number of available free probes means that, at any time, free targets compete for the remaining available probes and thus the binding probability decreases as the number of bound targets grows. If we denote the number of probes in a spot by n_p, a simple model for $p_b(i)$ is given by

$$p_b(i) = \left(1 - \frac{n_b(i)}{n_p}\right) p_b = \frac{n_p - n_b(i)}{n_p} p_b, \tag{2}$$

where p_b denotes the probability of the event where a target bounds to a probe assuming an unlimited abundance of the probes. [Note that, if the number of probes were infinite, one could approximate the binding probability by a constant.] On the other hand, probes depletion does not affect the release probability and it is reasonable to assume that the probability of an event where a bound target molecule gets released from a probe does not change between time intervals, i.e., $p_r(i) = p_r$, for all i.

By combining (1) and (2) and letting $\Delta t \to 0$, we arrive to the following differential equation,

$$\begin{aligned} \frac{dn_b}{dt} &= (n_t - n_b)\frac{n_p - n_b}{n_p}p_b - n_b p_r \\ &= n_t p_b - \left[(1 + \frac{n_t}{n_p})p_b + p_r\right] n_b + \frac{p_b}{n_p} n_b^2. \end{aligned} \tag{3}$$

[We note that in [28], the number of hybridized target-probe pairs is modeled by a rate equation similar to the nonlinear differential equation (3) (although derived differently). However, [28] employs the model only to analyze the equilibrium (i.e., the steady-state) of the reaction, and do not study kinetics of the hybridization process.]

Note that in (3), only $n_b = n_b(t)$, while all other quantities are constant parameters, albeit unknown. Before proceeding any further, we will find it useful to denote

$$\alpha = (1 + \frac{n_t}{n_p})p_b + p_r, \quad \beta = n_t p_b, \quad \gamma = \frac{p_b}{n_p}. \tag{4}$$

Clearly, from (4) we can express p_b, p_r, and n_p as $p_b = \frac{\beta}{n_t}$, $p_r = \alpha - (1 + \frac{n_t}{n_p})p_b$, and $n_p = \frac{p_b}{\gamma}$. Moreover, using (4), we can write (3) as

$$\frac{dn_b}{dt} = \beta - \alpha n_b + \gamma n_b^2 = \gamma(n_b - \lambda_1)(n_b - \lambda_2), \tag{5}$$

where λ_1 and λ_2 are introduced for convenience and are given by

$$\lambda_{1,2} = \frac{n_p}{2}\left(\frac{p_r}{p_b} + 1 + \frac{n_t}{n_p}\right) \pm \frac{n_p}{2}\sqrt{\left(\frac{n_t}{n_p} - 1\right)^2 + \left(\frac{p_r}{p_b} + 1\right)^2 + 2\frac{n_t p_r}{n_p p_b} - 1}.$$

Note that $\gamma = \beta/(\lambda_1\lambda_2)$. The solution to (5) is found as

$$n_b(t) = \lambda_1 + \frac{\lambda_1(\lambda_1 - \lambda_2)}{\lambda_2 e^{\beta(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})t} - \lambda_1}. \tag{6}$$

We should point out that (3) describes the change in the amount of target molecules, n_b, captured by the probes in a single probe spot of the microarray. Similar equations, possibly with different values of the parameters n_p, n_t, p_b, and p_r, hold for other spots and other targets.

From (6) it follows that

$$\beta = n_t p_b = \left.\frac{dn_b}{dt}\right|_{t=0}. \tag{7}$$

Thus, the slope of the hybridization curve at $t = 0$ contains information about the amount of the target. Note that, in the real-time microarray experiments, we actually observe a decrease in the light intensity of fluorescent tags as targets bind to probes and quenchers "turn-off" the light, which is essentially information about $n_p - n_b$, not n_b; nevertheless, since

$$\left.\frac{dn_b}{dt}\right|_{t=0} = -\left.\frac{d(n_p - n_b)}{dt}\right|_{t=0},$$

we can indeed estimate the amount of targets from the early-stage hybridization data. This allows for broader dynamic range than that of conventional microarrays since by not waiting for steady-state of the reaction we alleviate the effect of saturation. Moreover, detection in real-time microarrays is potentially much faster than in conventional microarrays – the former may be done within minutes from the start of the hybridization process, while the latter requires hybridization to reach steady-state which may take several hours.

On a related note, inverse of the time constant reflecting how fast $n_b(t)$ in (6) reaches steady-state is

given by

$$\begin{aligned}\tau_{n_b}^{-1} &= -\beta\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_2}\right)\\ &= \gamma\left(\lambda_1-\lambda_2\right)\\ &= p_b\sqrt{\left(\frac{n_t}{n_p}-1\right)^2+\left(\frac{p_r}{p_b}+1\right)^2+2\frac{n_t p_r}{n_p p_b}-1}. \qquad (8)\end{aligned}$$

Clearly, $\tau_{n_b}^{-1}$ is a function of n_t/n_p. In fact, if $n_t >> n_p$, $\tau_{n_b}^{-1}$ is a linear function of the amount of targets since, in this case, $\tau_{n_b}^{-1} \approx p_b n_t/n_p$. Now, the larger the number of targets, n_t, the faster the reaction since more targets compete for n_p probes. For the same reason, the smaller the number of available probes, n_p, the faster the reaction. This can be used to further expand the dynamic range of a real-time microarray system. In particular, the dynamic range provided by a single probe spot is limited by the span of observable reaction rates – say, from seconds to hours. On the other hand, by having several probe spots with different amounts of probe molecules, we can observe a broader range reaction rates than with just one spot.

