A Statistical Model for Microarrays, Optimal Estimation Algorithms, and Limits of Performance

Haris Vikalo, Babak Hassibi, and Arjang Hassibi, *Member, IEEE*

***Abstract*—DNA microarray technology relies on the hybridization process, which is stochastic in nature. Currently, probabilistic cross hybridization of nonspecific targets, as well as the shot noise (Poisson noise) originating from specific targets binding, are among the main obstacles for achieving high accuracy in DNA microarray analysis. In this paper, statistical techniques are used to model the hybridization and cross-hybridization processes and, based on the model, optimal algorithms are employed to detect the targets and to estimate their quantities. To verify the theory, two sets of microarray experiments are conducted: one with oligonucleotide targets and the other with complementary DNA (cDNA) targets in the presence of biological background. Both experiments indicate that, by appropriately modeling the cross-hybridization interference, significant improvement in the accuracy over conventional methods such as direct readout can be obtained. This substantiates the fact that the accuracy of microarrays can become exclusively noise limited, rather than interference (i.e., cross-hybridization) limited. The techniques presented in this paper potentially increase considerably the signal-to-noise ratio (SNR), dynamic range, and resolution of DNA and protein microarrays as well as other affinity-based biosensors. A preliminary study of the Cramer–Rao bound for estimating the target concentrations suggests that, in some regimes, cross hybridization may even be beneficial—a result with potential ramifications for probe design, which is currently focused on minimizing cross hybridization. Finally, in its current form, the proposed method is best suited to low-density arrays arising in diagnostics, single nucleotide polymorphism (SNP) detection, toxicology, etc. How to scale it to high-density arrays (with many thousands of spots) is an interesting challenge.**

***Index Terms*—Cross hybridization, DNA microarrays, maximum *a posteriori*, maximum likelihood, minimum-mean-square-error (MMSE) estimation, Poisson noise, quantum-limited SNR, shot noise, statistical modeling.**

I. INTRODUCTION

OVER the past decade, high-throughput assay technologies have gained a lot of attention in the genomic research community. DNA microarrays, in particular, have attracted much interest due to the large scale and parallel nature of their experiments, as well as the richness of the information obtained by them. This stands in contrast to traditional techniques that are capable of analyzing only a small number of genes at a time.

DNA microarrays [1], [2] (which are, essentially, massively parallel affinity-based biosensors) 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.

A DNA microarray is basically an affinity-based biosensor where the binding is based on hybridization, a process in which complementary DNA (cDNA) strands specifically bind to each other creating structures in a lower energy state. Typically, the surface of a DNA microarray consists of an array (grid) of spots, each containing identical single-stranded DNA oligonucleotide capturing probes, whose locations are fixed during the process of hybridization and detection. Each single-stranded DNA capturing probe has a length of 25–70 bases, depending on the exact platform and application [1]. In the DNA microarray detection process, the mRNA targets that need to be quantified are initially used to generate fluorescent labeled cDNA, which are applied to the microarray afterwards. Under appropriate experimental conditions (i.e., temperature and salt concentration), labeled cDNA molecules that are the perfect match to the microarray probes will hybridize, i.e., bind to the complementary capturing oligos. Nevertheless, there will always be a number of nonspecific bindings since cDNA may nonspecifically cross-hybridize to probes that are not the perfect match but are rather only partial complements (having mismatches). It is important to understand that this particular phenomenon, i.e., nonspecific binding, is inherent to all affinity-based biosensors such as DNA or protein microarrays and also inevitable, given that it originates from the probabilistic and quantum mechanical nature of molecular interactions and biochemical bonds present in these systems [3]. Finally, the fluorescent intensities at each spot are measured to obtain an image, having correlation to the hybridization process, and thus the gene expression levels.

Manuscript received May 1, 2005; revised January 7, 2006. This work was supported in part by an award from the David and Lucille Packard Foundation for Science and Engineering, a Grubstake Award from the California Institute of Technology, and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jaakko Astola.

The authors are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: hvikalo@caltech.edu; hassibi@caltech.edu; arjang@caltech.edu).

Digital Object Identifier 10.1109/TSP.2006.873716

Today, the sensitivity, dynamic range and resolution of the DNA microarray data is limited by cross hybridization [6] (which may be interpreted as interference), in addition to several other sources of noise and systematic error in the detection procedure [7]. 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 ([6], [7]). In addition, the measurements are also corrupted by the noise due to imperfect instrumentation and other biochemistry independent noise sources.

Typically, cross hybridization is considered to be hurtful and often attempted to be suppressed by creating more specific probes. For instance, in the design of DNA microarrays, the capturing probes are often selected so that the sequences of nucleotides that comprise them are as unique as possible, and different from others as much as possible [5]. Nevertheless, if the application requires distinguishing among similar targets, cross hybridization is certainly present and perhaps limiting the accuracy. This may often be the fundamental limitation in microarrays designed for diagnostics and single nucleotide polymorphism (SNP) detection, for instance.

One of the main challenges for a precise target detection and quantification is the correct identification and modeling of the noise sources, and the consecutive incorporation of the noise model in the design of optimal estimators. While the former has recently been experimentally studied (see, e.g., [6]), the latter is still largely unexplored. In this paper, we describe the hybridization and cross-hybridization processes by Markov chains, similar to the techniques employed for modeling affinity based sensors in [3], which suggests that these biosensors have a quantum-limited SNR. Using the stationary distribution of the Markov chains, we formulate a statistical model of the microarray measurements. We note that a statistical modeling of DNA microarray data, using the experimentally measured correlation between observed hybridization intensity and calculated free energy of hybridization, was also proposed in [9]. In a related work [8], another statistical model that addresses individual probe-specific effects and automatic detection and handling of outliers and image artifacts, was proposed. However, we believe our model and algorithms go well beyond those of [8] and [9].

In our model, the biological noise is modeled as shot noise, thus accounting for the inherent fluctuations of the measured signal. We consider various criteria for the design of optimal algorithms for the detection of the presence and the estimation of the quantity of the target molecules. In particular, we consider the maximum-likelihood, maximum *a posteriori*, and constrained least-squares criteria. Therefore, instead of trying to suppress the cross hybridization, we essentially exploit it. This results in an increase in the signal-to-interference-and-noise ratio (SINR), and accordingly the precision of the microarray becomes limited by only the inherent noise, getting closer to its fundamental quantum-limited SNR.

The paper is organized as follows. In Section II, we develop a probabilistic model of the DNA microarray. Based on this model, we consider several algorithms for optimal detection of the target concentrations in Section III and derive the Cramer–Rao lower bound (CRLB) on the minimum mean-square error of all estimators. In Sections IV and V, we test the performance of the proposed technique on two sets of experiments: one involving oligonucleotide targets and the other cDNA targets in rich biological background. Finally, conclusions and directions for future work (such as how to scale the calibration experiments to high-density arrays) are discussed in Section VI.

Preliminary results of this paper were first reported in [10].

II. Probabilistic DNA Microarray Model

We consider an $m \times m$ DNA microarray, with $M \leq m^2$ different types of oligonucleotide probes attached to its surface. In other words, a particular oligonucleotide probe may be present at more than one spot of the array. Each probe is particularly designed to capture one of the possible targets in the sample that is required to be detected and quantified. We will assume that a total of n molecules of N different types of cDNA targets, $N \leq M$, each consisting of $c_1, c_2, \ldots, c_N$ molecules $(\sum_{i=1}^{N} c_i = n)$, are present in the sample that is applied to the microarray in the hybridization phase. For any target, there may be more than one spot on the $m \times m$ array where the complementary probes are located; we denote the number of spots with probes that are complements to the target of the type i by M_i, and note that $\sum_{i=1}^{M} M_i = m^2$. (For notational simplicity, we will in fact assume that $M = m^2$.) The array is scanned after the system has reached biochemical equilibrium. The resulting image has information about the number of targets captured at each spot and the goal is to detect which targets are present and to estimate their unknown concentrations c_i.

