14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

ON PROBE-LEVEL INTERFERENCE AND NOISE MODELING IN GENE EXPRESSION MICROARRAY EXPERIMENTS

Paul G. Flikkema

Control Engineering Laboratory
Helsinki University of Technology, Espoo 02015 Finland
phone: +358 9 451 5241, fax: +358 9 451 5208, email: paul.flikkema@tkk.fi

ABSTRACT

This paper describes a signal processing model of gene expression microarray experiments using oligonucleotide technologies. The objective is to estimate the expression transcript concentrations modeled as an analog signal vector. This vector is received via a cascade of two noisy channels that model noise (uncertainty) before, during, and after hybridization. The second channel is also mixing since transcript-probe hybridization is not perfectly specific. The gene expression levels are estimated based on a second-order statistical model that incorporates biological, sample preparation, hybridization, and optical detection noises. A key feature is the explicit modeling of gene-specific and non-specific hybridization in which both have deterministic and random components. The model is applied to the processing of probe pairs as used in Affymetrix arrays, and comparison of currently used methods with the optimum Gauss-Markov estimator. In general, the estimation performance is a function of the hybridization noise characteristics, probe set design and number of experimental replicates, with implications for integrated design of the experimental process.

1. INTRODUCTION

DNA array technologies enable the simultaneous estimation of the expression levels of multiple genes in biological tissues. Now about 10 years old, microarrays can measure the expression of tens of thousands of genes and have become the central measurement tool for experimental research in disciplines ranging from systems biology to drug discovery. Every living cell produces mRNA (messenger RNA) coded by the DNA of its genes in the process of transcription; this mRNA is then translated into proteins which orchestrate the myriad functions of the cell.

In a microarray experiment, the mRNA transcripts are extracted from a biological sample and converted through a series of steps to complementary RNA (cRNA) that is washed over the microarray chip. The chip has numerous probe sites (often called spots or wells), each containing a DNA sequence associated with a particular gene. The key step is called hybridization, wherein probes bind preferentially with their target transcripts to form double-helix duplexes. The cRNA molecules are labeled with a reporter molecule that fluoresces upon illumination with a laser from a scanning confocal microscope; a probe's fluorescence signal is an increasing function of the amount of cRNA that has become chemically bound, or hybridized, to the probe's DNA. The optical signals are read by the microscope and processed to form estimates of gene activity. Every step of this experimental process can introduce significant noise, and the expense of microarray assays implies small sample sizes (often only three).

There are several approaches for synthesis of the DNA for a gene and deposition at a site on the microarray. One of the more popular techniques is based on depositing short (20 to 60 nucleotide) DNA sequences at each probe. A gene is associated with multiple probes, and the processing combines the optical signals from the probeset—all the probes for a gene—to compute the expression level estimate. At least two types of arrays are in mass production: inkjet-spotted arrays (Agilent) and chemically-deposited arrays (Affymetrix). Recently, researchers have prototyped an ink-jet device for in-laboratory synthesis of custom arrays [1], greatly expanding possibilities for the design of probes, probe sets, and arrays that are optimized for specific experiments.

Numerous approaches have been developed for the processing of the data, but most efforts have been characterized by the construction and linking of various ad hoc statistical approaches. Roughly, three separate steps are performed [2]. First, background/signal adjustment is performed to remove optical scanning artifacts and hybridization noise. The second step is called normalization and attempts to remove systematic variations between arrays. The final "probe summarization" step combines the resulting probe signals for a gene to estimate its expression level. Due to the widespread use of Affymetrix arrays, most efforts have focused on averages of probe signals or certain differences [3]. A signal processing framework similar in spirit to the one in this paper uses a binary matrix to model the presence or absence of oligos to study the design of arrays with with multi-target probes [4]. A mixed linear model that emphasizes differential experimental effects (e.g., cell lines and treatments) is presented in [5]; however, transcript/probe interaction effects are considered fixed, and random effects are assumed normal. Other related work includes stochastic molecular models of the hybridization analysis for cDNA arrays [6], and iterative methods for joint estimation of hybridization parameters and expression levels [7, 8]. The latter techniques are based on blind models; however, most experiments are sample-starved, and significant progress is being made in the determination of hybridization parameters from theory and controlled experiments [9, 10].

