# MODELING THE KINETICS OF HYBRIDIZATION IN MICROARRAYS

*H. Vikalo*[1], *B. Hassibi*[1], *M. Stojnic*[1], *and A. Hassibi*[2]

[1]Department of Electrical Engineering, California Institute of Technology, Pasadena, CA
[2]Department of Electrical and Computer Engineering, University of Texas, Austin, TX

## ABSTRACT

Conventional fluorescent-based microarrays acquire data after the hybridization phase. In this phase the targets analytes (i.e., DNA fragments) bind to the capturing probes on the array and supposedly reach a steady state. Accordingly, microarray experiments essentially provide only a single, steady-state data point of the hybridization process. On the other hand, a novel technique (i.e., real-time microarrays) capable of recording the kinetics of hybridization in fluorescent-based microarrays has recently been proposed in [5]. The richness of the information obtained therein promises higher signal-to-noise ratio, smaller estimation error, and broader assay detection dynamic range compared to the conventional microarrays. In the current paper, we develop a probabilistic model of the kinetics of hybridization and describe a procedure for the estimation of its parameters which include the binding rate and target concentration. This probabilistic model is an important step towards developing optimal detection algorithms for the microarrays which measure the kinetics of hybridization, and to understanding their fundamental limitations.

## 1. INTRODUCTION

A DNA microarray [1]-[3] is an affinity-based biosensor where the binding is based on hybridization, a chemical processes in which single DNA strands specifically bind to each other creating structures in a lower energy state. DNA microarrays are primarily used to measure gene expression levels, i.e., to quantify the process of transcription of DNA data 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, researchers may be able to infer critical information about functionality of the cells or the whole organism. Accordingly, a perturbation from the typical expression levels is often an indication of a disease; thus DNA microarray experiments may provide valuable insight into the genetic causes of diseases. Indeed, one of the ultimate goals of DNA microarray technology is to allow development of molecular diagnostics and creation of personalized drugs.

Today, the sensitivity, dynamic range and resolution of the DNA microarrays is limited by cross-hybridization [4] (which may be interpreted as interference), in addition to several other sources of noise and systematic errors in the detection procedure. The number of hybridized molecules varies due to the probabilistic nature of the 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 the interference (i.e., cross-hybridization) becomes the dominating limiting factor of the signal strength [4]. Additionally, the measurements are also corrupted by the noise due to imperfect instrumentation and other biochemistry independent noise sources.

Acquiring larger amount of the useful data (e.g., observing the entire hybridization process) would improve the signal-to-noise ratio in and the performance of microarrays. However, the conventional fluorescent-based DNA microarray are incapable of providing such additional data. There, the measured signal emanates from the fluorescently labeled target molecules which have hybridized to the probes at the surface of the microarray. Typically, the 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 (e.g., 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.

Recently, we have developed a novel *real-time microarray* (RT-$\mu$Array) system, capable of evaluating the abundance of multiple targets in a sample by performing the real-time detection of the target-probe binding events [5]. This system samples fluorescent signals emanating from the probes capturing quencher-labeled targets in the solution and thus does not require any washing step. The RT-$\mu$Array systems may employ various time averaging schemes to suppress the Poisson noise and fluctuation of the target bindings. Due to all these advantages, the RT-$\mu$Array systems achieve higher signal-to-noise ratio, potentially significantly smaller estimation error, and broader detection dynamic range compared to the conventional microarrays.

The paradigm shift in data acquisition, from measuring single steady-state data point in the conventional mi-

1-4244-0999-3/07/$25.00 

croarrays to obtaining full hybridization kinetics in the RT-$\mu$Array systems, requires novel detection algorithms. These need to be preceded by the development of probabilistic models of the hybridization process. [We note that quantification of targets in the RT-$\mu$Array systems can be performed by means of estimating the parameters of the hybridization kinetics – in particular, the binding rate.] There has been a significant amount of prior work on modeling hybridization (see, e.g., [6], [7]) and on probabilistic modeling of hybridization in microarrays (see, e.g., [4], [8], 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[9], and the study of the hybridization process in a fluorescence-based system with a single surface-bound probe and a single target in [10].

In this paper, we study the hybridization process measured by the RT-$\mu$Array [5]. We develop the probabilistic model of the process and propose an estimator of the model parameters.

## 2. A PROBABILISTIC MODEL OF THE HYBRIDIZATION PROCESS

For the models developed in this section, we assume that the hybridization in the microarrays under consideration is reaction-rate limited, rather than diffusion-limited. This is a reasonable assumption for the sample volumes used.