3 Estimating parameters of the model

In this section, we outline a procedure for the estimation of parameters of the model developed in Section 2. Ultimately, by observing the hybridization process, we would like to obtain n_t, the number of target molecules. In addition, to fully characterize the hybridization process (including the computation of the reaction rate), we also need to find the parameters p_b , p_r, and n_p. However, we do not have direct access to $n_b(t)$ in (6), but rather to $y_b(t) = kn_b(t)$, where k denotes a transduction coefficient. In particular, we

observe

$$y_b(t) = \lambda_1^* + \frac{\lambda_1^*(\lambda_1^* - \lambda_2^*)}{\lambda_2^* e^{\beta^*(\frac{1}{\lambda_1^*} - \frac{1}{\lambda_2^*})t} - \lambda_1^*}, \tag{9}$$

where $\lambda_1^* = k\lambda_1$, $\lambda_2^* = k\lambda_2$, and $\beta^* = k\beta$. For convenience, we also introduce

$$\gamma^* = \frac{\beta^*}{\lambda_1^*\lambda_2^*} = \frac{\gamma}{k}, \text{ and } \alpha^* = \gamma^*(\lambda_1^* + \lambda_2^*) = \alpha. \tag{10}$$

From (9), it follows that

$$\beta^* = \left.\frac{dy_b}{dt}\right|_{t=0}. \tag{11}$$

Assume, without a loss of generality, that λ_1^* is the smaller and λ_2^* the larger of the two, i.e., $\lambda_1^* = \min(\lambda_1^*, \lambda_2^*)$ and $\lambda_2^* = \max(\lambda_1^*, \lambda_2^*)$. From (9), we find the steady-state of $y_b(t)$,

$$\lambda_1^* = \lim_{t\to\infty} y_b(t). \tag{12}$$

So, from (11) and (12) we can determine β^* and λ_1^*, two out of the three parameters in (9). To find λ_2^*, γ^*, and β^*, one needs to estimate $\tau_{n_b}^{-1} = \gamma^*(\lambda_1^* - \lambda_2^*)$ from the acquired data (more on this in Subsection 3.1) and use it in combination with (10). Then, we may attempt to use (4) to obtain p_b, p_r, n_p, and n_t from α^*, β^*, and γ^*. However, (4) provides only 3 equations while there are 4 unknowns that need to be determined. Therefore, we need at least 2 different experiments to find all of the desired parameters. Assume that the arrays and the conditions in the two experiments are the same except for the target amounts applied. Denote the target amounts by n_{t_1} and n_{t_2}; on the other hand, it is reasonable to assume that p_b and p_r remain the same in the two experiments. Let the first experiment yield α_1^*, β_1^*, and γ_1^*, and the second one yield α_2^*, β_2^*,

and γ_2^*, where $\gamma_2^* = \gamma_1^*$. Then it can be shown that

$$p_b = \frac{\beta_1^*\gamma_1^* - \beta_2^*\gamma_2^*}{\alpha_1^* - \alpha_2^*}, \;\; p_r = \alpha_1^* - p_b - \frac{\beta_1^*\gamma_1^*}{p_b}. \tag{13}$$

Moreover,

$$n_p = \frac{p_b}{k\gamma_1^*}, \;\; n_{t_1} = \frac{\beta_1^*\gamma_1^*}{p_b^2} n_p, \quad n_{t_2} = \frac{\beta_2^*\gamma_2^*}{p_b^2} n_p. \tag{14}$$

Note that n_p, n_{t_1}, and n_{t_2} in (13) - (14) are known within the transduction coefficient k, where $k = \frac{y_b(0)}{n_p}$. To find k and thus unambiguously quantify n_p, n_{t_1}, and n_{t_2}, we need to perform a calibration experiment (i.e., an experiment with a known amount of targets n_t).

3.1 Estimating the amount of targets via least-squares

At the early stage of the hybridization reaction, the quadratic term in (5) can be neglected and we can write

$$\frac{dn_b}{dt} = \beta - \alpha n_b, \tag{15}$$

where α and β are given by (4). The solution to (15) is given by $n_b(t) = C(1 - e^{-t/\tau})$. The amount of unbound probes (which we measure), is given by

$$y_b(t) = Ce^{-t/\tau}, \tag{16}$$

where $\tau = 1/\alpha$ and $C = \beta/\alpha$. The amount of targets, n_t, can be estimated from C and τ. In particular, in a comparative experimental trial where a test sample containing n_{t_1} of a target is compared against a reference

sample containing n_{t_2} of the same target, we can write

$$\frac{n_{t_1}}{n_{t_2}} = \frac{\beta_1}{\beta_2} = \frac{C_1/\tau_1}{C_2/\tau_2} = \frac{C_1}{C_2}\frac{\tau_2}{\tau_1},$$

where $\{C_1, \tau_1\}$ and $\{C_2, \tau_2\}$ are the parameters of the model (16) for the early stage of the target's hybridization process in the test and the reference sample, respectively. [We should also note that for $n_t \gg n_p$, the measured signal follows (16) not only at the early stage but throughout the reaction. This holds since as seen from (8), for $n_t \gg n_p$ we have $\tau_{n_b}^{-1} \approx \alpha = p_b n_t/n_p$.]