In this paper, we develop an integrated signal processing model that captures key aspects of the process from experiment to transcript level estimates, with focus on characterization of the signal of interest and sources of noise and interference. Our experiment model admits the coloring of bio-

This work was funded by the Helsinki University of Technology and a grant from the Northern Arizona University Strategic Alliance for Bioscience Research and Education.

logical noise by the hybridization. The hybridization model explicitly captures deterministic and random components of binding affinity and cross-talk between probe signals; from this, we show that the performance of several existing approaches is unnecessarily sensitive to cross-hybridization effects. We also describe a simple non-parametric estimator that does not require inference of distributions and provides reliability information that can exploited in higher-level inference tools, e.g., for the inference of network structure. The next section describes the model. Optimum processing of the data is outlined in Section 3, followed by its application to processing of probe pairs from Affymetrix GeneChip arrays, and concluding remarks.

2. OLIGO MICROARRAY EXPRESSION EXPERIMENT MODEL

Let the biological sample contain target transcripts k, $k = 1,\ldots,K$ with molar concentrations $\{x_k\}_{k=1}^{K}$. Here K is the number of targets, or genes to be considered. The goal is to estimate the x_k's from the optical signals when the array is scanned.

2.1 Biological Noise

The first measurement uncertainty is biological noise. This noise is significant in microarray experiments [11], and arises from several sources. Genotypic noise is caused by genomic variations within the sampling population, (e.g., if the sample uses littermates), while variation among genetically identical individuals is called phenotypic noise. One source of phenotypic noise is alternative splicing, in which pre-mRNA is edited by cutting and pasting operations after transcription but before translation. Other noise is due to differences (of up to 35%) in response that may be caused by variations in a cell's dynamically available resources [12]. These noises can vary from target to target, depending on, for example, spatial variation of expression in the sample. They are normally not separable in microarray experiments, but their characterization is an active research area spurred by new techniques to measure responses of individual cells [13]. Their sum is modeled as a zero-mean random variable β_k for each target transcript. Hence the signal plus biological noise for target k is $x_k + \beta_k$. Clearly, these noises will be correlated between genes (e.g., in operons); this is modeled with a covariance matrix Σ_β.

Samples are prepared using RNA extraction, reverse transcription to cDNA, transcription to fluorescence-labeled cRNA, and preparation of the hybridization cocktail. The accompanying sample preparation noise appears to be small compared with other noises [14], and will be neglected here for convenience. However, these procedures may also add bias, resulting in inter-chip variations. Other inter-chip differences may result from differences an array manufacturing or processing. Thus there has been a large amount of effort in normalization to allow accurate comparisons between arrays [15], from simple mean adjustments to distribution-matching via quantile equalization. Moreover, the use of controls (transcripts with complementary probes that are spiked into the cocktail in controlled concentrations) can further quantify these variations. In the following, we will assume that such normalization has been performed.

2.2 Hybridization

In oligonucleotide arrays, each target transcript is specified by a set of probe molecules corresponding to the gene, with each probe molecule a string of 20 to 60 nucleotides. For example, in the Affymetrix GeneChip design, probes are 25 nucleotides long, and there are 32 to 40 probes per target transcript. Each probe on the chip consists of hundreds or thousands of identical nucleotide sequences.

The log signal present at probe i of probeset j due to signal from transcript k, caused by the binding of a transcript fragment to the probe to form a duplex sequence, can be modeled as a bilinear function of (1) the log component concentration up to saturation of the probe site and (2) the transcript-level hybridization affinity of the two sequences [16, 9, 4]. Thus the expression signal at probe j of probeset i due to gene k is $h(i,j;k)x_k$, where the coefficient $h(i,j;i)$ models what is often called gene-specific binding (GSB). The hybridization affinity is a non-trivial function of the length and composition of the probe sequences and the fluorescent labeling of the cRNA [17, 18, 9]. In practice, these parameters can be estimated using sequence-based methods, or computations based on the free energy changes of duplex formation [10, 7]. With these definitions, the gene-specific signal at probe (i,j) can be written as $s_{ij} = h(i,j;i)[x_i + \beta_i]$.

