ON LIMITS OF PERFORMANCE OF DNA MICROARRAYS

H. Vikalo, B. Hassibi, and A. Hassibi

California Institute of Technology, Pasadena, CA
e-mail: hvikalo,hassibi,arjang@caltech.edu

ABSTRACT

DNA microarray technology relies on the hybridization process which is stochastic in nature. Probabilistic cross-hybridization of non-specific targets, as well as the shot-noise originating from specific targets binding, are among the many obstacles for achieving high accuracy in DNA microarray analysis. In this paper, we use statistical model of hybridization and cross-hybridization processes to derive a lower bound (viz., the Cramer-Rao bound) on the minimum mean-square error of the target concentrations estimation. A preliminary study of the Cramer-Rao bound for estimating the target concentrations suggests that, in some regimes, cross-hybridization may, in fact, be beneficial—a result with potential ramifications for probe design, which is currently focused on minimizing cross-hybridization.

1. INTRODUCTION

DNA microarrays [1, 2] are affinity-based biosensors where the binding is based on hybridization, a process in which complementary DNA strands specifically bind to each other creating structures in a lower energy state. Typically, the surface of a DNA microarray contains an array 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 complementary DNA (cDNA) which are applied to the microarray afterwards. Under appropriate experimental conditions, labeled cDNA molecules that are a perfect match to the microarray probes will hybridize, i.e., bind to the complementary capturing oligos. Nevertheless, there will always be a number of non-specific bindings since cDNA may non-specifically cross-hybridize to probes that are not a perfect match but are rather only partial complements (having mismatches). It is important to understand that this particular phenomenon, i.e., non-specific 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 present in these system [3]. Finally, the fluorescent labels in each spot are measured to obtain an image, having correlation to the hybridization process, and thus the gene expression levels.

Today, the sensitivity, dynamic range and resolution of the DNA microarray data is limited by cross-hybridization [4] (which may be interpreted as interference), in addition to several other sources of noise and systematic error in the detection procedure [5]. 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 cross-hybridization becomes the dominating limiting factor of the signal strength ([4], [5]). Additionally, 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 [6]. 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.

2. 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

This work was supported in part by the National Science Foundation under grant no. CCR-0133818 and CCR-0326554, by the David and Lucille Packard Foundation, and by Caltech's Lee Center for Advanced Networking.

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 $\sum_{i=1}^{M} M_i = m^2$. The array is scanned after the system has reached bio-chemical 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 non-specific cross-hybridization with probes whose nucleotide sequences are only partly matches with the target. We assume that both hybridization and cross-hybridization are random events. Let q_{li} and n_{li} denote the probability of binding and the total number of bound target molecules of type i to probe l, respectively. 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

$$p(n_{li} = x) = \left(\begin{array}{c} c_i \\ x \end{array} \right) q_{li}^{x} (1 - q_{li})^{c_i - x} \; . \qquad (1)$$

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_{li} into a $m^2 \times 1$ column vector $\mathbf{n} = \left[\begin{array}{cccc} n_1 & n_2 & \ldots & n_{m^2} \end{array} \right]^T$, the measurement obtained from a DNA micro-array is $\mathbf{s} = \mathbf{n} + \mathbf{v}$, where $\mathbf{v}$ is the noise due to imperfect instrumentation 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 $\mathbf{c} = \frac{1}{m^2} \left[\begin{array}{ccc} c_1 & \ldots & c_N \end{array} \right]^T$ we may write

$$\mathbf{s} = Q\mathbf{c} + \mathbf{w} + \mathbf{v}, \qquad (2)$$

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

$$\Sigma_{\mathbf{w}} = \text{diag}(\sum_{i=1}^{N} q_{1i}(1 - q_{1i}) c_i, \ldots, \sum_{i=1}^{N} q_{m^2 i}(1 - q_{m^2 i}) c_i). \qquad (3)$$

Equation (2) 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. Matrix Q can be obtained either from calibration experiments or via analytical expressions such as ΔG, melting temperature, etc. (see, e.g., [7]). Furthermore, note 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.

