Optimal Estimation of Gene Expression Levels in Microarrays

Haris Vikalo[a], Arjang Hassibi[b] and Babak Hassibi[a]
[a]California Institute of Technology
[b]Stanford University

***Abstract*— Microarray technology relies on the hybridization process, which is stochastic in nature. However, current measurement and detection techniques do not fully exploit this stochastic nature nor do they consider it in data analysis. In this paper, we propose a probabilistic model of the DNA microarray and employ this model for optimal estimation of gene expression levels. Simulation results indicate significant improvement in the reliability of the estimates over the direct readout of the data.**

I. Introduction

Recently, high-throughput assay technologies have gained much attention in the genomic research community. DNA microarrays, in particular, have attracted much interest due to the large scale, parallel nature of the experiments, and the richness of the information that they provide. This stands in contrast to traditional techniques capable of analyzing only a small number of genes at a time. DNA microarrays [1] are primarily used to measure gene expression levels, i.e., the transcription of the DNA data into messenger RNA molecules (mRNA). DNA microarray technology is based on hybridization, a process in which complementary DNA strands specifically bind to each other. Typically, the surface of a DNA microarray contains a grid of different single stranded DNA oligonucleotide probes, whose locations are fixed during the process of hybridization and detection. The target mRNA that needs to be detected is first used to generate fluorescent labeled cDNA which is then applied to the microarray. The labeled cDNA molecules that are a perfect match to the microarray probes bind to the complementary oligos. However, there will be a number of non-specific bindings since cDNA may cross-hybridize to probes that are not a perfect match but rather only partial complements.

There has been a lot of work on employing statistical analysis tools for interpretation of microarray measurements (see [2] and the references therein). In this paper, we are interested in a more fundamental problem – the one of optimal estimation of the gene expressions. The number of hybridized molecules varies due to the probabilistic nature of the hybridization. This noise is Poisson-like at high expression levels, and more complex at low expression levels where non-specific binding becomes more significant [3]. We describe hybridization and cross-hybridization processes by Markov chains, similar to the techniques used in modeling affinity based sensors in [4]. Using the stationary distribution of the Markov chains, we formulate a statistical model of the microarray readout. The biological noise is modeled as the shot noise thus accounting for the inherent fluctuations of the measured signal. The detection problem is posed as the maximum likelihood optimization. Preliminary simulation results that we present indicate significant improvement over direct readout.

II. Model

We consider an $m \times m$ microarray, with m^2 types of oligonucleotide probes attached to its surface. A total of N molecules of n different types of cDNA targets, with concentrations $c_1, c_2, \ldots, c_n$, $\sum_{i=1}^{n} c_i = N$, are applied to the microarray. Our goal is to estimate c_i's from a scanned image of the array that gives information about the location and number of hybridized and cross-hybridized probes. The measurement is taken after the system reaches equilibrium.

For simplicity, assume that each cDNA fragment may hybridize to only one of the oligonucleotide probes, while it may engage in k non-specific bindings (cross-hybridizations). Diffusion movement of the unbound cDNA molecules is modeled as a random walk accross the array, and their distribution is assumed to be uniform on the array [4].

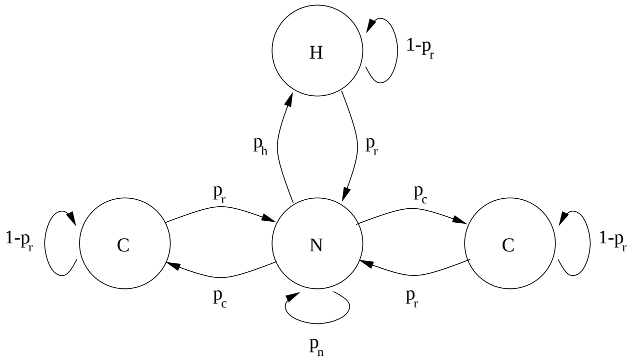

Fig. 1. **Markov chain modeling states of a target molecule on a microarray with one specific and $k = 2$ non-specific binding sites. The hybridized state is denoted by 'H', cross-hybridized states are denoted by 'C', the unbound state is denoted by 'N'.**

