A PROBABILISTIC MODEL FOR INHERENT NOISE AND SYSTEMATIC ERRORS OF MICROARRAYS

Arjang Hassibi[a] *and Haris Vikalo*[b]

[a] Department of Electrical Engineering, Stanford University, CA 94305, USA
[b] Department of Electrical Engineering, California Institute of Technology, CA 91125, USA

***Abstract* — A probabilistic model for the measurement noise in microarray systems is presented. This model includes the inherent Poisson noise of the assay, as well as the systematic errors which are typically introduced during microarray fabrication and detection processes. The model presented here formulates not only the uncertainty of the measured expression levels, but also the contribution of the each procedural step to the overall detection signal-to-noise ratio (SNR).**

I. INTRODUCTION

Gene expression microarrays measure the expression level of thousands of genes simultaneously, providing a massively-parallel affinity-based detection platform in life science research [1]. Unfortunately, the high level of uncertainty associated with each microarray experiment often obscures some of the important characteristics of the biological processes of interest. The expression level uncertainty in such systems, fundamentally originates from the probabilistic characteristics of the detection process, all the way from sample extraction and mRNA purification to hybridization and imaging [2,3]. Currently, there are various techniques which increase the accuracy and signal-to-noise ratio (SNR) of the estimated values [4]. Nonetheless, all techniques rely on either comparative methods [5], or mathematical algorithms which introduce confidence zones by excluding the *unreliable* data and outliers [6]. Independent of the method utilized, the degree in which the SNR is improved in both approaches is still limited by the inherent microarray noise.

In this study we investigate and further model the underlying biochemical and fabrication noise sources which limit the SNR in microarrays. The result of this study, can not only improve the efficiency of the estimation algorithms, but also give design insights into the setup of the different microarray-based experiments.

II. MODEL

A. Systematic noise vs. inherent noise

We define systematic errors as the unwanted deviations from the intended detection protocol. If these errors are accurately evaluated, in theory, they can be compensated by post experiment data processing. If not, they result in a particular type of measurement uncertainty, typically referred to as systematic noise. Examples of systematic noise sources are fluidic handling errors or non-uniformities in spotting the probes.

We define inherent noise of the detection system as the *unavoidable* uncertainties even with ideal detection where no systematic error exists. Inherent noise is basically inevitable since it originates from the stochastic nature of molecular-level interactions. Poisson noise sources in microarrays [2] and image sensor detection shot-noise [1] are examples of such noise sources.

B. Sample extraction and mRNA purification

In a typical microarray experiment, targets (m different mRNA molecules) are initially extracted from a sample and subsequently purified. Let K_E denote the fraction of the original volume which is extracted (i.e., the extracted volume is K_E times the original sample volume). Furthermore, assume that the original sample has n_i targets of the i^{th} mRNA type, and that the purification process has the yield of Y_p. Now we can model this procedure to be a *random deletion process*. Consequently, the number of the i^{th} targets obtained after extraction and purification, $X_{P,i}$, becomes a random variable with a Gaussian distribution when n_i is large (generally the case in typical microarray assays). This type of variation or noise is inherent to microarrays and conceptually very similar to *partition* noise (see Table I for details).

If the extracted volume has a systematic error with zero mean and standard deviation of σ_E, we can use *Burgess variance theorem* [7] to reassess the distribution of $X_{P,i}$ as shown in Table I.

C. Reverse Transcriptase (RT)

The goal of RT process is to generate cDNA copies from all different mRNA target molecules. The process is basically another random deletion process, similar to extraction, with yield (survival probability) of Y_{RT}, defined as the probability of creating a single cDNA molecule from individual mRNA molecules. We again employ Burgess variance theorem to find the mean and variance of the total number of cDNA molecules $X_{RT,i}$ generated from target i (see Table I).

Step	Expected Value	Variance
Extraction and Purification ($X_{P,i}$)	$\overline{X}_{P,i} = n_i K_E Y_P$	$\sigma^2_{P,i} = n_i \left[\begin{matrix} (\sigma_E K_E Y_P)^2 \\ + K_E Y_P (1 - K_E Y_P) \end{matrix} \right]$
Reverse Transcription ($X_{RT,i}$)	$\overline{X}_{RT,i} = \overline{X}_{P,i} Y_{RT}$	$\sigma^2_{RT,i} = \sigma^2_{P,i} Y_{RT} + \overline{X}_{P,i} Y_{RT} (1 - Y_{RT})$
Hybridization (X_i)	$\overline{X}_i = \sum_{k=1}^{m} q_{i,k} \overline{X}_{RT,k}$	$\sigma_i^2 = \sum_{k=1}^{m} q_{i,k}^2 \sigma^2_{RT,k} + \sum_{k=1}^{m} \overline{X}_{RT,k} q_{i,k} (1 - q_{i,k})$

Table I: Microarray noise formulations.