Hybridization is subject to errors from at least two sources. First, in oligo arrays the cRNA is chemically fragmented into oligo segments to provide better affinity to the probes, since the binding of full-length transcripts to the attached probe oligos would be very inefficient. The fragmentation results in the cutting of the transcripts at random locations. The second source is binding of non-target transcript fragments, referred to as non-specific binding. We model these errors using a deterministic and random component for each probe signal. The deterministic component captures inter-transcript interference (ITI), the contributions to the signal via $h(i,j;k)x_k$, $k \neq i$ from all non-target transcripts in the sample. Both gene-specific binding and ITI are noisy, so the hybridization noise is modeled as an additive random vector η (whose number of elements is equal to the total number of probes) with covariance matrix Σ_η.

2.3 Scanning and Chip-Induced Effects

The final component of the model is optical noise due to laser scanning and intensity measurement. It is considered additive and zero-mean, and can be effectively removed by background adjustment [19]. Moreover, it is a trivial extension of the model and will subsequently be absorbed into η, so the complete model is (1).

2.4 Experiment Model

At this point, the composite signal at probe (i,j) can be written as

$$s_{ij} = \sum_{k=1}^{K} h(i,j;k)[x_k + \beta_k] + \eta_{ij}, \qquad (1)$$

or in matrix form (stacking probe sets and probes within sets) as

$$s = H[x + \beta] + \eta, \qquad (2)$$

where we call H the hybridization matrix.

Summarizing, the signal vector x is first corrupted by biological noise. This is followed by hybridization, where the

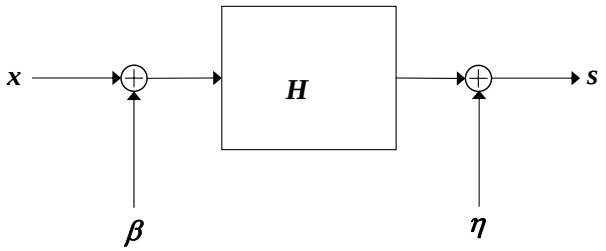

Fig. 1. Block diagram of microarray experiment model. The transcript expression vector (signal) vector x is corrupted by biological noise β, followed by hybridization, with mixing by the hybridization matrix H and corruption by η.

signal and biological noise components are mixed, and hybridization noise is added (Figure 1). Hence the model is structured as two cascaded communication channels: the first adds noise, while the second also causes interference. Note that the ITI for probe j of target gene i is $\sum_{k\neq i} h(i,j;k)[x_k + \beta_k] + \eta_{ij}$.

Due to the low signal-to-noise ratio of microarray experiments, experimental protocols call for multiple replicates of the biological samples. Hence the signal for probe set i of replicate r is the vector

$$s_i^{(r)} = H_i[(x^{(r)} + \beta_i^{(r)}] + \eta_i^{(r)}.$$

The composite signal for all probe sets and all replicates can be formed from stacking the outputs, again yielding (1).

2.5 Hybridization Model Inference

In microarray experiments, the expression measurement of over 10,000 genes is not uncommon, making experimental determination of $h(i,j;k)$ for every probe in every probe set for every gene impossible. However, the developing theory of nucleotide duplex formation may provide theoretical models. In addition, genes that cause significant inter-transcript interference can be found using BLAST searches [7]. Both considerations motivate a modified model. The set of known genes $\mathscr{G}$ is partitioned into two sets: $\mathscr{G}^{(1)}$, composed of genes of interest and those causing them significant ITI, and $\mathscr{G}^{(2)} = \mathscr{G} - \mathscr{G}^{(1)}$, which includes all other genes (including unknown genes). This imposes corresponding partitions of $x = [x^{(1)\prime} : x^{(2)\prime}]'$ (where $(\cdot)'$ denotes transpose), $\beta = [\beta^{(1)\prime} : \beta^{(2)\prime}]'$, and $H = [H^{(1)} : H^{(2)}]$, so that

$$s = \sum_{l=1}^{2} H^{(l)}[x^{(l)} + \beta^{(l)}] + \eta.$$

Consider target gene i. With $h_i^{(1)}$ denoting column i of $H^{(1)}$, the signal is

$$s_i = h_i^{(1)} x_i^{(1)} + \sum_{k\neq i} h_k^{(1)} x_k^{(1)} + \gamma + \eta,$$

where first term is the signal from the target, the second term is the explicitly modeled fixed ITI, and γ models the combined effect of the remaining ITI and biological noises.