Remark: Note that we 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. Thus, the parameters of the system model are constant and do not depend on the number of target molecules that are bound to different probes.

3. OPTIMAL ESTIMATION OF TARGET CONCENTRATIONS

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

$$\max_{\mathbf{c} \geq 0} p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}|\mathbf{c}), \qquad (4)$$

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 $\Sigma_s = \sigma^2 I + \Sigma_{\mathbf{w}}$. The optimization (4) is equivalent to the minimization

$$\min_{\mathbf{c} \geq 0} \left[(\mathbf{s} - Q\mathbf{c})^* \Sigma_s^{-1} (\mathbf{s} - Q\mathbf{c}) + \log \det \Sigma_s \right] . \qquad (5)$$

Note that the above problem is highly nonlinear and non-convex (because the c_i are present in both $\mathbf{c}$ and Σ_s). 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 obtained by solving

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

We tested our hypotheses regarding the statistical model and verified performance of the estimation algorithms on *real* microarray data obtained through a set of experiments. The oligonucleotide probes in these experiments are from a commercial set chosen from genes of the bacterium *Escherichia coli*; the oligonucleotide targets are custom designed. We omit the specifications of the experiments due to the lack of space and refer interested reader to [8] for details.

Figure 1 shows measured and estimated signal in an experiment where two targets were applied to a microarray. Due to cross-hybridization, direct readout implies presence of four targets. The estimation algorithm correctly detects presence of only two targets, and precisely recovers their concentrations.

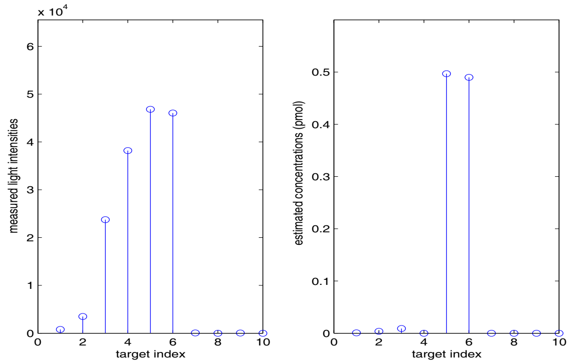

Fig. 1. **Measured and estimated signal in an experiments where two targets were applied to a microarray. Due to cross-hybridization, direct readout implies presence of four targets. The estimation algorithm correctly detects presence of only two targets, and precisely recovers their concentrations.**

4. LIMITS OF PERFORMANCE

The minimum mean-square error of *any* estimation procedure is lower bounded by the Cramer-Rao bound [9]. Assuming an unbiased estimator, the Cramer-Rao lower bound (CRLB) on the minimum mean-square error of estimating a parameter c_i is given by

$$E\left(\hat{\mathbf{c}}_i - \mathbf{c}_i\right)^2 \geq [F^{-1}]_{ii}, \tag{6}$$

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

$$F_{ij} = -E_{\mathbf{s}} \frac{\partial^2}{\partial c_i \partial c_j} \log p_{\mathbf{s}|\mathbf{c}}(\mathbf{s}). \tag{7}$$

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

$$L(\mathbf{c}) = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma_s - \frac{1}{2}(\mathbf{s} - Q\mathbf{c})^T \Sigma_s^{-1}(\mathbf{s} - Q\mathbf{c}).$$

After computing the Hessian (details omitted for brevity), we obtain

$$F = Q^T \Sigma_s^{-1} Q + \frac{1}{2}(Q - Q \odot Q)^T \Sigma_s^{-2}(Q - Q \odot Q). \tag{8}$$

Our end result therefore is $E\left(\hat{\mathbf{c}}_i - \mathbf{c}_i\right)^2 \geq$

$$\left[\left(Q^T \Sigma_s^{-1} Q + \frac{1}{2}(Q - Q \odot Q)^T \Sigma_s^{-2}(Q - Q \odot Q)\right)^{-1}\right]_{ii}. \tag{9}$$

4.1. Comparison with direct readout

Note that, being unbiased, the maximum-likelihood estimate (5) achieves the Cramer-Rao bound in (9). In most current applications of micro-arrays, 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

$$E_{\mathbf{s}}(\mathbf{s} - \mathbf{c})(\mathbf{s} - \mathbf{c})^T = (Q - I)\mathbf{c}\mathbf{c}^T(Q - I)^T + \Sigma_s. \tag{10}$$

Comparing (10) with (9) for a given system model and concentrations, provides a measure of the improvement of the techniques proposed in this paper over the methods that employ direct readout.