Let u_i denote the number of unbound molecules of type i, $i = 1, \ldots, n$. A cDNA in a close proximity to its matching probe oligonucleotide will hybridize with probability $p_{H,i}$. Therefore, a fraction of the unbound targets of type i that is being captured is $u_i p_{H,i}/m^2$. This fraction is constant at the equilibrium; however, at a given time instant, any particular molecule may be in a captured or in a released state. Therefore, the probability that a particular unbound target is going to be captured by its matching probe is $p_{h,i} = p_{H,i}/m^2$. Similarly, the probability that any particular target cDNA will cross-hybridize is $p_{c,i} = p_{C,i}/m^2$. The probability that a target is released is denoted by p_r. [For simplicity, we assume that all cross-hybridizations are equally likely and that the probability of release is the same for both specific and non-specific binding.] The Markov chain that models transitions between possible states of a target with one specific and $k = 2$ non-specific binding sites is shown in Figure 1. The probability $p_n = 1 - kp_c - p_h$ in Figure 1 denotes the likelihood that an unbound target remains free.

Let $\mu_i = [\mu_{i,1} \ \ \mu_{i,2} \ \ \ldots \ \ \mu_{i,k+2}]^T$ be a vector whose components are fractions of the total number of the targets of type i that are in one of the $k+2$ states of the Markov chain in Figure 1. In particular, let $\mu_{i,1}$ denote the number of hybridized molecules, $\mu_{i,j}$, $j = 2, \ldots, k+1$, denote the numbers of cross-hybridized molecules in each of the k non-specific binding sites, and let $\mu_{i,k+2}$ be the number of unbound molecules. In equilibrium, we are interested in finding the components of the vector μ_i such that $\mu_i = P_i \mu_i$, $\mathbf{1}^T \mu_i = c_i$, and thus

$$\mu_i = \left[\begin{array}{c} I - P_i \\ \mathbf{1}^T \end{array} \right]^{\dagger} \cdot \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ c_i \end{array} \right], P_i = \left[\begin{array}{cccc} 1 - p_r & \ldots & 0 & p_h \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 - p_r & p_c \\ p_r & \ldots & p_r & p_n \end{array} \right]$$

where $(\cdot)^\dagger$ denotes a pseudoinverse.

Every state in the Markov chain for the type i target corresponds to a probe to which some of the cDNA targets of type i may bind (except for the last state which collects remaining unbound molecules). Let $\mathcal{L}_i$ denote the set of indices l_j^i, $1 \leq l_j^i \leq m^2$, $1 \leq j \leq k+1$, that indicate those probes that are associated with one of the binding states in the Markov chain. In particular, let l_1^i denote the probe on the microarray to which target i binds specifically, and let $l_2^i, \ldots, l_{k+1}^i$ denote the probes on the microarrays to which target i binds non-specifically.

Let μ_i^{k+1} denote the vector comprising first $k+1$ components of μ_i, and let $\mathbf{q}_i^{k+1} = \mu_i^{k+1}/c_i$. Then its first component, $q_i^{k+1}(1)$, is the probability that a target molecule of type i is hybridized, while $q_i^{k+1}(j)$, $2 \leq j \leq k+1$ are the probabilities that a target molecule of type i is cross-hybridized. Define the $m^2 \times 1$ vector $\mathbf{q}_i$ such that

$$q_i(l_j^i) = \left\{ \begin{array}{c} q_i^{k+1}(j), l_j^i \in \mathcal{L}_i, \\ 0, \ \text{otherwise}. \end{array} \right.$$