3. EXPRESSION LEVEL ESTIMATION

The (log) linear model (2) can be simplified. By collecting the (colored) noise as $n = H\beta + \eta$ with mean zero and covariance matrix $\Sigma_n = H\Sigma_\beta H' + \Sigma_\beta$, the model becomes

$$s = Hx + n.$$

Notice that no assumptions have been made on the distribution of either the signal or the noise. This is in contrast to parametric approaches that assume normal [5], log-normal [15], or gamma [16] distributions. In practice, x is often treated as deterministic unknown; in this case the optimum estimator (in terms of the estimation error variance for each transcript) is the Gauss-Markov (generalized least-squares) estimator

$$\hat{x}_{GM} = \Sigma_{GM} H' \Sigma_n^{-1} s,$$

which is unbiased with error covariance

$$\Sigma_{GM} = (H'\Sigma_n^{-1} H)^{-1}. \tag{3}$$

This estimator only requires knowledge of first and second moments, which are easier to estimate from controlled experiments than joint distributions. Note that one is free to fit any distribution to these statistics in subsequent processing, if desired.

4. OPTIMUM PROBE-PAIR PROCESSING

Probe sets in the Affymetrix GeneChip arrays consist of 16-20 pairs of probes; each pair consists of a PM probe that is perfectly matched to a nucleotide sequence found in the target gene, and a MM (mismatch) probe that differs from the PM probe at one site in the middle of the sequence. The MM probe was designed to provide a measure of non-specific binding signal that could be subtracted from the PM signal [19]. However, it is common for the MM signal to exceed the PM signal in a significant number of probe pairs, so a number of techniques have been developed that use only the PM signal (see the review in [3]).

To explore this, we simplify the model (1) to a single probe pair. In correspondence with other models, we ignore biological noise and include ITI in the hybridization noise, whose covariance matrix is

$$\Sigma_\eta = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{4}$$

where σ^2 is the hybridization noise variance and ρ is the cross-hybridization noise correlation coefficient. This model of hybridization noise is consistent with results [18, 2] demonstrating that cross-hybridization is noisy, but is neither uncorrelated nor identical between probes. We note that for convenience we have assumed that the hybridization noise variances are equal; however, all results can be generalized to arbitrary variances. For example, a probe's noise variance could be a function of its specific and non-specific binding affinities.

Letting the hybridization affinities for the PM and MM probes be h_1 and h_2 respectively, the probe pair signal is

$$s = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} x + \eta. \tag{5}$$

From (3), the error variance of Gauss-Markov estimator $\hat{x}_{GM}$ is

$$\sigma^2_{GM} = \frac{\sigma^2(1-\rho^2)}{h_1^2+h_2^2-2\rho h_1 h_2}. \quad (6)$$

The PM and PM $-$ MM signals are $h_1 x+\eta$ and $(h_1-h_2)x+\eta_1-\eta_2$, respectively, so

$$\sigma^2_P = \frac{\sigma^2}{h_1^2}, \quad \sigma^2_{P-M} = \frac{2\sigma^2(1-\rho)}{(h_1-h_2)^2}, \quad (7)$$

for the unbiased PM and PM $-$ MM estimators $\hat{x}_P$, $\hat{x}_{P-M}$.

We use the unbiased versions of all three estimators; this requires that the hybridization affinities and the hybridization noise covariance matrix are known from prior controlled experiments. Their error variances are compared as a function of ρ in Figure 2. In the figure, we have set the ratio of binding affinities $h_2/h_1 = 0.7$ based on experimental results in [19, 16]. Note that this ratio can be considered the relative binding efficiency of the mismatch and perfect-match probes. Results are normalized ($\sigma^2_P = 1$), and the noise power is kept constant and equal for both probes, so that increasing noise coherence results in greater squared-sum noise power.