In general, in addition to hybridization to its matching oligonucleotide probe, each target molecule of type i may also engage in nonspecific cross hybridization with probes whose nucleotide sequences are only partial matches with the target. In particular, for each target i, we will denote by k_i the number of nonspecific cross hybridizations. In our model, we assume that both hybridization and cross hybridization are random events. Accordingly, let p_i^h denote the probability that a target of type i hybridizes to its matching probe. If we assume that p_i^H is the probability that target i hybridizes to its matching probe *when it is in the proximity of its matching probe*, then we can write

$$\begin{aligned} p_i^h &= p_i^H \cdot \text{Prob}(\text{target } i \text{ is in proximity of} \\ &\qquad\qquad \text{its matching probe}) \\ &= p_i^H \cdot \frac{1}{m^2} \end{aligned}$$

where we have used the fact that the target molecules are undergoing a random walk to deduce the $1/m^2$ factor. The reason for expressing p_i^h in terms of p_i^H is that the latter is what depends on factors such as the chemistry and probe and target sequences. For example, p_i^H can be estimated from the target and probe sequences, as well as the hybridization conditions, using the concepts of ΔG (Gibbs free energy change) and melting temperature (see, e.g., [11] and [12]). However, these give only

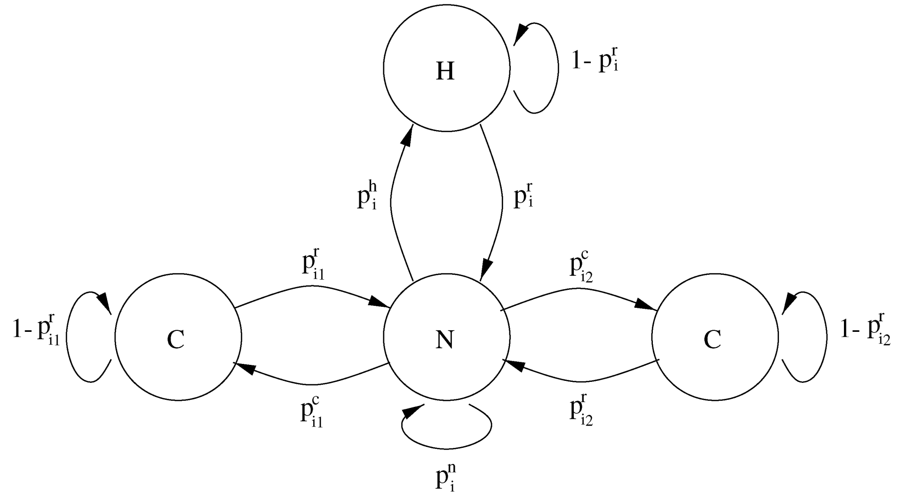


Fig. 1. Markov chain modeling states of a target molecule on a microarray with one specific and $k = 2$ nonspecific binding sites. The hybridized state is denoted by "H," cross-hybridized states are denoted by "C," and the unbound state is denoted by "N."

approximate values and further refinements are currently under investigation.

Furthermore, let p_{ij}^c denote the probability that a target of type i cross-hybridizes to a probe of type j. Similarly, we may write $p_{ij}^c = p_{ij}^C/m^2$, where p_{ij}^C is the probability that target i cross-hybridizes with probe j *when it is in its proximity*. We should note that cross hybridization is not necessarily reciprocal: in other words, in general $p_{ij}^c \neq p_{ji}^c$. In fact, cross hybridization need not even be mutual: If a target of type i cross-hybridizes to a probe of type j, the target j does not necessarily cross-hybridize to a probe of type i, i.e., it could be that $p_{ji}^c = 0$ even though $p_{ij}^c > 0$. Finally, the diffusion of the unbound target molecules is modeled as a random walk across the array [13]. Thus, in equilibrium, the distribution of the molecules is assumed to be uniform on the array [3].

If all we have is a probability of binding (i.e., hybridization and cross hybridization) then, if enough probes are present, eventually all the target molecules would bind to the probes. However, this is not the case since both hybridization and cross hybridization are reversible processes: once a target molecule is bound to a probe there is a nonzero probability that it will be released. We denote the release probability for hybridization (i.e., the probability that target i is released from probe i) by p_i^r and for cross hybridization (i.e., the probability that target i is released from probe j) by p_{ij}^r. In this sense, any target molecule of type i can be in one of $k_i + 2$ states: one state corresponding to hybridization to probe i, k_i states corresponding to cross hybridization to probes j, and one unbound state. The transition probabilities between these states are given by the probabilities p_i^h, p_{ij}^c, p_i^r, and p_{ij}^r. The corresponding Markov chain model is depicted in Fig. 1 for an example where $k_i = 2$. The probability $p_i^n = 1 - \sum_j p_{ij}^c - p_i^h$ in Fig. 1 denotes the likelihood that an unbound target remains free.

Remark 1: At this point, we need to mention an important assumption in our work. We will assume that the probabilities p_i^h and p_{ij}^c are *constant*. In other words, they do not depend on the number of target molecules that are bound to different probes. It is certainly conceivable that if there are not enough probes, and/or if there are too many target molecules, then as more targets bind to probes there will be less probes available for binding and so the binding probabilities p_i^h and p_{ij}^c will decrease. This will essentially lead to *saturation*. Therefore, in our model we will restrict ourselves to the case where saturation is not met, i.e., we will assume that the concentration of target molecules relative to the number of probes is low.[1]

Remark 2: Recently, correlation between probes on the chip (locational dependency) has been studied (see, e.g., [14]). We have not directly incorporated these into our model, although in principle it is possible to do so by adjusting the values of the p_i^h and p_{ij}^c in accordance with the location of those probes. The calibration experiments which we use to finetune the model in Sections IV and V do, in fact, make the corrections required by macroscopic issues such as the correlation between the probes.

What we are interested in is the probability that a given molecule of type i is in any of the aforementioned $k_i + 2$ states, once we have reached equilibrium. Let us denote this by the probability vector $\mu_i = [\mu_{i,1}\ \mu_{i,2}\ \ldots\ \mu_{i,k_i+2}]^T$, where $\mu_{i,1}$ is the probability of being in the hybridized state, $\mu_{i,j}, 2 \leq j \leq k_i + 1$ is the probability of being in the jth cross-hybridized state, and μ_{i,k_i+2} is the probability of being unbound. These probabilities are clearly given by the stationary distribution of the Markov chain, i.e., they satisfy

$$\mu_i = P_i \mu_i, \quad \mathbf{1}^T \mu_i = 1$$

where $\mathbf{1}$ denotes the vector of all 1's, and where the transition matrix P_i is given by

$$P_i = \begin{bmatrix} 1 - p_i^r & 0 & \cdots & 0 & p_i^h \\ 0 & 1 - p_{i1}^r & \cdots & 0 & p_{i1}^c \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 - p_{ik_i}^r & p_{ik_i}^c \\ p_i^r & p_{i1}^r & \cdots & p_{ik_i}^r & p_i^n \end{bmatrix}.$$

[1] We should mention that modeling the case where the binding probabilities p_i^h and p_{ij}^c are a function of the number of molecules already in a bounded state is quite interesting and will allow one to study microarrays when the target concentrations are high. However, the model and resulting estimation algorithms become quite more complicated.

Since what is measured in a microarray is (an indication of) the number of molecules bound to any particular probe, let us now turn our attention from target molecules to probes. Thus, consider the lth probe, $l = 1, 2, \ldots, m^2$ and let the number of target molecules of type i that are bound to it be given by n_{li}. Clearly, the total number of molecules bound to probe l is given by $n_l = \sum_{i=1}^{N} n_{li}$. Each n_{li} is an independent binomial random variable, one of which corresponds to hybridization and the remaining to (possible) cross hybridizations. Let us denote by q_{li} the probability that a target of type i is bound to probe l. These can be readily found from the earlier computed $\mu_{i,j}$s. In fact

$$q_{li} = \begin{cases} \mu_{i,1}, & \text{if target } i \text{ hybridizes with probe } l \\ \mu_{i,j_i}, & \text{if } l \text{ is the } j_i\text{th probe target } i \\ & \quad \text{cross-hybridizes to } (2 \le j_i \le k_i) \\ 0, & \text{otherwise} \end{cases}.$$

Since the total number of target molecules of type i that are available is given by c_i, the distribution of n_{li} is given by

$$p(n_{li} = x) = \binom{c_i}{x} q_{li}^x (1 - q_{li})^{c_i - x}. \tag{1}$$

Since the number of molecules involved is large, this is well approximated by a Gaussian random variable with the same mean $q_{li}c_i$ and variance $q_{li}(1 - q_{li})c_i$. Furthermore, since the n_{li} are independent, n_l is well approximated by a Gaussian random variable with mean $\sum_{i=1}^{N} q_{li}c_i$ and variance $\sum_{i=1}^{N} q_{li}(1-q_{li})c_i$.

Arranging the n_l into a $m^2 \times 1$ column vector $\mathbf{n} = [n_1 \; n_2 \; \ldots \; n_{m^2}]^T$, the measurement obtained from a DNA microarray is

$$\mathbf{s} = \mathbf{n} + \mathbf{v} \tag{2}$$

where $\mathbf{v}$ is the noise due to imperfect instrumentation (e.g., read noise of scanner or camera) and other biochemistry independent noise sources and can be well modeled as having iid Gaussian entries with zero mean and variance σ^2. Recall further that $\mathbf{n}$ also can be represented as having independent Gaussian entries with mean $\sum_{i=1}^{N} q_{li}c_i$ and variance $\sum_{i=1}^{N} q_{li}(1 - q_{li})c_i$. Thus, defining the $N \times 1$ column vector $\mathbf{c} = (1/m^2)[c_1 \; c_2 \; \ldots \; c_N]^T$, we may write the microarray master equation

$$\mathbf{s} = Q\mathbf{c} + \mathbf{w} + \mathbf{v} \tag{3}$$

where Q is the matrix with (l, i) component q_{li} and $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance matrix

$$E\mathbf{w}\mathbf{w}^T = \operatorname{diag}\left(\sum_{i=1}^{N} q_{1i}(1-q_{1i})c_i, \ldots, \sum_{i=1}^{N} q_{m^2 i}(1-q_{m^2 i})c_i\right). \tag{4}$$

The master equation (3) is the relationship between the measured signal $\mathbf{s}$ and the unknown target concentrations $\mathbf{c}$. Note that once Q and σ^2 are given, the model is fully specified. (Q can be obtained either by direct measurements or through knowledge of the probabilities p_i^h, p_{ij}^c, p_i^r and p_{ij}^r.) Note also that the unknown concentrations (the c_i) are also present in the covariance matrix of $\mathbf{w}$. In fact, this means that we have a shot-noise model.