Furthermore, define the matrix $Q = [\mathbf{q}_1 \ \mathbf{q}_2 \ \ldots \ \mathbf{q}_{m^2}]$. Then the microarray measurement model can be written as

$$\mathbf{s} = Q\mathbf{c} + \mathbf{w} + \mathbf{v}, \qquad (1)$$

where $\mathbf{c} = [c_1 \ \ \ldots \ \ c_{m^2}]^T$ is the $m^2 \times 1$ vector of input concentrations of cDNA target molecules, and $\mathbf{s}$ is the $m^2 \times 1$ vector of measured light intensities. Furthermore, $\mathbf{w}$ is the $m^2 \times 1$ vector that describes the inherent fluctuations in the measured signal. These fluctuations are due to the probabilistic nature of the hybridization process and depend on the signal intensity (essentially, $\mathbf{w}$ is the vector of shot-noise). In particular, if $q_{i,j} > 0$, then a target molecule of type j binds (whether specifically or non-specifically) to a probe of type i. This is a Bernoulli event whose variance is $q_{i,j}(1 - q_{i,j})$. Since $s_i = \sum_{j=1}^{m^2} q_{i,j} c_j$, the variance of signal fluctuations is

$$\sigma_{w,i}^2 = \sum_{j=1}^{m^2} q_{i,j}(1 - q_{i,j}) c_j.$$

We further assume that the fluctuations are Gaussian, i.e., each entry of $\mathbf{w}$ has Gaussian distribution $\mathcal{N}(0, \sigma_{w,i}^2)$. Finally, $\mathbf{v}$ in (1) is the $m^2 \times 1$ vector whose components can practically be assumed to have iid Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and represent the noise due to imperfect instrumentation and other biochemistry independent noise sources.

III. Optimal Estimation of Gene Expression Levels

The maximum-likelihood (ML) estimate of the input concentrations maximizes the probability $p(\mathbf{s}|\mathbf{c})$, and is given by

$$\hat{\mathbf{c}} = Q^{-1}\mathbf{s}. \qquad (2)$$

As an example, we simulate an 8×8 microarray, and apply $n = 6$ types of cDNA targets. Furthermore, $\mathbf{c} = [10000, 20000, 20000, 20000, 10000, 20000]$, and thus the total number of target molecules is $N = 100000$. The probability of hybridization of a cDNA target to the matching probe is assumed to be $p_H = 0.8$, while the probability of cross-hybridization to any one of $k = 3$ other probes is assumed to be $p_C = 0.1$. The probabilities of release from both the hybridized and the cross-hybridized states for all targets are $p_r = 0.02$. The simulations are run sufficiently long so that the steady-states of the Markov chains have been reached.

Table 1 shows the numbers of the target molecules bound to probes on the array. Clearly, due to the non-specific binding as well as the stochastic nature of both the hybridization and cross-hybridization, information about the original concentrations of the targets is lost. The ML solution (2), however, recovers absolute values of the original concentrations, as illustrated in Table 2.

0	0	4684	0	0	1649	0	0
0	0	872	0	818	0	0	864
0	0	797	6504	378	0	1203	0
0	6649	0	0	375	838	810	0
0	0	0	0	7858	0	0	0
0	0	0	0	0	0	0	0
0	0	3103	0	0	374	6768	0
0	0	0	0	869	385	0	0

TABLE I

Number of captured cDNA targets accross the array

0	0	11503	7	16	0	18	16
23	15	42	6	86	18	13	39
15	7	21	19433	40	10	73	12
0	19860	9	0	20	52	55	13
8	1	15	18	20988	0	18	15
11	22	13	19	8	2	0	9
22	10	9322	6	7	22	20216	4
7	18	15	17	185	43	27	9

TABLE II

Estimated input concentrations of the targets.

References

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

[2] W. Zhang and I. Shmulevich (editors), *Computational and Statistical Approaches to Genomics*, Kluwer, 2002.

[3] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," in *PNAS*, October 29, 2002, pp. 14031-14036.

[4] A. Hassibi, T. Lee, R. Navid, R. Dutton, and S. Zahedi, "Effects of scaling on the SNR and speed of biosensors," *Proc. Internat. Conf. of EMBS*, September 2004.