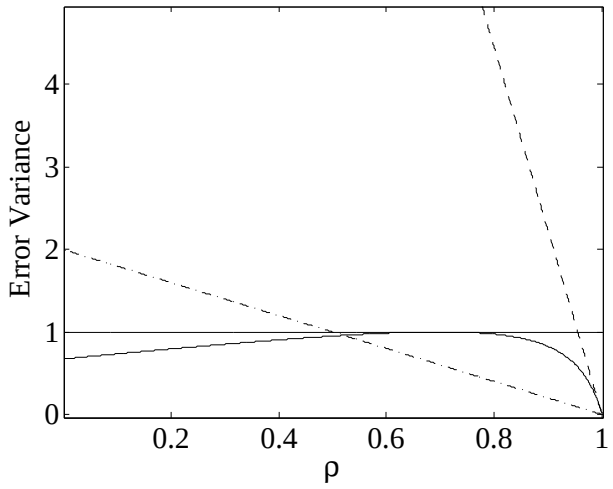

Fig. 2. Error variance performance of probe-pair processing estimators as a function of cross-hybridization correlation ρ: unbiased perfect match $\hat{x}_P$ (solid horizontal line); Gauss-Markov $\hat{x}_{GM}$ (solid curve); unbiased PM $-$ MM $\hat{x}_{P-M}$ (dashed); biased PM $-$ MM $\hat{x}_{B:(P-M)}$ (dash-dot).

Estimators based on PM $-$ MM implicitly assume perfect noise coherence and thus have large error variance for all but large values of ρ. For example, the model of [20] is (using our nomenclature) $s_i = r + h_i x + n_i$ for the PM ($i=1$) and MM ($i=2$) probes, where r is a baseline response. The resulting PM $-$ MM signal is

$$(h_1-h_2)x+\varepsilon,$$

where $\varepsilon = n_1 - n_2$ is re-coined as a single random variable so that hybridization noise covariance effects are not accounted for. The relative superiority of $\hat{x}_P$ over $\hat{x}_{P-M}$ for low cross-hybridization covariances as seen from Figure 2 may explain why PM-only summaries (e.g., [21]) have been preferred for low expression levels. However, PM-only summaries discard

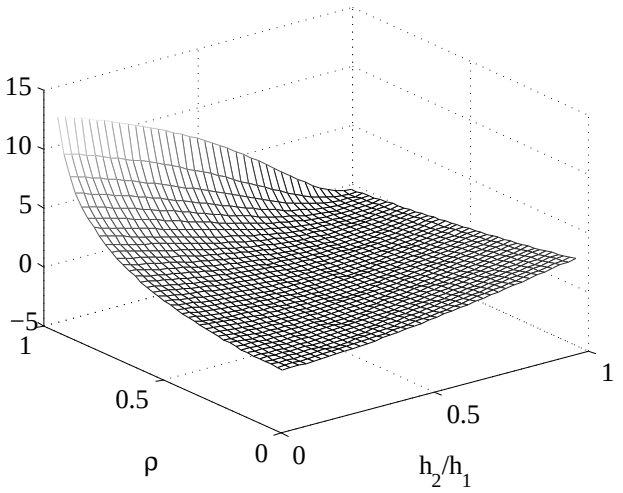

Fig. 3. Gain in dB of Gauss-Markov estimator vs. unbiased PM estimator as a function of h_2/h_1 and ρ.

useful information from the MM probe that is helpful at low and high noise coherence levels. From a signal processing perspective, both the PM $-$ MM and PM estimators are deficient. In contrast, the Gauss-Markov estimator optimally combines both probe signals for any value of noise coherence. By treating the PM and MM signals as equals (except for differences in affinity), it also handles cases where the the MM signal exceeds the PM signal. The cross-hybridization noise correlation coefficient must be known, but only up to its second-order statistics.

Figure 2 also shows that the performance of the biased PM $-$ MM estimator $\hat{x}_{B:(P-M)}$ is significantly better than that for $\hat{x}_{(P-M)}$. This variance/bias (precision/accuracy) tradeoff, often used to qualitatively compare probe summarization approaches, is now simply a function of the parameterized probe pair model (4),(5).