The real-time microarray system samples the signal (i.e., the light intensity) of the probe spots at certain time intervals (multiples of Δ, say) and thus obtains the sequence

$$y_n = Ce^{-n\Delta/\tau} + v(n\Delta),$$

where $v(t)$ denotes the noise. Assume that the length of the sequence $\{y_n\}$ is N. To estimate τ and C, we solve the inconsistent linear system of equations,

$$\underbrace{\begin{bmatrix} \log y_b(1) \\ \log y_b(2) \\ \vdots \\ \log y_b(N) \end{bmatrix}}_{z} = \underbrace{\begin{bmatrix} 1 & -\Delta \\ 1 & -2\Delta \\ \vdots & \\ 1 & -N\Delta \end{bmatrix}}_{H} \cdot \underbrace{\begin{bmatrix} \log C \\ 1/\tau \end{bmatrix}}_{x}.$$

A straightforward solution minimizing the mean-square error is given by $\hat{x} = (H^*H)^{-1}H^*z$. This can be implemented via the computationally efficient recursive least-squares (RLS) algorithm (e.g., [33]).

4 Cross-hybridization

Focusing on the early phase of the hybridization process and its reaction rate opens up the possibility of suppressing cross-hybridization, an event where interfering targets bind to probes designed to test another target. When a single target analyte is present, the number of available probe molecules, or equivalently the light intensity of a probe spot, decays exponentially with time according to (16). If, in addition to hybridization of the target of interest, a number of other targets cross-hybridize to the same probe spot, the light intensity of the probe spot will decay as the sum of several exponentials,

$$I(t) = \sum_{k=0}^{K} C_k e^{-\alpha_k t}, \tag{17}$$

where index $k = 0$ corresponds to the desired target, and $k = 1, \ldots, K$ correspond to the cross-hybridizing analytes. The reaction rates for the different analytes differ due to different numbers of analytes, binding probabilities, etc. (we omit explicit expressions for brevity). Therefore, if we can estimate the reaction rates from (17), we should be able to determine the number of molecules for each of the analytes binding to the spot.

The real-time microarray system samples the signal and obtains the sequence

$$y_n = I(n\Delta) + v(n\Delta) = \sum_{k=0}^{K} C_k e^{-n\Delta\alpha_k} + v(n\Delta),$$

for $n = 0, 1, \ldots, T$, where T is the total number of samples, and $v(t)$ represents the measurement noise. Defining $u_k = e^{-\Delta\alpha_k}$, we may write

$$y_n = \sum_{k=0}^{K-1} C_k u_k^n + v(n). \tag{18}$$

The goal is to

(i) determine the value of K (i.e., how many analytes are binding to the probe spot),

(ii) estimate the values of the pairs $\{C_k, u_k\}$ for all $k = 0, \ldots, K-1$, and

(iii) determine the number of copies of each analyte.

To solve (i), i.e., to determine the number of exponential components in a noisy signal, the measurements are used to form the so-called Hankel matrix of the form

$$\begin{bmatrix} y_{T/2} & y_{T/2-1} & \cdots & y_1 & y_0 \\ y_{T/2+1} & y_{T/2} & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T/2+1} & y_{T/2} \end{bmatrix}.$$

When y_n is the sum of K exponentials, the above Hankel matrix has rank K, i.e., only K nonzero eigenvalues. When y_n is noisy, the standard practice is to compute the singular values of the Hankel matrix and estimate K as being the number of significant singular values.

The problem of determining the number of exponential signals in noisy measurements, and estimating the individual rates of each component, is a classical one in signal processing and is generally referred to as system identification (see, e.g., [29], [30], [31], [32], and the references therein). The basic idea is that, when the signal y_n is the sum of K exponentials, it satisfies a Kth order homogenous difference equation

$$y_n + h_1 y_{n-1} + \cdots + h_{K-1} y_{n-K+1} + h_K y_{n-K} = 0, \tag{19}$$

whose characteristic equation

$$z^K + h_1 z^{K-1} + \cdots + h_{K-1} z + h_K = 0 \tag{20}$$

has roots equal to the u_k in (18). Therefore, to find the u_k, from which we determine the rates α_k and thereby the amounts of targets present, we first must find the coefficients h_1, h_2, ..., h_K. In a noiseless scenario, Prony's method (see [29] and the references therein) provides an exact solution: using the measured data sequence $\{y_n\}$, from (19) we write

$$\underbrace{\begin{bmatrix} y_{K-1} & y_{K-2} & \cdots & y_1 & y_0 \\ y_K & y_{K-1} & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-K+1} & y_{T-K} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_K \end{bmatrix}}_{\mathbf{h}} = - \underbrace{\begin{bmatrix} y_K \\ y_{K+1} \\ \vdots \\ y_T \end{bmatrix}}_{\mathbf{b}} \tag{21}$$

for $T = 2K - 1$, which is solved to obtain $\mathbf{h}$. The u_k are then obtained as the roots of (20). Finally, to find $\mathbf{c} = [C_0 \;\; C_1 \;\; \ldots \;\; C_{K-1}]$, we solve the system $V\mathbf{c} = \mathbf{b}$, where the Vandermonde matrix,

$$V = \begin{bmatrix} 1 & \cdots & 1 \\ z_1 & \cdots & z_K \\ \vdots & \ddots & \vdots \\ z_1^{T-1} & \cdots & z_K^{T-1} \end{bmatrix}$$

spans the K-dimensional data subspace.