III. Optimal Estimation of Target Concentrations

In this section, we state a few criteria that may be used for recovering the unknown vector $\mathbf{c}$ in the microarray master equation (3). Furthermore, we derive a lower bound (viz., the Cramer–Rao bound) on the minimum mean-square error of the target concentrations estimation. We also offer some discussions of the results.

Maximum-Likelihood Criterion: The maximum-likelihood (ML) estimate of the input concentrations maximizes the probability $p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}|\mathbf{c})$, i.e., it is obtained by solving the optimization problem

$$\max_{\mathbf{c} \ge 0} p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}|\mathbf{c}), \tag{5}$$

where, due to Gaussian distribution of both $\mathbf{w}$ and $\mathbf{v}$, we have

$$p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}|\mathbf{c}) = \frac{1}{(2\pi)^{M/2} \det(\Sigma_s)^{1/2}} e^{-\frac{1}{2}(\mathbf{s}-Q\mathbf{c})^T \Sigma_s^{-1} (\mathbf{s}-Q\mathbf{c})}$$

where the covariance matrix Σ_s is given by the equation shown at the bottom of the page. The optimization (5) is equivalent to the minimization

$$\min_{\mathbf{c} \ge 0} \left[(\mathbf{s} - Q\mathbf{c})^* \Sigma_s^{-1} (\mathbf{s} - Q\mathbf{c}) + \log\det\Sigma_s\right] \tag{6}$$

or more spelled out

$$\min_{c_i \ge 0} \sum_{l=1}^{m^2} \left[\frac{\left(s_l - \sum_{i=1}^{N} q_{li}c_i\right)^2}{\sigma^2 + \sum_{i=1}^{N} q_{li}(1 - q_{li})c_i} + \log\left(\sigma^2 + \sum_{i=1}^{N} q_{li}(1 - q_{li})c_i\right)\right]. \tag{7}$$

Note that the above problem is highly nonlinear and nonconvex. It can be, at best, solved via some iterative procedure. A good initial condition for any such iterative method can be found from the deterministic least-squares solution described further below.

Maximum a posteriori *Criterion:* In many cases, one may have prior information about the target concentrations. In this case, one would want to use the maximum *a posteriori* (MAP)

$$\Sigma_s = \begin{bmatrix} \sigma^2 + \sum_{i=1}^{N} q_{1i}(1 - q_{1i})c_i & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma^2 + \sum_{i=1}^{N} q_{m^2 i}(1 - q_{m^2 i})c_i \end{bmatrix}.$$

estimate that maximizes $p(\mathbf{c}|\mathbf{s}) = p(\mathbf{s}|\mathbf{c})p(\mathbf{c})/p(\mathbf{s})$, i.e., it is obtained by solving the optimization

$$\max_{\mathbf{c}\geq 0} p(\mathbf{s}|\mathbf{c})p(\mathbf{c}).$$

Here, this reduces to

$$\min_{\mathbf{c}\geq 0} \left[(\mathbf{s}-Q\mathbf{c})^*\Sigma_s^{-1}(\mathbf{s}-Q\mathbf{c}) + \log\det\Sigma_s - \log p(\mathbf{c})\right] \quad (8)$$

where $p(\mathbf{c}) = p(c_1,\ldots,c_N)$ is the *a priori* information about the joint presence of the different targets. The *a priori* information in the MAP estimation therefore accommodates potential use of information obtained previously by some other means, i.e., it allows for biological data fusion.

Deterministic Least-Squares Criterion: The deterministic least-squares (LS) solution is obtained by solving the following optimization problem:

$$\min_{\mathbf{c}\geq 0} \|\mathbf{s} - Q\mathbf{c}\|^2. \quad (9)$$

Although this criterion does not have as nice a stochastic interpretation, it is a quadratic program that can be solved exactly via efficient convex optimization techniques (e.g., the reflective Newton method—see [15]). In other words, the inequality constraints $c_i \geq 0$ do not pose a problem. In fact, any other prior information (such as upper and lower bounds on the concentrations, saturation, etc.) that can be cast as inequality (or, more generally, convex) constraints can be readily incorporated into the method and solution.

As mentioned earlier, the solution obtained from deterministic least-squares is often a very good initial condition for iterative methods used for solving the ML and/or MAP problems.

A. Limits of Performance

The minimum mean-square error (MMSE) of *any* estimation procedure is lower bounded by the Cramer–Rao bound [16]. We compute and use this bound to characterize the limits of achievable performance of target quantification in microarrays.

Assuming an unbiased estimator, the CRLB on the MMSE of estimating a parameter c_i is given by

$$E(\hat{c}_i - c_i)^2 \geq [F^{-1}]_{ii} \quad (10)$$

where the Fisher information matrix F is given by the negative of the expected value of the Hessian matrix of $\log p_{\mathbf{s}|\mathbf{c}}(\mathbf{s})$. In other words, the entries of F are given by

$$F_{ij} = -E_{\mathbf{s}}\frac{\partial^2}{\partial c_i \partial c_j}\log p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}). \quad (11)$$

Since the expectation is over only $\mathbf{s}$, F (and hence the CRLB) is a function of $\mathbf{c}$. We shall further find it convenient to define the entries of the Hessian matrix H as

$$H_{ij} = \frac{\partial^2}{\partial c_i \partial c_j}\log p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}).$$

Note now that H is a function of both $\mathbf{s}$ and $\mathbf{c}$.

In our case, the function whose second derivative we desire is

$$\begin{aligned} L(\mathbf{c}) &= \log\left(p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}|\mathbf{c})\right) \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma_s - \frac{1}{2}(\mathbf{s}-Q\mathbf{c})^T\Sigma_s^{-1}(\mathbf{s}-Q\mathbf{c}). \end{aligned}$$

Rather than attempt to compute the Hessian by evaluating two consecutive derivatives, we shall find it more convenient to do so by perturbing $\mathbf{c}$ around two of its components, say c_i and c_j, and noting that to second order

$$\begin{aligned} L(\mathbf{c}+\mathbf{e}_i\delta c_i + \mathbf{e}_j\delta c_j) &= L(\mathbf{c}) + [\delta c_i \quad \delta c_j]\begin{bmatrix}\frac{\partial L(\mathbf{c})}{\partial c_i} \\ \frac{\partial L(\mathbf{c})}{\partial c_j}\end{bmatrix} \\ &\quad + \frac{1}{2}[\delta c_i \quad \delta c_j]\begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj}\end{bmatrix}\begin{bmatrix}\delta c_i \\ \delta c_j\end{bmatrix} \end{aligned} \quad (12)$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the ith and jth unit vectors with ones in the ith and jth components, respectively, and zeros elsewhere. To determine the expansion (12), we will find it useful to write the covariance matrix Σ_s as

$$\Sigma_s = D_0 + \sum_{i=1}^{N} D_i c_i$$

where $D_0 = \sigma^2 I_{m^2}$, and where

$$D_i = \begin{bmatrix} q_{1i}(1-q_{1i}) & & & \\ & q_{2i}(1-q_{2i}) & & \\ & & \ddots & \\ & & & q_{m^2 i}(1-q_{m^2 i}) \end{bmatrix}.$$

Furthermore, note that we can write

$$\begin{aligned} L(\mathbf{c}+\mathbf{e}_i\delta c_i+\mathbf{e}_j\delta c_j) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}L_1(\mathbf{c}+\mathbf{e}_i\delta c_i+\mathbf{e}_j\delta c_j) \\ &\quad - \frac{1}{2}L_2(\mathbf{c}+\mathbf{e}_i\delta c_i+\mathbf{e}_j\delta c_j) \end{aligned}$$

where

$$L_1(\mathbf{c}+\mathbf{e}_i\delta c_i+\mathbf{e}_j\delta c_j) = \log\det(\Sigma_s + D_i\delta c_i + D_j\delta c_j)$$

and

$$\begin{aligned} L_2(\mathbf{c}+\mathbf{e}_i\delta c_i+\mathbf{e}_j\delta c_j) &= (\mathbf{s}-Q\mathbf{c}-Q\mathbf{e}_i\delta c_i - Q\mathbf{e}_j\delta c_j)^T \\ &\quad \times(\Sigma_s + D_i\delta c_i + D_j\delta c_j)^{-1}(\mathbf{s}-Q\mathbf{c}-Q\mathbf{e}_i\delta c_i - Q\mathbf{e}_j\delta c_j). \end{aligned}$$

To find the contribution of L_1 to the Hessian, we use

$$\log\det(I+X) \approx \mathrm{tr}X - \frac{1}{2}\mathrm{tr}X^2 + o(X^2)$$

to write

$$\begin{aligned} &L_1(\mathbf{c}+\mathbf{e}_i\delta c_i+\mathbf{e}_j\delta c_j) \\ &\quad = \log\det\Sigma_s + \log\det\left(I + \Sigma_s^{-1}D_i\delta c_i + \Sigma_s^{-1}D_j\delta c_j\right) \\ &\quad \approx \log\det\Sigma_s + \mathrm{tr}\left(\Sigma_s^{-1}D_i\delta c_i + \Sigma_s^{-1}D_j\delta c_j\right) \\ &\qquad - \frac{1}{2}\mathrm{tr}\left[\Sigma_s^{-1}D_i\Sigma_s^{-1}D_i(\delta c_i)^2 + 2\Sigma_s^{-1}D_i\Sigma_s^{-1}D_j\delta c_i\delta c_j \right. \\ &\qquad \left. + \Sigma_s^{-1}D_j\Sigma_s^{-1}D_j(\delta c_j)^2\right]. \end{aligned} \quad (13)$$