Assume that the hybridization process starts at $t = 0$, and consider discrete time intervals of the length $\Delta t$. Consider the change in the number of bound target molecules 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 target molecules, $n_b(i)$ and $n_b(i+1)$ are the numbers of bound target molecules at $t = i\Delta t$ and $t = (i+1)\Delta t$, respectively, and where $p_b(i)$ and $p_r(i)$ denote the probabilities of a target molecule binding to and releasing from a capturing probe during the $i^{th}$ time interval, respectively. Hence,

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

It is reasonable to assume that the probability of the target release does not change between time intervals, i.e., $p_r(i) = p_r$, for all $i$. On the other hand, the probability of forming a target-probe pair depends on the availability of the probes on the surface of the array. If we denote the number of probes in a spot by $n_p$, then we can model this probability as

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

where $p_b$ denotes the probability of forming a target-probe pair assuming an unlimited abundance of probes.

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

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

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

$$\begin{aligned} \alpha &= (1 + \tfrac{n_t}{n_p})p_b + p_r, \\ \beta &= n_t p_b, \qquad \gamma = \tfrac{p_b}{n_p}. \end{aligned} \quad (4)$$

Clearly, from (4),

$$p_b = \frac{\beta}{n_t},\ n_p = \frac{p_b}{\gamma},\ \text{and}\ p_r = \alpha - (1 + \frac{n_t}{n_p})p_b.$$

Using (4), we can write (3) as

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

where $\lambda_1$ and $\lambda_2$ are introduced for convenience and are given by

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

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

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

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. Moreover, (3) can be extended to model kinetics of both hybridization and cross-hybridization (i.e., non-specific binding). For instance, if we assume that the signal measured by a particular probe spot consists of a hybridization and a cross-hybridization component, they can be described by the following system of coupled differential equations,

$$\begin{aligned} \frac{dn_{b,h}}{dt} &= (n_h - n_{b,h})\frac{n_p - n_{b,h} - n_{b,c}}{n_p} p_h - n_{b,h} p_{r,h}, \\ \frac{dn_{b,c}}{dt} &= (n_c - n_{b,c})\frac{n_p - n_{b,h} - n_{b,c}}{n_p} p_c - n_{b,c} p_{r,c}, \end{aligned}$$

where $n_{b,h}$ and $n_{b,c}$ denote the number of specific and non-specific targets bound to probes, $n_h$ and $n_c$ denote the total number of specific and non-specific targets, and where $p_h$ and $p_c$ denote the probabilities of forming specific and non-specific target-probe pairs given an unlimited abundance of the probe molecules while $p_{r,h}$ and $p_{r,c}$ denote the probabilities of breaking those pairs, respectively. We refrain from a more detailed study of the cross-hybridization process in this paper.

## 3. ESTIMATING PARAMETERS OF THE MODEL

Here we outline a procedure for estimation of the parameters of the model developed in Section 2. Ultimately, by observing the hybridization process, we would like to obtain $p_b$ , $p_r$, $n_p$, and $n_t$. 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

$$y_b(t) = \lambda_1^* + \frac{\lambda_1^*(\lambda_1^* - \lambda_2^*)}{\lambda_2^* e^{\beta^*(\frac{1}{\lambda_1^*} - \frac{1}{\lambda_2^*})t} - \lambda_1^*}, \tag{7}$$

where

$$\lambda_1^* = k\lambda_1, \lambda_2^* = k\lambda_2, \text{ and } \beta^* = k\beta.$$

For convenience, we also introduce

$$\gamma^* = \frac{\beta^*}{\lambda_1^*\lambda_2^*} = \frac{\gamma}{k}, \text{ and } \alpha^* = \gamma^*(\lambda_1^* + \lambda_2^*) = \alpha. \tag{8}$$

From (5), it follows that

$$\beta^* = \left.\frac{dy_b}{dt}\right|_{t=0}. \tag{9}$$

Assume, without a loss of generality, that $\lambda_1^*$ is the smaller and $\lambda_2^*$ the larger of the two, i.e., $\lambda_1^* = \min(\lambda_1^*, \lambda_2^*)$ and $\lambda_2^* = \max(\lambda_1^*, \lambda_2^*)$. From (7), we find the steady-state of $y_b(t)$,

$$\lambda_1^* = \lim_{t\to\infty} y_b(t). \tag{10}$$

So, from (9) and (10) we can determine $\beta^*$ and $\lambda_1^*$, two out of the three parameters in (7). To find the remaining one, $\lambda_2^*$, one needs to fit the curve (7) to the experimental data.