In practice, the measured data is noisy and thus we require robust estimation of the u_k. To this end, we

may use a variety of different techniques including – but not limited to – total least squares (TLS), ESPRIT, modified Prony's method, etc.

The TLS approach [29], in particular, addresses limitations of the ordinary least-squares (LS) solution to (21) (given by $\hat{\mathbf{h}} = -(B^*B)^{-1}B^*\mathbf{b}$). The LS limitations arise from the assumption that the data matrix B is noise free. In the TLS approach, one forms the $(T-K+1)\times(K+1)$ Hankel matrix

$$\begin{bmatrix} \mathbf{b} & B \end{bmatrix} = \begin{bmatrix} y_K & y_{K-1} & \cdots & y_1 & y_0 \\ y_{K+1} & y_K & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T-K+1} & y_{T-K} \end{bmatrix} \tag{22}$$

and then identifies the $\mathbf{h} = [h_1 \ \ldots \ h_K]$ as the $(K+1)^{st}$ right singular vector of (22) (for more details, see [29]). The complexity of the TLS approach is essentially determined by computing the singular value decomposition (SVD) of (22), which requires $O(K^3)$ computations.

The SVD is often the first step in the ESPRIT algorithm [31], too, where it is used to obtain a $T \times K$ matrix U which spans the signal subspace. Let U_1 denote the matrix comprising all but the last row of U, and let U_2 denote the matrix comprising all but the first row of U. It can be shown (see [31] for details) that the eigenvalues of $\Psi = U_1^{\dagger}U_2$ are good estimates of the u_k in (18), where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of its argument. The complexity of performing the SVD, computing the Ψ, and finding the eigenvalues of Ψ, is $O(T^3)$, $O(TK^2)$, and $O(K^3)$, respectively.

In Figure 5, the left plot compares the performances of the TLS approach, the ESPRIT algorithm, and the so-called modified Prony algorithm [32] (for brevity, we omit the discussion on the modified Prony algorithm and refer the interested reader to [32]). We plot the mean-square error as a function of the signal-to-noise ratio (SNR) (the SNR is computed with signal energy averaged over the time interval $0 < t < 200$min, the

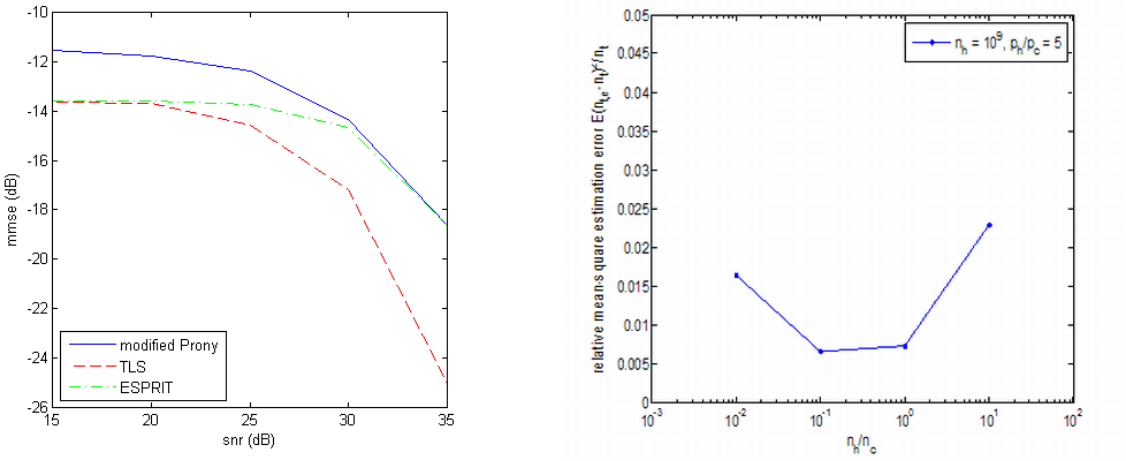

Figure 5: *Mean-square estimation error (relative) plotted as a function of the ratio* n_h/n_c*, given* $p_h/p_c = 5$*,* $n_h = 10^9$.

same interval during which the samples processed by the estimation algorithms were collected). We consider the situation where two targets bind to the same probe spot – one due to hybridization, and the other due to cross-hybridization. The parameters of the systems are chosen so as to mimic realistic experimental scenarios; in particular: $n_p = 1.6 \times 10^{11}$, $n_{t1} = n_{t2} = 10^{10}$, $p_{b1} = 4 \times 10^{-2}$, $p_{b2} = 2 \times 10^{-2}$. The TLS algorithm performs the best, followed by the ESPRIT algorithm, and the modified Prony approach.