Comparing (13) and (12), it is clear that

$$\frac{\partial^2 L_1(\mathbf{c})}{\partial c_i \partial c_j} = -\frac{1}{2}\mathrm{tr}\left(\Sigma_s^{-1} D_i \Sigma_s^{-1} D_j\right). \qquad (14)$$

We now shift our attention to $L_2(\mathbf{c} + \mathbf{e}_i\delta c_i + \mathbf{e}_j \delta c_j)$. To find its contribution to the Hessian H, we use

$$(I + A)^{-1} \approx I - A + A^2 + o(A^2)$$

to obtain

$$\begin{aligned}(\Sigma_s &+ D_i\delta c_i + D_j \delta c_j)^{-1}\\ &= \left(I + \Sigma_s^{-1} D_i \delta c_i + \Sigma_s^{-1} D_j \delta c_j\right)^{-1}\Sigma_s^{-1}\\ &\approx \left[I - \Sigma_s^{-1}D_i\delta c_i - \Sigma_s^{-1} D_j \delta c_j + 2\Sigma_s^{-1}D_i\Sigma_s^{-1}D_j\delta c_i \delta c_j\right.\\ &\qquad \left. + \Sigma_s^{-1}D_i\Sigma_s^{-1}D_i(\delta c_i)^2 + \Sigma_s^{-1}D_j\Sigma_s^{-1}D_j(\delta c_j)^2\right]\Sigma_s^{-1}.\end{aligned}$$

Putting this back in the expression for $L_2(\mathbf{c} + \mathbf{e}_i\delta c_i + \mathbf{e}_j \delta c_j)$, it is not too difficult to identify

$$\begin{aligned}\frac{\partial^2 L_2(\mathbf{c})}{\partial c_i \partial c_j} &= (\mathbf{s} - Q\mathbf{c})^T \Sigma_s^{-1} D_i \Sigma_s^{-1} D_j \Sigma_s^{-1}(\mathbf{s} - Q\mathbf{c})\\ &\quad + (s - Q\mathbf{c})^T \Sigma_s^{-1} D_i Q \mathbf{e}_j + \mathbf{e}_i^T Q^T \Sigma_s^{-1} Q \mathbf{e}_j. \qquad (15)\end{aligned}$$

Using $E_{\mathbf{s}}(\mathbf{s} - Q\mathbf{c}) = \mathbf{0}$ and $E_{\mathbf{s}}(\mathbf{s} - Q\mathbf{c})(\mathbf{s} - Q\mathbf{c})^T = \Sigma_s$ to obtain the expectation of (15) and combining the result with (14) yields

$$E_{\mathbf{s}}\frac{\partial^2 L(\mathbf{c})}{\partial c_i \partial c_j} = -\mathbf{e}_i^T Q^T \Sigma_s^{-1} Q \mathbf{e}_j - \frac{1}{2}\mathrm{tr}\left(\Sigma_s^{-1} D_i \Sigma_s^{-1} D_j\right).$$

The (i, j) entry of the Fisher information matrix is therefore given by

$$F_{ij} = \mathbf{e}_i^T Q^T \Sigma_s^{-1} Q \mathbf{e}_j + \frac{1}{2}\mathrm{tr}\left(\Sigma_s^{-1} D_i \Sigma_s^{-1} D_j\right).$$

Note that with our definition of the diagonal matrices D_i we have

$$\begin{aligned}\mathrm{tr}\left(\Sigma_s^{-1} D_i \Sigma_s^{-1} D_j\right) &= \sum_{k=1}^{m^2} \frac{1}{\sigma_{s,k}} D_{i,k} \frac{1}{\sigma_{s,k}} D_{j,k}\\ &= \sum_{k=1}^{m^2} \frac{q_{ki}(1 - q_{ki}) q_{kj}(1 - q_{kj})}{\sigma_{s,k}^2}\end{aligned}$$

which is readily identified as the (i, j) component of the matrix $(Q - Q \odot Q)^T \Sigma_s^{-2}(Q - Q \odot Q)$, where $\odot$ represents the direct product, $(A \odot B)_{ij} = (A)_{ij}(B)_{ij}$. Note that $\sigma_{s,k}$ denotes the (k, k) component of Σ_s and $D_{i,k}$ is the (k, k) component of D_i. Therefore, we can write

$$F = Q^T \Sigma_s^{-1} Q + \frac{1}{2}(Q - Q \odot Q)^T \Sigma_s^{-2}(Q - Q \odot Q). \qquad (16)$$

Our end result, therefore, is

$$\begin{aligned}E(\hat{c}_i - c_i)^2 \geq \Bigg[\Bigg(&Q^T \Sigma_s^{-1} Q + \frac{1}{2}(Q - Q \odot Q)^T\\ &\times \Sigma_s^{-2}(Q - Q \odot Q)\Bigg)^{-1}\Bigg]_{ii}. \qquad (17)\end{aligned}$$

1) Comparison With Direct Readout: Note that, being unbiased, the ML estimate (7) achieves the Cramer–Rao bound in (17). In most current applications of microarrays, one assumes that $N = m^2$ and estimation is performed by direct readout. In this case it is easy to see that the mean-square error of direct readout is given by

$$E_{\mathbf{s}}(\mathbf{s} - \mathbf{c})(\mathbf{s} - \mathbf{c})^T = (Q - I)\mathbf{c}\mathbf{c}^T (Q - I)^T + \Sigma_s. \qquad (18)$$

Comparing (18) with (17) for a given system model and concentrations provides a measure of the improvement of the techniques proposed in this paper over the currently widely used methods that employ direct readout.

B. Effect of Cross Hybridization

In current microarray technology a great deal of effort is put into the design of the probes (often using some time-consuming form of combinatorial optimization) in such a way so as to minimize the effect of cross hybridization. In some important applications, such as SNP detection, the desired targets are inherently similar and so eliminating the effect of cross hybridization may not be completely possible.

Moreover, using the algorithms described in this paper, it may be that cross hybridization can be turned to one's advantage. Take, for simplicity, the extreme case where our sample has only a single target, i.e., $N = 1$. If an array has been designed that has no cross hybridization then, assuming the target present is the first target, it will only bind to probe site number one and not to any of the other sites. The Fisher matrix from (17) therefore becomes

$$F_{11}^{nc} = \frac{q_{11}^2}{\sigma^2 + q_{11}(1 - q_{11})c_1} + \frac{1}{2}\cdot\frac{q_{11}^2(1 - q_{11})^2}{\left(\sigma^2 + q_{11}(1 - q_{11})c_1\right)^2}. \qquad (19)$$

Assume now that the array does have cross hybridization, i.e., that target 1 can bind to probe k with probability q_{k1}. The Fisher matrix now becomes

$$\begin{aligned}F_{11}^{c} &= \sum_{k=1}^{m^2}\left[\frac{q_{k1}^2}{\sigma^2 + q_{k1}(1 - q_{k1})c_1}\right.\\ &\qquad\qquad \left. + \frac{1}{2}\cdot\frac{q_{k1}^2(1 - q_{k1})^2}{\left(\sigma^2 + q_{k1}(1 - q_{k1})c_1\right)^2}\right]\\ &= F_{11}^{nc} + \sum_{k=2}^{m^2}\left[\frac{q_{k1}^2}{\sigma^2 + q_{k1}(1 - q_{k1})c_1}\right.\\ &\qquad\qquad \left. + \frac{1}{2}\cdot\frac{q_{k1}^2(1 - q_{k1})^2}{\left(\sigma^2 + q_{k1}(1 - q_{k1})c_1\right)^2}\right]\\ &> F_{11}^{nc}.\end{aligned}$$

In other words, the existence of cross hybridization improves the accuracy of our estimate of target 1.

Of course, as one increases the number of targets beyond $N = 1$, one would expect the improvement in accuracy to diminish and, in fact, for large enough N for the accuracy to degrade compared to the case of no hybridization. However, for what value of N this transition occurs depends very much on the values of the parameters σ^2 and Q, on the concentration of the targets c_i and on the number of probes m^2.

To illustrate this, consider an artificial example where we have N targets that hybridize to their corresponding probes with probability $q_{ii} = q$ and that cross-hybridize to all other (m^2-1) probes with probability $q_{ij} = \beta$, $i \neq j$. Furthermore, assume that the concentration of all N targets are identical, i.e., $c_i = c$, for $i = 1, \ldots, N$. (The reason for choosing such symmetric parameters is that it will allow us to explicitly compute the inverse of the Fisher matrix F. We hope it will also give some insight into the more general setting.)