Having determined $\lambda_1^*$, $\lambda_2^*$, and $\beta^*$, we use (8) to obtain $\alpha^*$ and $\gamma^*$. Then, we should use (4) to obtain $p_b$, $p_r$, $n_p$, and $n_t$ from $\alpha^*$, $\beta^*$, and $\gamma^*$. However, (4) gives us 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, $p_b$ and $p_r$ remain the same in the two experiments. Let the first experiment yield $\alpha_1^*$, $\beta_1^*$, and $\gamma_1^*$, and the second one yield $\alpha_2^*$, $\beta_2^*$, and $\gamma_2^*$, where $\gamma_2^* = \gamma_1^*$. Then it can be shown that

$$p_b = \frac{\beta_1^*\gamma_1^* - \beta_2^*\gamma_2^*}{\alpha_1^* - \alpha_2^*}, \tag{11}$$

and

$$p_r = \alpha_1^* - p_b - \frac{\beta_1^*\gamma_1^*}{p_b}. \tag{12}$$

Moreover,

$$n_p = \frac{p_b}{k\gamma_1^*}, \tag{13}$$

and

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

The following comments are in order. First, note that in (12)-(13) only the data obtained from one of the experiments (i.e., $\alpha_1^*$, $\beta_1^*$, and $\gamma_1^*$) are used for the parameter estimation. As an alternative, we could repeat (12)-(13) using $\alpha_2^*$, $\beta_2^*$, and $\gamma_2^*$, and then find $p_r$ and $n_p$ as the averages of their respective estimates. On another note, quantities (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 $n_t$).

## 4. EXPERIMENTAL VERIFICATION

In this section, we describe the experiments designed to test the validity of the proposed model and demonstrate the parameter estimation procedure. To this end, two DNA microarray experiments are performed. The custom $8 \times 9$ arrays contain 25mer probes printed in 3 different probe densities. The targets are Ambion mRNA Spikes, applied to the arrays with different concentrations. The concentrations used in the two experiments are 80ng/50$\mu$l and 16ng/50$\mu$l.

The signal measured in the first experiment, where 80ng of the target is applied to the array, is shown in Figure 1. The smooth line shown in the same figure represents the fit obtained according to (7). In the second experiment, 16ng of the target is applied to the array. The measured signal, and the corresponding fit obtained according to (7), are both shown in Figure 2.
Applying (11)-(14), we obtain

$$p_b = 1.9 \times 10^{-3}, \quad p_r = 2.99 \times 10^{-5}.$$

Furthermore, we find that

$$n_{t_1}/n_{t_2} = \beta_1^*/\beta_2^* = 3.75. \tag{15}$$

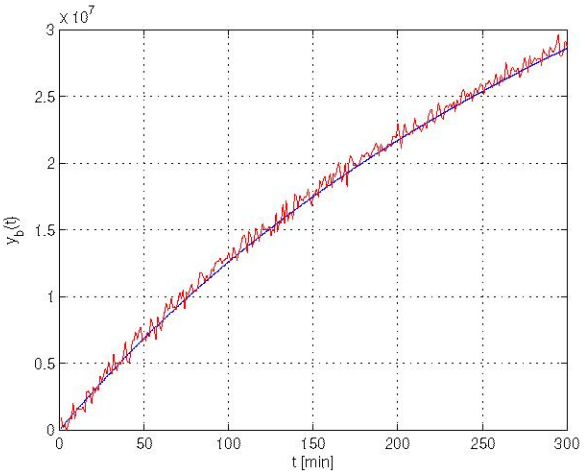


**Fig. 1**. *The measured signal from 80ng of Ambion Spike 3 applied to a microarray, and the mathematical fit according to (6).*

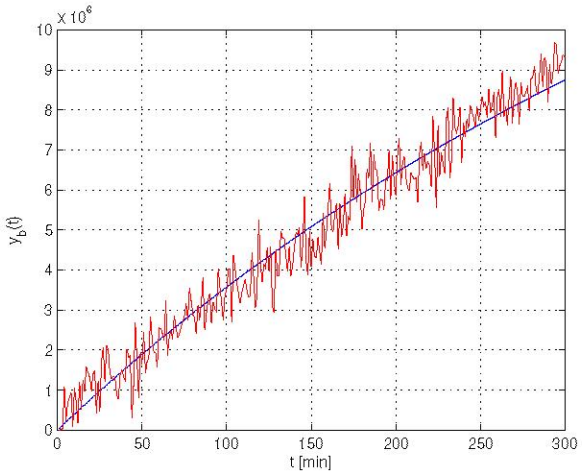


**Fig. 2**. *The measured signal from 16ng of Ambion Spike 3 applied to a microarray, and the mathematical fit according to (6).*

Note that the above ratio is relatively close to its true value, $80/16 = 5$. Finally, 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 $n_p = 1.6 \times 10^{11}$.

## 5. SUMMARY AND CONCLUSION

In this paper, we studied the kinetics of hybridization in microarrays. We developed a probabilistic model which encapsulates the hybridization process, and showed how to estimate the parameters of the model. Moreover, we presented experimental data to verify the validity of the model and demonstrate its applicability to the target quantification.

## 6. REFERENCES

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

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

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

[4] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," in *Proc. Natl. Acad. Sci. USA*, October 29, 2002, 14031-14036.

[5] A. Hassibi, H. Vikalo, and B. Hassibi, "Real-time microarrays," in preparation for submission to *PNAS*, 2007.

[6] J. SantaLucia, Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics", *Proc. Natl. Acad. Sci. USA*, 95, 1998, 1460-1465.

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

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

[9] 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.

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