Figure 3 shows $\frac{\sigma^2_P}{\sigma^2_{GM}}$, the gain in precision (inverse error variance) in dB of the Gauss-Markov estimator relative to the unbiased PM estimator. In particular, it shows the gain for two special cases of interest. First, the gain in precision over when only the PM probe is used is given by (from (6),(7))

$$h_2 = 0: \quad \frac{\sigma^2_P}{\sigma^2_{GM}} = \frac{1}{(1-\rho^2)},$$

showing how PM-only designs fail to exploit knowledge of the cross-hybridization noise correlation. Another case of interest is when $h_2 = h_1$, i.e., when both probes have the same hybridization affinities. In this case we have

$$h_2 = h_1: \quad \frac{\sigma^2_P}{\sigma^2_{GM}} = \frac{2}{(1+\rho)},$$

demonstrating that the Gauss-Markov estimator has a gain of up to 3 dB. This quantifies the true cost of the combination of current Affymetrix probe-pair designs and PM-only summaries. Specifically, neglecting 50% of the probes–the MM probes—in Affymetrix chips costs up to 3 dB in error variance.

5. CONCLUSIONS

The primary aim of this paper is to describe a signal processing model that captures complete microarray experiments. The introduction of a statistical model for biological noise as well as hybridization and cross-hybridization effects places the comparison of different approaches on a unified foundation, and the explicit modeling of biological noise may help explain experimental results. The hybridization matrix provides the connection between the probe/array design and the performance of target transcript estimation. The ability to jointly characterize deterministic and random hybridization effects, along with new custom oligo array technologies [1], opens the door to new approaches for joint experiment and array design [4]. For example, instead of using a fixed PM/MM pair design for all probes, probesets could be designed to achieve error variance specifications that vary depending on interest in the target transcript.

REFERENCES

[1] C. Lausted et al., "POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer," *Genome Biology*, vol. 5, July 2001.

[2] S. C. Choe et al., "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset," *Genome Biology*, vol. 6, 2005.

[3] R. Irizarry et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, 2003.

[4] I. Shmulevich et al., "Data extraction from composite oligonucleotide microarrays," *Nucleic Acids Research*, vol. 31, 2003.

[5] T. Chu, B. Weir, and R. Wolfinger, "A systematic statistical linear modeling approach to oligonucleotide array experiments:35-51," *Math. Biosci.*, vol. 176, pp. 35–51, 2002.

[6] H. Vikalo, A. Hassibi, and B. Hassibi, "Optimal estimation of gene expression levels in microarrays," in *GENSIPS 2005*, 2005.

[7] J. C. Huang, Q. D. Morris, T. R. Hughes, and B. J. Frey, "GenXHC: a probabilistic generative model for cross-hybridization compensation in high-density genome-wide microarray data," *Bioinformatics*, vol. 21, pp. i222–i231, 2005.

[8] D. Dueck et al., "Iterative analysis of microarray data," in *Proc. Forty-Second Allerton Conf. on Comm., Control and Comp.*, 2004.

[9] R. Mei et al., "Probe selection for high-density oligonucleotide arrays," *PNAS*, vol. 100, pp. 11237–11242, Sept. 2003.

[10] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Research*, vol. 33, 2005.

[11] J. McClintick et al., "Reproducibility of oligonucleotide arrays using small samples," *BMC Genomics*, vol. 4, 2003.

[12] M. Elowitz et al., "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183–1186, 2002.

[13] A. Colman-Lerner et al., "Regulated cell-to-cell variation in a cell-fate decision system," *Nature*, vol. 437, pp. 699–706, 2005.

[14] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *PNAS*, vol. 99, pp. 14031–6, Oct. 2002.

[15] Z. Wu and R. Irizarry, "A statistical framework for the analysis of microarray probe-level data," JHU Biostatistics Working Paper 73, 2005.

[16] X. Liu, M. Milo, N. Lawrence, and M. Rattray, "A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips," *Bioinformatics*, vol. 21, pp. 3637–3644, 2005.

[17] A. Relogio et al., "Optimization of oligonucleotide-based DNA microarrays," *Nucleic Acids Research*, vol. 30, 2002.

[18] F. Naef and M. O. Magnasco, "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays," *Phys. Rev. E*, vol. 68, 2003.

[19] Z. Wu and R. Irizarry, "Stochastic models inspired by hybridization theory for short oligonucleotide arrays," in *Proc. RECOMB 2004*, 2004.

[20] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *PNAS*, vol. 98, pp. 31–36, Jan. 2001.

[21] R. Irizarry et al., "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research*, vol. 31, 2003.