On the right plot in Figure 5, we focus on the best performing one of the three considered algorithms, the TLS, and study its minimum mean-square estimation error over the range of ratios n_{t1}/n_{t2}. The probability of hybridization is assumed to be 5 times greater than the probability of cross-hybridization (i.e., $p_{t1}/p_{t2} = 5$). The number of hybridizing target molecules is $n_{t1} = 10^9$, while the number of cross-hybridizing molecules is varied. The simulation studies indicate potentially successful suppression of cross-hybridization over 3 orders of magnitude of n_{t1}/n_{t2}.

5 Experimental Verification

In this section, we present a series of experimental results which demonstrate the data acquisition and estimation in real-time microarrays. The microarrays were manufactured and the materials for experiments was prepared in the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology. The hybridization data is acquired with a Zeiss Pascal laser scanning microscope. The details of the experiments are given below.

Example 1 [Oligo targets.]

For the first set of experiments, we designed and printed a number of custom 6×6 microarrays, and employed them to test a set of oligo targets. For each target analyte there are multiple probe spots printed on an array, where different spots have different densities of probe molecules. The probes were labeled with Cy5 dyes, and the targets with BlackHole$^{\text{TM}}$ quenchers.

We consider two experiments and the data acquired therein; in the first experiment, 2ng/50μl of the target is applied to the microarray, whereas in the second experiment 0.2ng/50μl of the target is applied. Let us focus on one of the targets and two of the probe spots designed to test that target. One of the probe spots contains twice as many probe molecules as the other one; we refer to the former as the *high density* and to the latter as the *low density* probe spot. The hybridization process data acquired at the low and high density probe spots is shown in Figure 6 and Figure 7, respectively.

We employ the least-squares approach of Section 3.1 to process the signal measured in the early part of the reaction and compute the corresponding time constants (since the starting light intensities are the same, $C_1 = C_2$, and we normalized the signal). The computed time constants are illustrated with the exponential fit $e^{-t/\tau}$, shown as the dashed curves in Figure 6 and Figure 7. As discussed in Section 3.1, the ratio of the time constants should correspond to the ratio of the target amounts in the respective experiments (in particular, as

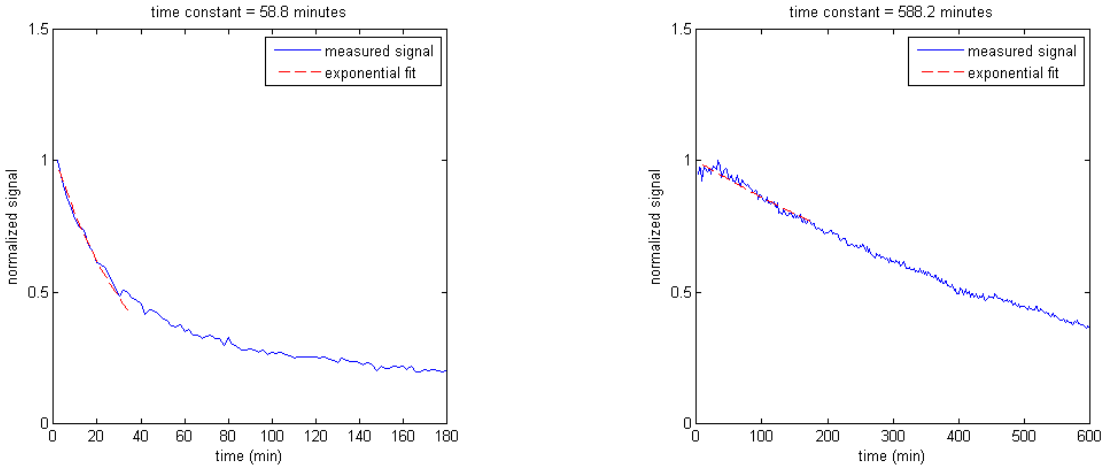

Figure 6: *The signal measured at a low density probe spot in an experiment with* 2*ng (left) and* 0.2*ng (right) of the oligo target applied to an array. The dashed line represents the exponential fit according to (16), where the time constant is computed using the least-squares as described in Section 3.1.*

stated in Section 3.1, $n_{t_1}/n_{t_2} = \tau_2/\tau_1$). This is indeed the case: the ratio of the time constants of the signal measured at the low density spots in the two experiments shown in Figure 6 is $\tau_2/\tau_1 = 10.0$. Moreover, the ratio of the time constants of the signal measured at the high density spots in the two experiments shown in Figure 7 is $\tau_2/\tau_1 = 11.6$. On the other hand, the ratio of the amounts of the target in the two experiments is precisely $n_{t_1}/n_{t_2} = 10$ (recall that the amounts of the target in the two experiments are $n_{t1} = 2$ng and $n_{t2} = 0.2$ng, respectively). This implies that we can accurately estimate relative ratio of the number of targets in a test and a reference sample by comparing the time constant of the hybridization process in the test sample with the time constant of the hybridization process in the reference sample.