With these parameters, it is not difficult to see that

$$\Sigma_s = \sigma_s I_{m^2}, \quad \sigma_s = \left(\sigma^2 + q(1-q)c + (N-1)\beta(1-\beta)c\right)$$

and that (after some straightforward algebra)

$$F = \begin{bmatrix} a & b & \ldots & b \\ b & a & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ b & \ldots & b & a \end{bmatrix} = (a-b)I_N + \mathbf{1} \cdot b \cdot \mathbf{1}^T$$

where

$$a = \frac{q^2}{\sigma_s} + \frac{q^2(1-q)^2}{2\sigma_s^2} + (m^2-1)\left[\frac{\beta^2}{\sigma_s} + \frac{\beta^2(1-\beta)^2}{2\sigma_s^2}\right]$$

$$b = \frac{2q\beta}{\sigma_s} + \frac{q(1-q)\beta(1-\beta)}{\sigma_s^2} + (m^2-2)\left[\frac{\beta^2}{\sigma_s} + \frac{\beta^2(1-\beta)^2}{2\sigma_s^2}\right].$$

Now, inverting a matrix of the form of F above is straightforward since

$$\begin{aligned} F^{-1} &= \left((a-b)I_N + \mathbf{1} \cdot b \cdot \mathbf{1}^T\right)^{-1} \\ &= \frac{1}{a-b} I_N - \frac{1}{a-b}\mathbf{1}\frac{1}{b^{-1} + \frac{\mathbf{1}^T\mathbf{1}}{a-b}}\mathbf{1}^T\frac{1}{a-b} \\ &= \frac{1}{a-b}\left[I_N - \frac{\mathbf{1} \cdot b \cdot \mathbf{1}^T}{a+(N-1)b}\right]. \end{aligned}$$

Therefore

$$[F^{-1}]_{11} = \frac{1}{a-b} \cdot \frac{a+(N-2)b}{a+(N-1)b}. \tag{20}$$

This is the CRLB that should be compared with the one without cross hybridization in (19). Fig. 2 does this comparison for the parameters $\sigma^2 = 1000$, $c = 500$, $m^2 = 100$ (i.e., a 10 × 10 array), $q = 0.3$, and $\beta = 0.01$. As can be seen from the figure, cross hybridization is, in fact, beneficial when the number of targets is $N \leq 6$. Therefore, our artificial example seems to indicate that there is benefit in having cross hybridization in scenarios where the number of targets of interest in a given sample is much less than the number of probes on the array.

IV. Experimental Verification (I): Oligonucleotide Targets

In this section, we describe a set of experiments designed specifically to test our hypotheses regarding the statistical model and to verify the performance of the estimation algorithms on this experimental data.

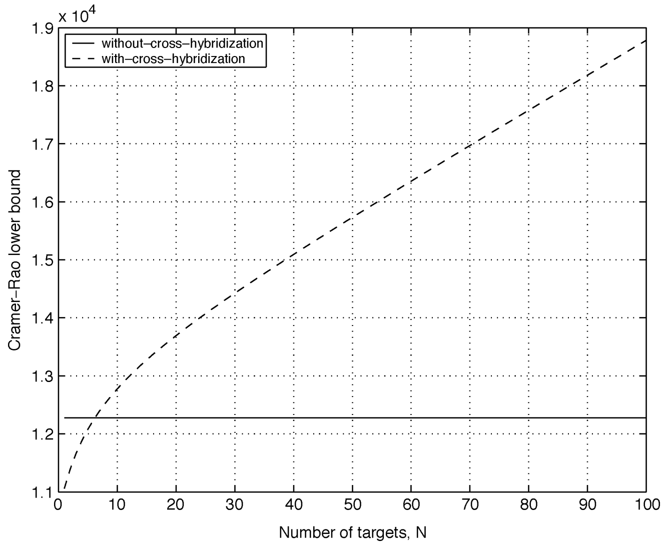


Fig. 2. CRLB with and without cross hybridization as a function of the number of target types N. The parameters are $\sigma^2 = 1000$, $c = 500$, $m^2 = 100$ $q = 0.3$, and $\beta = 0.01$.

Description of the Experiments: We started with a commercial set of oligonucleotide probes chosen from 96 genes of the bacterium *Escherichia coli* (specifically, the *E. coli* Array-Ready Oligo Set sample purchased from Operon Technologies, Huntsville, AL); denote this set by

$$\mathcal{P}_{96} = \{P_1, P_2, \ldots, P_{96}\}.$$

Each probe is a 70 bases long (i.e., a 70-mer) and, even though the set is commercial and designed with minimization of cross hybridization in mind, there are many pairs of probes that are mutually similar. We selected a subset of ten such probes, i.e., the probes are selected so that there is some cross correlation between the sequences of nucleotides comprising them. More specifically, we proceeded in the following manner. We selected the first probe as $p_1 = P_1$. To find the second probe, we used the sequence alignment functions in Matlab's Bioinformatics Toolbox to find one that had significant cross correlation with p_1. Call this probe p_2. In the process of determining p_2, we also designed two targets, t_1 and t_2, which are 25 mers such that they are Watson–Crick complements of certain subsequences of p_1 and p_2, respectively, and such that they have high cross correlation with a certain subsequence of the other probe.

After this, we proceeded in a sequential manner by determining a probe, say p_i, that has significant cross correlation with the probes selected earlier, $\{p_1, \ldots, p_{i-1}\}$, and in doing so designed a 25-mer target t_i that hybridizes perfectly to p_i, yet has high cross correlation with certain subsequences of the earlier probes $\{p_1, \ldots, p_{i-1}\}$.

At the end of this process, we obtained a subset of ten probes

$$\mathcal{P}_{10} = \{p_1, p_2, \ldots, p_{10}\}$$

as well as a set of ten targets

$$\mathcal{T}_{10} = \{t_1, t_2, \ldots, t_{10}\}.$$

The targets were highly purified and fluorescently labeled with Cy5 Cyanine dyes.

Two types of 10 × 10 arrays were designed:

- Type 1, which has all 96 probes from $\mathcal{P}_{96}$

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}	P_{19}	P_{20}
P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}	P_{27}	P_{28}	P_{29}	P_{30}
P_{31}	P_{32}	P_{33}	P_{34}	P_{35}	P_{36}	P_{37}	P_{38}	P_{39}	P_{40}
P_{41}	P_{42}	P_{43}	P_{44}	P_{45}	P_{46}	P_{47}	P_{48}	P_{49}	P_{50}
P_{51}	P_{52}	P_{53}	P_{54}	P_{55}	P_{56}	P_{57}	P_{58}	P_{59}	P_{60}
P_{61}	P_{62}	P_{63}	P_{64}	P_{65}	P_{66}	P_{67}	P_{68}	P_{69}	P_{70}
P_{71}	P_{72}	P_{73}	P_{74}	P_{75}	P_{76}	P_{77}	P_{78}	P_{79}	P_{80}
P_{81}	P_{82}	P_{83}	P_{84}	P_{85}	P_{86}	P_{87}	P_{88}	P_{89}	P_{90}
P_{91}	P_{92}	P_{93}	P_{94}	P_{95}	P_{96}	P_1	P_2	P_3	P_4

- Type 2, which contains only the probes from $\mathcal{P}_{10}$

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
p_{10}	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
p_9	p_{10}	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
p_8	p_9	p_{10}	p_1	p_2	p_3	p_4	p_5	p_6	p_7
p_7	p_8	p_9	p_{10}	p_1	p_2	p_3	p_4	p_5	p_6
p_6	p_7	p_8	p_9	p_{10}	p_1	p_2	p_3	p_4	p_5
p_5	p_6	p_7	p_8	p_9	p_{10}	p_1	p_2	p_3	p_4
p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_1	p_2	p_3
p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_1	p_2
p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_1

Determiming Parameters of the Statistical Model: To apply the estimation algorithms discussed in Section III, we need to determine matrix Q in (3). To this end, the probabilities of hybridization and cross hybridization of each of the 10 targets to any of the ten probes are determined based on both of the following:

- analytical expressions (ΔG, melting temperature, etc.; see, e.g., [12]);
- calibration experiments, where only one target is applied to a microarray, and its binding to each probe is quantified.

The melting temperature is used to get a rough estimate of the desired probabilities. Then, the calibration experiments are used to finetune them. We performed two sets of calibration experiments where the target quantity was 2 pmol in 50 μlit.

The final measurement obtained by the experiment is a 16-b image (scanned by GenePix scanner by Axon Instruments, Inc., Foster City, CA) with the intensities of the pixels ranging between 0 and 65 535. These intensities are correlated to the hybridization process. The results of the calibration experiments are summarized in the matrix R shown below.

$$R = kQ$$

$$= \begin{bmatrix} 65 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 3 & 55 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 62 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 22 & 3 & 60 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 28 & 23 & 48 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 32 & 6 & 54 & 0 & 0 & 0 & 0 \\ 9 & 4 & 6 & 4 & 41 & 34 & 56 & 0 & 0 & 0 \\ 1 & 1 & 1 & 3 & 4 & 25 & 46 & 40 & 0 & 0 \\ 0 & 0 & 2 & 0 & 8 & 0 & 5 & 2 & 63 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 16 & 61 & 46 \end{bmatrix} \times 10^3.$$

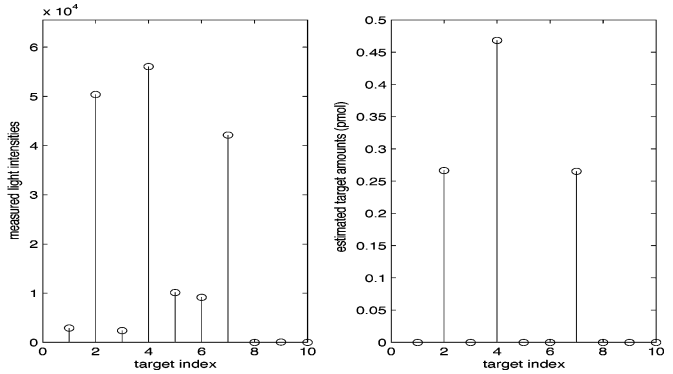


Fig. 3. Measured and estimated signal, T_1 mixture.