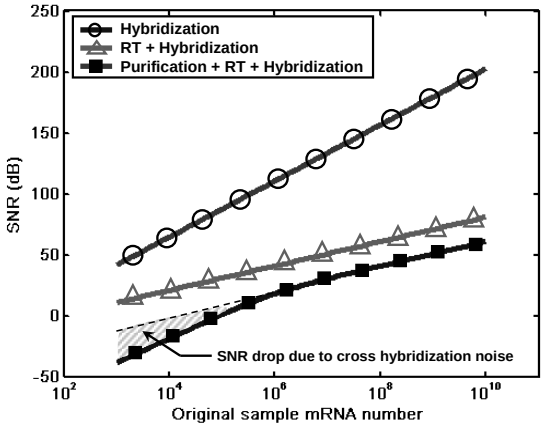

Figure 1: SNR with different detection procedures.

D. Array fabrication

Fabrication of the array involves spotting or synthesis of the ssDNA probes on a planar surface. The capturing process in affinity-based sensors (microarrays) is inherently probabilistic [8] where the statistics of the capturing process defines the specificity of the probe. In typical microarrays, the probe specificity, imperfect synthesis, and surface non-uniformities, all contribute to the nonspecific capturing events and systematic errors in capturing efficiency. To model these, we can define the capturing probability q_{ij} for the capturing of molecule j at location i. This basically results in a Markov model for the hybridization, where we can calculate the statistics of the captured targets at equilibrium [8].

E. Hybridization and cross-hybridization

It has been reported that the dominant measurement noise in microarrays is caused by the hybridization, which depends on the target expression level [2,8]. This Poisson-like noise indicates that the number of hybridized molecules varies due to the probabilistic nature of the hybridization. Beside the fluctuations of the number of specific bindings (i.e., the hybridized molecules), non-specific bindings (i.e., the process of cross-hybridization) may also take place which further degrade the certainty of the measurement. Therefore, at any time, a target particle may be *i*) hybridized to the perfectly complementary probe, *ii*) cross-hybridized to a partially complementary probe, or *iii*) free. It can be shown that the microarray measurement model, given all of the aforementioned noise processes, can be written as follows:

$$X = Qn + w, \qquad (1)$$

where n denotes the vector of quantities of the mRNA target molecules in the original sample, and X denotes the vector of captured cDNA molecules. Vector w is the vector of Gaussian noise whose variance depends upon n, i.e., w is a shot-noise that models signal fluctuations due to sample preparation, RT, hybridization, and cross-hybridization. Its variance is computed from the stationary distribution of the previously mentioned Markov chain. The components of each vector and coupling matrix $Q \in R^{m \times m}$ are listed in Table I.

Using the proposed framework, in Fig. 1, we plotted SNR as a function of the original mRNA number. In this example, K_E =0.1, σ_E =0.03, Y_p=0.6, and Y_{RT}=0.2. As evident from Fig. 1, the random deletion processes basically dominates the SNR degradation on high concentration levels assuming capturing probability is of 10^{-2} in hybridization, whereas in low concentrations cross-hybridization dominates SNR (probability is assumed to be 10^{-4} with 10^6 background targets). This specific result matches the microarray empirical noise datasets which were previously reported [1,2], demonstrating the validity of this modeling approach.

REFERENCES

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

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

[3] Tseng, G.C. *et al*, "Issues in cDNA microarray analysis: Quality Filtering, Channel Normalization, Model of Variations and Assessment of Gene Effects," *Nucleic Acid Research*, 12:2549-2557, 2001.

[4] Zhang, W., and I. Shmulevich (*editors*), c*omputational and Statistical Approaches to Genomics*, Kluwer Academic Publishers, 2002.

[5] Chen, Y., E.R. Dougherty, and M.L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Optics*, pp. 364-374, 1997.

[6] Li, C., and W.H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proc. Natl. Acad. Sci*., 98:31-36**,** 2001.

[7] R.E. Burgess, "Homophase and Hetrophase Fluctuations in Semiconducting Crystals," *Disc. Faraday Soc.* 28:151, 1959.

[8] Hassibi, A. *et al*, "Biological Shot-Noise and Quantum-Limited SNR in Affinity-Based Biosensor," (to appear) *J. Applied Physics*, 2005.