Note that, in this example, conventional microarrays would not give reliable answers. For the low density spots in Figure 6, for instance, the reaction with the larger amount of target molecules reaches the steady-state in $2.5 - 3$ hours. The reaction with the smaller amount of target molecules takes $25 - 30$ hours to enter the equilibrium (the figure shows only 10 hours). A conventional microarray is typically left to hybridize for several hours, and then the corresponding measurements (a single data point for each spot) are

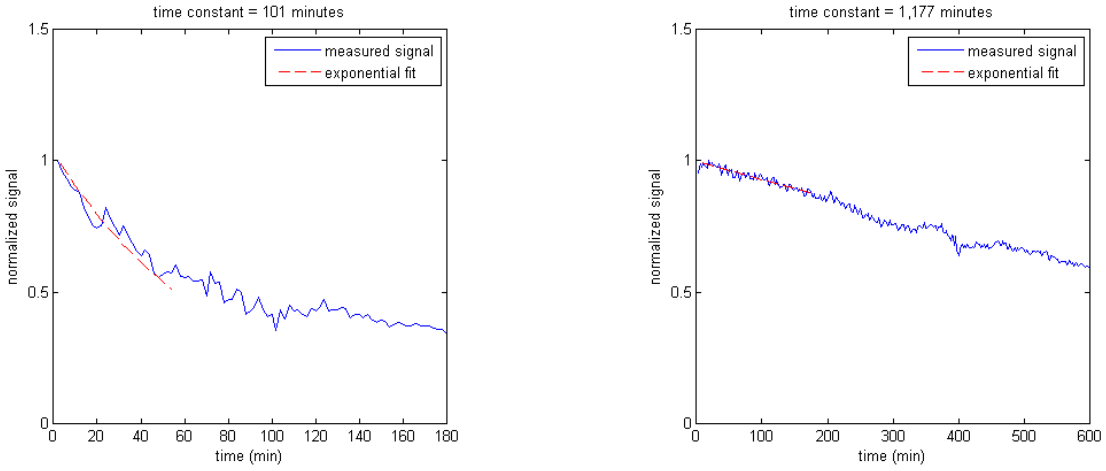

Figure 7: *The signal measured at a high density probe spot in an experiment with* $2ng$ *(left) and* $0.2ng$ *(right) of the oligo target applied to an array. The dashed line represents the exponential fit according to (16), where the time constant is computed using the least-squares as described in Section 3.1.*

compared against each other. But this implies that the conventional microarray technique would generate a result based on comparing a hybridization process which entered its equilibrium with another hybridization process which is far from its equilibrium – this certainly leads to quantitatively erroneous conclusions.

Example 2 [cDNA targets.]

For the following experiment, we used a number of cDNA targets. In particular, the targets were generated from The RNA Spikes^{TM}, a commercially available set of 8 purified *Escherichia Coli* RNA transcripts purchased from Ambion Inc. Lengths of the RNA sequences in the set are (750, 752, 1000, 1000, 1034, 1250, 1475, 2000), respectively. [These spikes are typically used for calibration purposes in conventional microarrays.] The RNA sequences were reverse transcribed to obtain the cDNA targets, which were then labeled with Cy5 dyes. Moreover, we designed 8 probes (25mer oligonucleotides) and printed slides where each probe was repeated in 6 different spots; hence, the printed slides have 48 spots.

We focus on two experiments, one where the concentrations of the targets was 80ng/50μl, and the other where the concentrations of the targets was 5 times smaller, i.e., 16ng/50μl. Consider the hybridization

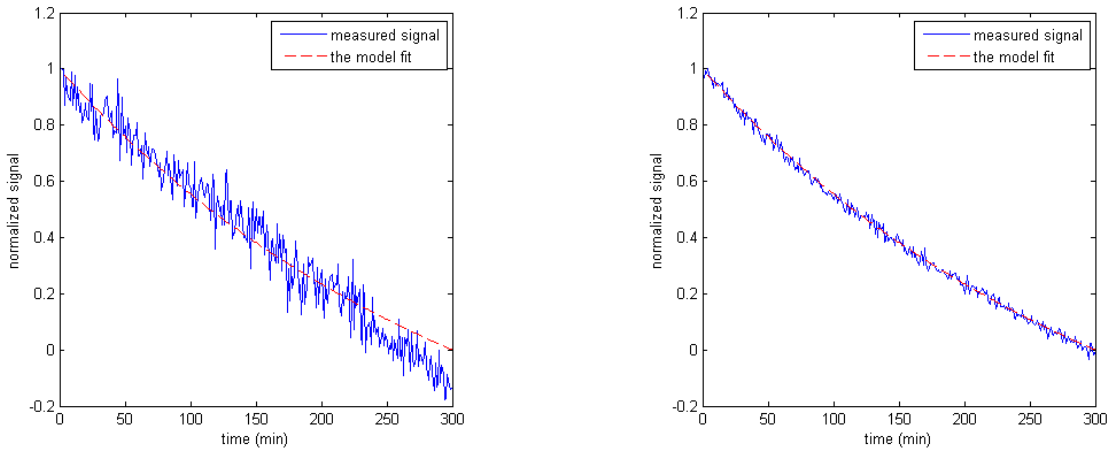

Figure 8: *The signal measured at a single probe spot in an experiment where* 80*ng (left) and 16ng (right) of the target is applied to the array. The smooth line represents the fit obtained using (9).*

data acquired by one of the probe spot in the two experiments. The signal measured in the first experiment, where 80ng of the target is applied to the array, is shown in the left plot of Figure 8. The dashed line shown in the same figure represents the fit obtained according to (9). In the second experiment, 16ng of the target is applied to the array. The measured signal, and the corresponding fit obtained according to (9), are both shown in the right plot of Figure 8.