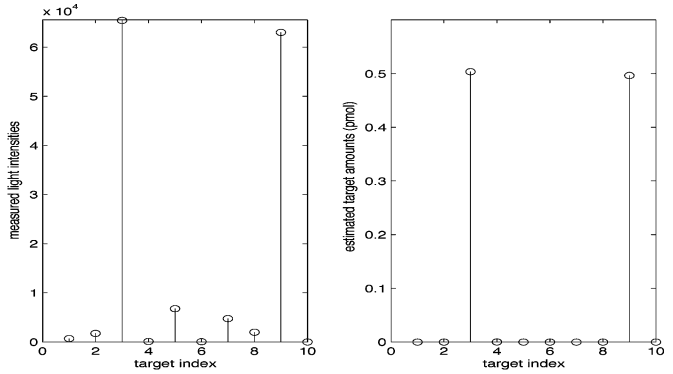


Fig. 4. Measured and estimated signal, T_2 mixture.

The matrix R is proportional to the probability matrix Q whose (i, j) component is the probability that target j binds to probe i. The reason for having the factor k is that we do not directly measure the number of molecules (as suggested by (3)) but rather light intensity. Therefore, k is essentially the factor that translates the concentration of target molecules to light intensity.

The peculiar (almost) lower-triangular structure of Q is an artifact of the sequential selection of the probes and targets in $\mathcal{P}_{10}$ and $\mathcal{T}_{10}$. Thus, t_j is likely to cross-hybridize to probes selected earlier, $i \leq j$, and not to ones selected later, $i > j$.

The Algorithm in Action: We tested the performance of the estimation algorithm in experiments where a mixture of 2 or 3 targets were applied to the designed microarrays. In particular, the mixtures

$$T_1 = \{t_2, t_4, t_7\}, \quad T_2 = \{t_3, t_9\}, \quad T_3 = \{t_5, t_6\}$$

were prepared, each with an equal amount of component target concentrations. The final concentrations of T_1, T_2, and T_3 were 1 pmol each and were applied to a 50 μlit microarray reaction buffer. Each experiment was replicated four times. GenePix and Matlab's Bioinformatics Toolbox were used for data analysis.

Figs. 3–5 show the measured signal and the estimated target quantities obtained from the constrained least-squares algorithm. Fig. 4 shows a very accurate estimation of the true target quantities of 50 pmol for targets t_3 and t_9. Fig. 3 shows a relatively good estimation of the target quantities of 33 pmol for targets t_2, t_4, and t_7. Note also that in both cases the artifacts due to cross hybridization are suppressed, i.e., no target is incorrectly detected (as a false positive).

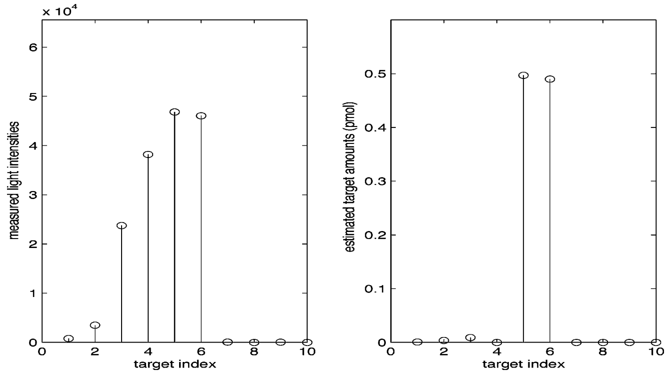


Fig. 5. Measured and estimated signal, T_3 mixture.

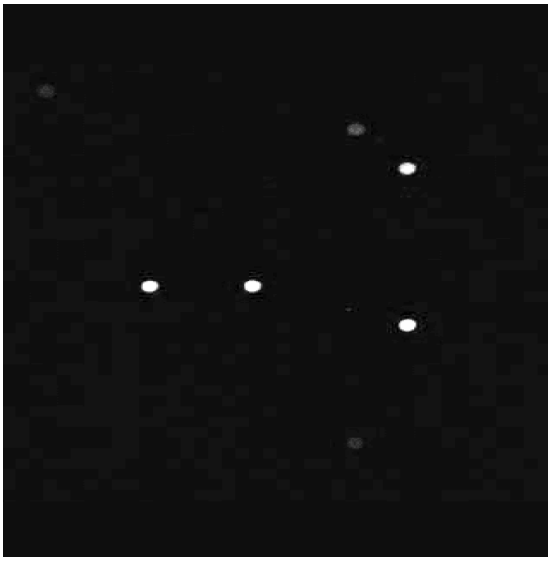


Fig. 6. Measured signal, T_3 applied to Type 1 microarray.

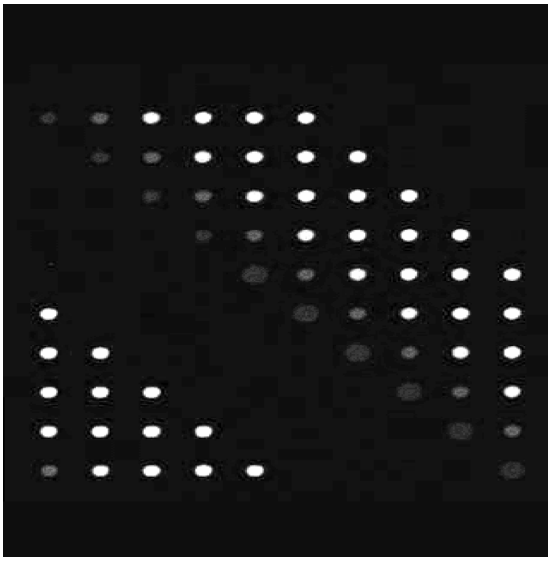


Fig. 7. Measured signal, T_3 applied to Type 2 microarray.

A particularly interesting experiment is the one where T_3 is applied to the microarray. The signal that was measured is shown in Fig. 6 and Fig. 7 and indicates a significant presence of binding to not only targets p_5 and p_6, but also to p_3 and p_4. The raw measured data vector is

$$[p1, \ldots, p_{10}] \\ = [800, 3520, 23760, 38200, 46820, 46060, 60, 0, 40, 0].$$

However, we know from our design of the experiment that the high levels of binding in spots 3 and 4 must be due to cross hybridization since T_3 contains only targets t_5 and t_6.

When our algorithm is applied to the measured data, it correctly identifies the presence of only two targets in the mixture, and quantifies them quite precisely as

$$[t1, \ldots, t_{10}] = [0, 0, 0, 0, 0.50, 0.49, 0, 0, 0, 0] \text{ pmol.}$$

V. Experimental Verification (II): cDNA Targets in Biological Background

In Section IV, we reported a set of microarrays experiments where the targets were synthesized 25-mer oligonucleotides. In this section, we present the results of an experiment wherein the targets are actual *E.Coli* cDNA molecules in a rich biological background (typical of actual microarray experiments).1

Description of the Experiments: The targets used in the experiment are generated from The RNA Spikes, a commercially available set of eight purified RNA transcripts purchased from Ambion, Inc., Austin, TX. The sizes of the RNA sequences are (750, 752, 1000, 1000, 1034, 1250, 1475, 2000), respectively. These spikes are used for calibration purposes in microarrays and so have been chosen such that the eight sequences have minimal correlation. We were therefore also very interested in whether we could observe cross hybridization effects in such a highly optimized set of targets. The RNA sequences were reverse transcribed to obtain cDNA targets, labeled with Cy5 dyes.

We designed 32 probes (25-mer oligonucleotides), four for each of the eight targets, and printed slides where each probe is repeated in ten different spots (hence, the slides have 320 spots). Denote the probes by p_{ij}, where $1 \leq i \leq 8$, $1 \leq j \leq 4$; therefore, p_{ij} denotes the jth probe for the ith target. Furthermore, let $\mathcal{P}_j = \{p_{ij}\}$, $1 \leq i \leq 8$ be the jth probe set. The probes are designed so that they cross-hybridize with one or more targets other than their intended ones.

The melting temperature is used to get a rough analytical estimate of the probabilistic model (i.e., of the matrix Q in the master equation (3)). Then, calibration experiments were performed in order to finetune the model. Two sets of eight calibration experiments were performed, wherein 2 ng of a single target was applied to a slide in every experiment. The experiments were done at $T = 24.8\,^{\circ}\mathrm{C}$, the data was acquired with a GenePix scanner by Axon Instruments and analyzed with GenePixPro 6.0 and Matlab's Bioinformatics Toolbox.