4.2. The effect of cross-hybridization

In current micro-array 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 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 so that it 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 (9) therefore becomes

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

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

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

and thus $F_{11}^{c} > F_{11}^{nc}$. 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

$$\left[F^{-1}\right]_{11} = \frac{1}{a-b} \cdot \frac{a + (N-2)b}{a + (N-1)b}. \qquad (12)$$

This is the CRLB that should be compared with the one without cross-hybridization in (11). Figure 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 seem 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.

5. SUMMARY AND CONCLUSIONS

We computed the Cramer-Rao bound for error of target concentrations estimation in DNA microarrays. The bound is derived assuming a statistical model for DNA microarrays based on a probabilistic description of the hybridization and cross-hybridization processes. The statistical model captures the shot noise nature of the noise in DNA microarrays that has been earlier observed experimentally [4].

Typically, probe design is based on minimizing the amount of cross-hybridization (see, e.g., [6] and the references therein). However, some preliminary studies of the Cramer-Rao bounds suggest that cross-hybridization may, in fact, be beneficial in certain scenarios. 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.

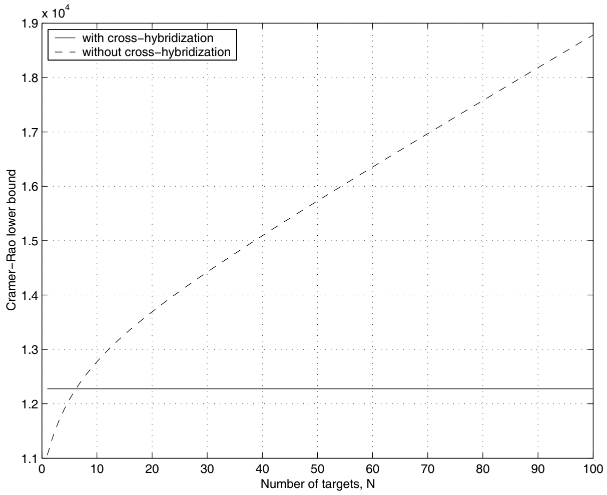

Fig. 2. The CRLB with and without cross-hybridization as a function of the number of target types N. The parameters are $\sigma^2 = 1000$, $c = 500$, $m^2 = 100$, $q = 0.3$, and $\beta = 0.01$.

6. REFERENCES

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

[2] U. R. Mller and D.V. Nicolau (Eds.), *Microarray Technology and Its Applications*, Springer, Berlin, Germany, 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.*, 97, 084701, 2005.

[4] Y. Tu, G. Stolovitzky, and U. Klein, “Quantitative noise analysis for gene expression microarray experiments,” in *Proceedings of National Academy of Science (PNAS)*, October 29, 2002, pp. 14031-14036.

[5] A. Hassibi and H. Vikalo, “A probabilistic model for inherent noise and systematic errors of microarrays,” *Digest of IEEE International Workshop on Genomic Signal Processing and Statistics*, 2005.

[6] L. Kaderali and A. Schliep, “An algorithm to select target specific probes for DNA chips,” *Bioinformatics*, 18(10), pp. 1340-1349, 2002.

[7] J. SantaLucia, Jr. and D. Hicks, ”The thermodynamics of DNA structural motifs”, *Annu. Rev. Biophys. Biomol. Struct.* 33, 415-40, 2004.

[8] H. Vikalo, B. Hassibi, and A. Hassibi, ”A statistical model for microarrays, optimal estimation algorithms, and limits of performance,” to appear in *IEEE Trans. on Signal Processing*.

[9] H. Cramer, *Mathematical Models of Statistics*, Princeton University Press, Princeton, NJ 1946.