By estimating the slopes of the hybridization signals, we find that

$$n_{t_1}/n_{t_2} = \beta_2/\beta_1 = 3.75. \tag{23}$$

Note that the above ratio is fairly close to its true value, $80/16 = 5$. Furthermore, from the acquired data we can estimate the parameters of the model developed in the previous sections. In particular, applying (13), we obtain $p_b = 1.9 \times 10^{-3}$, $p_r = 2.99 \times 10^{-5}$. Moreover, assuming that one of the experiments is used for calibration, we find that the value of the transduction coefficient is $k = 4.1 \times 10^{-4}$, and that the number of probe molecules in the observed probe spots is $n_p = 1.6 \times 10^{11}$.

Example 3 [cDNA targets in biological background.]

Finally, we repeated the experiments of Example 2 but with added biological background in order to emulate a realistic microarray experiment. The biological background employed is the total mouse DNA.

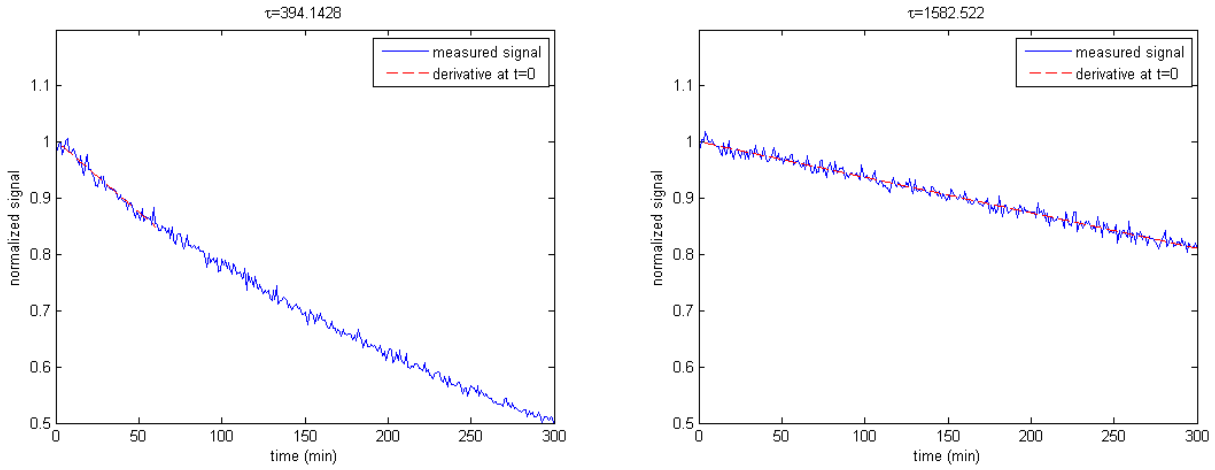

Figure 9: *The signal measured at a single probe spot in an experiment where* 80*ng (left) and 16ng (right) of the target, and* 2μ*g of the mouse DNA background, is applied to the array. The dashed line represents the derivative at* $t = 0$ *computed as (7).*

The signal measured in the first experiment, where 80ng of the target and 2μg of the mouse DNA is applied to the array, is shown in the left plot at Figure 9. The dashed line shown in the same plot represents the exponential fit according to (16), with the time constant computed via the least-squares as described in Section 3.1. In the second experiment, 16ng of the target and 2μg of the mouse DNA is applied to the array. The measured signal, and the corresponding exponential fit according to (16), are both shown in the right plot of Figure 9.

The ratio of the time constants of the measured signal in the two experiments is $\tau_2/\tau_1 = 4$, while the ratio of the amounts of targets is $n_{t1}/n_{t2} = 80/16 = 5$. This indicates robustness of the real-time microarrays with respect to the presence of rich biological background.

6 Summary and Conclusion

In this paper, we considered a novel real-time microarray system and the techniques for estimating the amounts of targets tested therein. Unlike the conventional ones which measure only the steady-state of a hybridization reaction, the real-time microarrays are capable of acquiring the entire kinetics of the reaction. We developed a probabilistic model for the kinetics of the hybridization process, and showed how to estimate the parameters of the model, including the amount of targets. Since the estimation is performed early in the hybridization process, the real-time microarray systems need not wait for the steady-state of the experiment; thus they have a significant speed advantage over the conventional microarrays.

In fact, many of the problems that affect conventional microarrays are circumvented in real-time microarrays. Since a real-time microarray can be scanned before a sample containing targets is applied to it, we can acquire information about the probe spots prior to an actual experiment; hence, we can correct for the inevitable variations occurring in the array fabrication process. In addition, the wealth of the data that the real-time microarrays provide enables us to deal with the hybridization noise and saturation. Ultimately, an increased amount of acquired data improves the accuracy, reliability, and dynamic range of the detection and quantification of targets.

Moreover, the real-time microarray data acquisition enables elimination of cross-hybridization. In particular, if more than one target binds to a microarray spot, each contributes an exponentially decaying component to the total signal acquired by the real-time microarray. Leveraging on the system identification ideas, we proposed techniques for separating the components of the composite signal, thus estimating the amounts of both the hybridizing as well as cross-hybridizing target analytes. This is a signal processing problem and we have solved it using advanced signal processing methods such as the total least-squares algorithm, modified Prony approach, the ESPRIT algorithm, etc.