To test the performance of the estimation algorithm, we performed a set of experiments where a mixture of three targets was spiked with a complex biological background (from mouse total RNA). In particular, we used a mixture $(r_1, r_2, r_3) = (1, 0.75, 1.5)$ ng and spiked it with 500 ng of the mouse RNA. The experiment was replicated four times. The

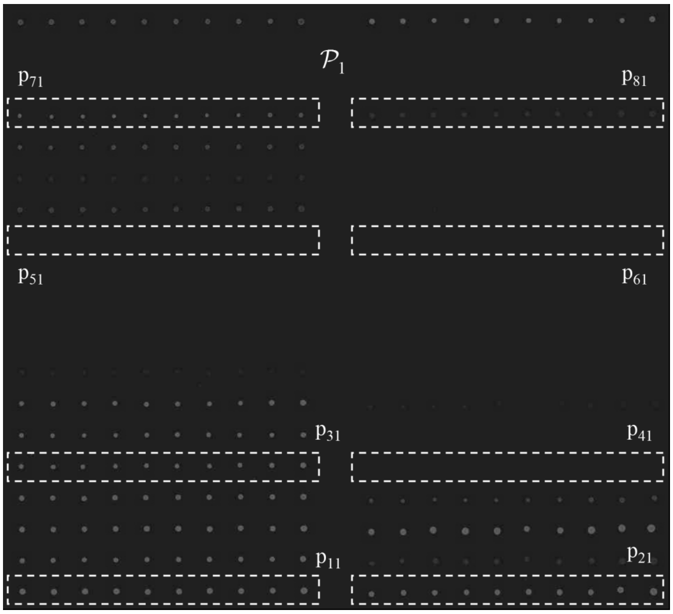


Fig. 8. Measured signal, $(r_1, r_2, r_3) = (1, 0.75, 1.5)$ ng of Ambion RNA Spikes in 500 ng of mouse RNA background. Locations of the probes from $\mathcal{P}_1$ are denoted in the image to give further description of the array.

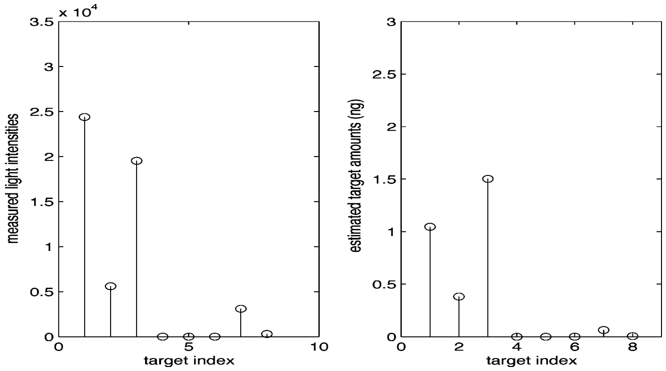


Fig. 9. Signal measured by the probe set $\mathcal{P}_1$ and the corresponding target amount estimates.

conditions of these experiments were the same as the conditions of the calibration experiments.

Fig. 8 shows the scanned image. In Fig. 9, we show the measured signal and the estimated values of the targets using the probe set $\mathcal{P}_1$. Similarly, Figs. 10–12 show the measured signals and the estimated values of the targets using the probe sets $\mathcal{P}_2 - \mathcal{P}_4$, respectively.

These figures show several interesting features. A direct readout of the signal obtained from Fig. 9 for probe set $\mathcal{P}_1$ might lead an observer to conclude that there is more target r_1 than target r_3 in the applied mixture. However, the constrained least-squares algorithm corrects this, recovers the true relation of the targets in the mixture, and gives fairly accurate estimates of their quantities. It also suppresses the cross-hybridization artifact that may lead one to erroneously believe in the presence of r_7 in the mixture. Similar remarks apply to Fig. 10 (obtained from the probe set $\mathcal{P}_2$) where the algorithm correctly estimates the presence of target r_2, even though this barely evident from the direct readout.

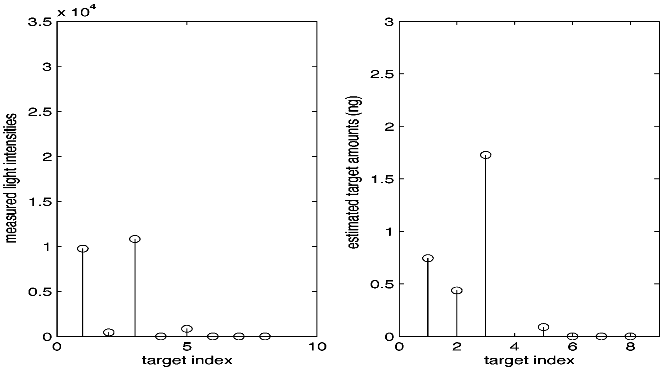


Fig. 10. Signal measured by the probe set $\mathcal{P}_2$ and the corresponding target amount estimates.

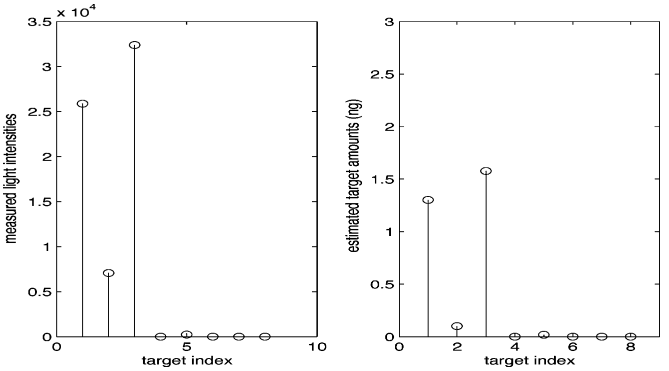


Fig. 11. Signal measured by the probe set $\mathcal{P}_3$ and the corresponding target amount estimates.

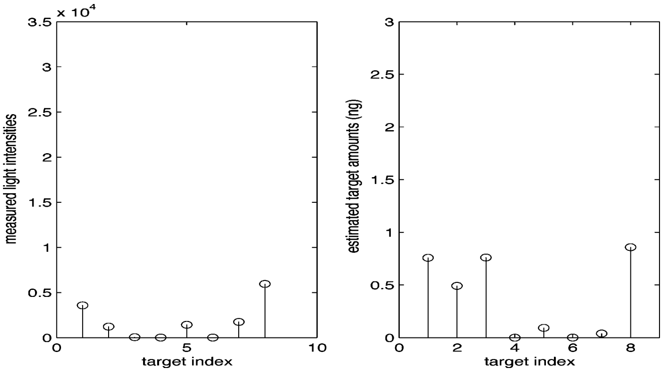


Fig. 12. Signal measured by the probe set $\mathcal{P}_4$ and the corresponding target amount estimates.

The presence of the biological background seems to most adversely affect the results of Figs. 11 and 12 obtained from probe sets $\mathcal{P}_3$ and $\mathcal{P}_4$, respectively. In particular, in Fig. 12, the algorithm incorrectly identifies the presence of target r_8. However, when all four sets of probes are used for estimation, the targets are estimated quite precisely, as indicated by Fig. 13.

The results thus demonstrate that our algorithm is fairly robust with respect to the presence of rich biological background. Moreover, it should be surely possible to further improve the performance of the algorithm by incorporating the presence of a biological background into our statistical model.

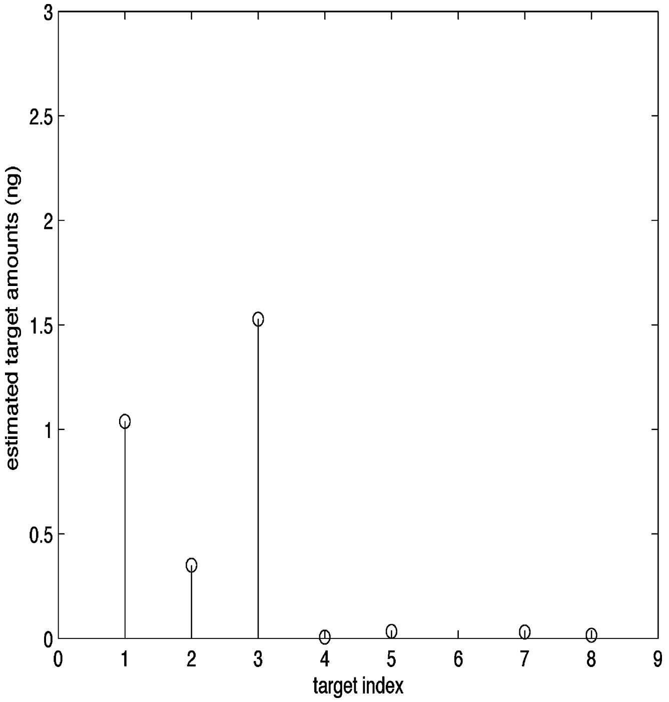


Fig. 13. Target amounts estimated using all probes.

VI. Summary and Conclusion

We developed a statistical model for DNA microarrays based on a probabilistic description of the hybridization and cross-hybridization processes. In particular, when the target concentrations are not too high, or if the number of probes per site is not too low—so that saturation does not occur—we show a linear relationship between the unknown target concentrations and the measured light intensities. This linear relationship is perturbed by additive white Gaussian noise consisting of two components, one of which has a variance proportional to the number of targets (and hence is shot noise). The shot noise nature of the noise in DNA microarrays has been earlier observed experimentally [6].