Finally, we presented extensive experimental results verifying the validity of the model and demonstrated that the amounts of targets can be estimated with high accuracy. The experimental results suggest robustness of the platform and the estimation methodology with respect to the presence of rich biological background.

7 Acknowledgements

The authors would like to thank Dr. Jose Luis Reichmann and Vijaya Rao of the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology for their help with the experiments presented in the paper.

References

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270(5235), October 1995, pp. 467-70.

[2] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proceedings of the National Academy of Sciences (PNAS)*, 93(20), October 1996, pp. 10614-9.

[3] D. Shalon, S. J. Smith, and P. O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Gen. Research*, 6(7), July 1996, pp. 639-45.

[4] J. DeRisi et. al., "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics*, 14(4), December 1996, pp. 457-60.

[5] A. P. Blanchard, R. J. Kaiser, and L. E. Hood, "High-density oligonucleotide arrays," *Biosensors & Bioelectronics*, 1996, 11:687-690.

[6] A. P. Blanchard, and L. E. Hood, "Sequence to array: probing the genome's secrets," *Nature Biotechnology*, 1996, 14:1649.

[7] M. Schena, *Microarray Analysis*, John Wiley & Sons, 2003.

[8] U. R. Mueller and D.V. Nicolau (Eds.), *Microarray Technology and Its Applications*, Springer, Berlin, Germany, 2005.

[9] M. Schena et. al., "Microarrays: biotechnology's discovery platform for functional genomics," *Trends in Biotechnology* 1998, 16, 301-306.

[10] D. D. Shoemaker et. al., "Experimental annotation of the human genome using microarray technology," *Nature*, 409(6822), 2001, pp. 922-927.

[11] W. Zhang and I. Shmulevich (Eds.), *Computational and Statistical Approaches to Genomics*, Kluwer Academic Publishers, 2002.

[12] J. Kononen et. al., "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine*, 4(7), July 1998, pp. 844-847

[13] M. J. Marton et. al., "Drug target validation and identification of secondary drug target effects using DNA microarrays," *Nature Medicine*, 4(11), Novemebr 1998, pp. 1293-301.

[14] J. Khan et. al., "Expression profiling in cancer using cDNA microarrays," *Electrophoresis*, 20(2), February 1999, pp. 223-9.

[15] C. A. Afshari, E. F. Nuwaysir, and J. C. Barrett, "Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation," *Cancer Research*, 59(19), October 1999, pp. 4759-60.

[16] U. Scherf et. al., "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, 24(3), March 2000, pp. 236-44.

[17] J. Marx, "DNA arrays reveal cancer in its many forms," *Science*, September 2000, 289: 1670-1672.

[18] D. T. Ross et. al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, 24(3), March 2000, pp. 227-35.

[19] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proceedings of the National Academy of Sciences (PNAS)*, October 29, 2002, 14031-14036.

[20] W. Zhang, I. Shmulevich, and J. Astola, *Microarray Quality Control*, John Wiley and Sons, 2004.

[21] H. Vikalo, B. Hassibi, and A. Hassibi, "A statistical model for microarrays, optimal estimation algorithms, and limits of performance," *IEEE Transactions on Signal Processing, Special Issue on Genomics Signal Processing*, vol. 54, no. 6, June 2006.

[22] H. Vikalo, A. Hassibi, and B. Hassibi, "Signal processing real-time microarrays," *Asilomar Conference Signals, Systems and Computers*, Pacific Grove, CA, November 2007 (invited paper).

[23] A. Hassibi, H. Vikalo, and B. Hassibi, "Real-time microarrays," in preparation for submission to *Proceedings of the National Academy of Sciences (PNAS)*, 2007.

[24] J. SantaLucia, Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics", *Proceedings of the National Academy of Sciences (PNAS)*, 95, 1998, 1460-1465.

[25] J. SantaLucia, Jr. and D. Hicks, "The thermodynamics of DNA structural motifs", *Annu. Rev. Biophys. Biomol. Struct.* 33, 2004, 415-440.

[26] D. I. Stimpson et. al., "Real-time detection of DNA hybridization and melting on oligonucleotide arrays by using optical wave guides," *Proc. Natl. Acad. Sci. USA*, vol. 92, July 1995, 6379-6383.

[27] M. R. Henry, P. W. Stevens, J. Sun, and D. M. Kelso, "Real-time measurements of DNA hybridization on microparticles with fluorescence resonance energy transfer," *Analyt. Bioch.*, no. 276, 1999, 204-214.

[28] G. A. Held, G. Grinstein, Y. Tu, "Modeling of DNA microarray data by using physical properties of hybridization," *PNAS*, 100, June 2003, pp. 7575-7580.

[29] E. M. Dowling et. al., "Exponential parameter estimation in the presence of known components and noise," *IEEE Transactions on Antennas and Propagation*, vol. 42, no. 5, May 1994.

[30] A. J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace based signal analysis using singular value decomposition," *Proceedings of the IEEE*, 81(9):1277-1308, September 1993.

[31] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. on Acous., Speech and Signal Processing*, vol. 37, no. 7, July 1989.

[32] M. R. Osborne and G. K. Smyth, "A modified Prony algorithm for fitting sums of exponential functions," *SIAM J. Sci. Statist. Comput.*, 16, 1995, pp. 119-138.

[33] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, 2000.