The statistical model can be fully described by knowing the probability of different targets binding to different probes. Though these probabilities can be somewhat estimated based on the target and probe sequences, e.g., using the concepts of ΔG and melting temperature (see, e.g., [12]), it appears that one needs some sort of calibration experiments to estimate them more accurately. Therefore, in its current form, our method is best suited to low-density arrays where the number of spots is not too large so that the number of calibration experiments is not too prohibitive. Fortunately, there are many applications for such arrays in diagnostics, SNP detection, toxicology, etc. (see, e.g., [18]–[21] and the references therein). Of course, it would be very interesting to see whether our method can be scaled to high-density arrays (with many thousands of spots) by coming up with models that require only a few calibration experiments.

In any event, once the probabilistic framework is in place, one may use a variety of statistical methods to estimate the target concentrations and we described ML, MAP, and constrained least-squares estimation. We also determined the Cramer–Rao bounds for estimation in DNA microarrays.

Our proposed algorithm differs from current methods (see, e.g., [4]) in that, rather than treating cross hybridization as noise, it views it as interference and does estimation while taking it into account. In fact, some preliminary studies of the Cramer–Rao bounds suggest that cross hybridization may, in fact, be beneficial. In particular, if we have only a few target types present in the sample (as is often the case in diagnostic applications), the existence of cross hybridization can lead to more accurate estimates of the target concentrations, simply because there are more sites where the targets can bind, thus increasing the signal strength. This result may have ramifications for probe design. [Currently, all probe design is based on minimizing the amount of cross hybridization (see, e.g., [17] and the references therein).]

Two sets of experiments were designed and performed, that confirmed the validity of the proposed model and the efficacy of the estimation techniques. The experiments included an example with a sample consisting of two oligonucleotide targets where existing techniques would detect the presence of four targets (the extra detected targets being an artifact of cross hybridization). Our algorithm, on the other hand, correctly detects only two targets and estimates their concentrations to remarkable accuracy. Results of a similar flavor were obtained in experiments with cDNA targets in the presence of a complex biological background.

The work described in this paper can be extended in several ways. One is to generalize the model to the case where the target concentrations are high and saturation at the probes may occur. This would require modeling the probability of binding to different probes as a function of the number of targets that are already bound to the probes. This will undoubtedly make the model nonlinear and will require modifications to the estimation algorithm. Another direction would be to study ways to more accurately compute the probabilities of various targets and probes binding (including possible local dependencies). As mentioned earlier, this may allow the method to scale to high-density arrays where extensive experimental calibration is not feasible. Also, a study of the robustness of the estimation algorithms to uncertainties in the statistical model should be useful. Our methods seemed fairly robust to the presence of a complex biological background. Nonetheless, one of the most interesting generalizations will be to study, both theoretically and experimentally, the proposed statistical framework for detection and quantification of targets in a complex biological background.

References

[1] M. Shena, *Microarray Analysis*. New York: Wiley, 2003.

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

[3] A. Hassibi, S. Zahedi, R. Navid, R. W. Dutton, and T. H. Lee, "Biological shot-noise and quantum-limited signal-to-noise ratio in affinity-based biosensors," *J. Appl. Phys.*, vol. 97, no. 084701, 2005.

[4] W. Zhang and I. Shmulevich, Eds., *Computational and Statistical Approaches to Genomics*. Norwell, MA: Kluwer, 2002.

[5] A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini, "Universal DNA tag systems: a combinatorial design scheme," in *Proc. 4th Annu. Int. Conf. Computational Molecular Biology*, Tokyo, Japan, 2000, pp. 65–75.

[6] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proc. Nat. Acad. Sci. (PNAS)*, pp. 14 031–14 036, Oct. 29, 2002.

[7] A. Hassibi and H. Vikalo, "A probabilistic model for inherent noise and systematic errors of microarrays," presented at the IEEE Int. Workshop Genomic Signal Processing Statistics, Newport, RI, May 22–24, 2005.

[8] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays," *Proc. Nat. Acad. Sci. (PNAS)*, vol. 98, pp. 31–36, 2001.

[9] G. A. Held, G. Grinstein, and Y. Tu, "Modeling of DNA microarray data by using physical properties of hybridization," *Proc. Nat. Acad. Sci. (PNAS)*, vol. 100, pp. 7575–7580, Jun. 2003.

[10] H. Vikalo, A. Hassibi, and B. Hassibi, "Optimal estimation of gene expression levels in microarrays," presented at the IEEE Int. Workshop Genomic Signal Processing Statistics, Newport, RI, May 22–24, 2005.

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

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

[13] H. C. Berg, *Random Walks in Biology*. Princeton, NJ: Princeton Univ. Press, 1993.

[14] B. J. Frey, Q. D. Morris, W. Zhang, N. Mohammad, and T. R. Hughes, "GenRate: A generative model that finds and scores genes by jointly accounting for the expression and genomic arrangement of putative exons," in *Proc. Pacific Symp. Biocomputing*, Jan. 2005, pp. 495–506.

[15] T. F. Coleman and Y. Li, "A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM J. Optim.*, vol. 6, no. 4, pp. 1040–1058, 1996.

[16] H. Cramer, *Mathematical Models of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.

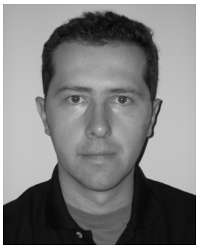
[17] L. Kaderali and A. Schliep, "An algorithm to select target specific probes for DNA chips," *Bioinformatics*, vol. 18, no. 10, pp. 1340–1349, 2002.

[18] F. de Longueville *et al.*, "Molecular characterization of breast cancer cell lines by a low density microarray," *Int. J. Oncology*, vol. 27, pp. 881–892, 2005.

[19] A. Rangel Lopez, "Low-density DNA microarray for detection of most frequent missense point mutations," *BMC Biotechnol.*, vol. 8, no. 5, Feb. 15, 2005.

[20] F. de Longueville *et al.*, "Use of a low-density microarray for studying gene expression patterns induced by hepatoxicants on primary cultures of rat hepatocytes," *Toxicolog. Sci.*, vol. 75, pp. 378–392, 2003.

[21] A. Bouchie, "Shift anticipated in DNA microarray market," *Nature Biotechnol.*, vol. 20, no. 8, 2002.



Haris Vikalo was born in Tuzla, Bosnia and Herzegovina. He received the B.S. degree from the University of Zagreb, Croatia, in 1995, the M.S. degree from Lehigh University, Bethlehem, PA, in 1997, and the Ph.D. degree from Stanford University, Stanford, CA, in 2003, all in electrical engineering.

In summer 1999, he held a short-term appointment at Bell Laboratories, Murray Hill, NJ. From January 2003 to July 2003, he was a Postdoctoral Researcher, and since July 2003, he has been an Associate Scientist at the California Institute of Technology, Pasadena. His research interests include genomic signal and information processing, wireless communications, estimation, and statistical signal processing.

Babak Hassibi was born in Tehran, Iran, in 1967. He received the B.S. degree from the University of Tehran, Iran, in 1989 and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 1993 and 1996, respectively, all in electrical engineering.

From October 1996 to October 1998, he was a Research Associate at the Information Systems Laboratory, Stanford University, and from November 1998 to December 2000, he was a Member of the Technical Staff in the Mathematical Sciences Research Center at Bell Laboratories, Murray Hill, NJ. Since January 2001, he has been with the Department of Electrical Engineering at the California Institute of Technology, Pasadena, where he is currently an Associate Professor. He has also held short-term appointments at Ricoh California Research Center, the Indian Institute of Science, and Linkoping University, Sweden. His research interests include wireless communications, robust estimation and control, adaptive signal processing, and linear algebra. He is the coauthor of the books *Indefinite Quadratic Estimation and Control: A Unified Approach to H^2 and H^∞ Theories* (New York: SIAM, 1999) and *Linear Estimation* (Englewood Cliffs, NJ: Prentice-Hall, 2000).

Dr. Hassibi is a recipient of an Alborz Foundation Fellowship, the 1999 O. Hugo Schuck Best Paper Award of the American Automatic Control Council, the 2002 National Science Foundation Career Award, the 2002 Okawa Foundation Research Grant for Information and Telecommunications, the 2003 David and Lucille Packard Fellowship for Science and Engineering, and the 2003 Presidential Early Career Award for Scientists and Engineers (PECASE). He has been a Guest Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY Special Issue on Space–Time Transmission, Reception, Coding and Signal Processing and is currently an Associate Editor for Communications of the IEEE TRANSACTIONS ON INFORMATION THEORY.

Arjang Hassibi (S'01–M'02) received the B.S. degree from the University of Tehran, Iran, in 1997 and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 2001 and 2005, respectively, all in electrical engineering.

Currently, he is a Postdoctoral Scholar at California Institute of Technology, Pasadena. His main areas of research are biosensor and bioelectronics, integrated biomedical systems, and biosensor modeling. He is also a cofounder of Xagros Technologies, Inc., Mountain View, CA, which is a molecular diagnostics company. He has also held research positions at Barcelona Design, Mountain View, CA, in 2000 and Panorama Research Institute, Sunnyvale, CA, between 2003 and 2005.