SPATIAL DEPENDENCE OF HYBRIDIZATION IN THE cDNA MICROARRAY

STEPHEN J. MERRILL, SANGEETA NELSON AND CRAIG A. STRUBLE

ABSTRACT. The cDNA microarray is an experiment in which a large number of molecules (here DNA associated with known genes) are affixed or “printed” to a slide or chip in an array of spots. The hybridization of the “probes” in the printed spots with mRNA-derived fluorescently labeled “targets” allows one to study gene expression in tissue. The experiment has many variations and difficulties, primarily concerning problems of reproducibility and interpretation. A barrier to understanding experiments of this type is the lack of a simplified model of the process to explore aspects of the analysis and variability inherent in the experiment.

In this paper, a mathematical model of one aspect of the experiment is developed, the hybridization of probes with targets. The model consists of several thousand Markov chains, each differing with respect to the transition matrix and the initial vector. The stochastic nature of the process (generally a small number of targets specific for each probe species) is mimicked by the stochastic nature of the chains. This model is applied to describe the dependence of the expected level of hybridization and the variance of that level on the location of a spot on the slide. In consideration of these results, spot position must be considered for any normalization method in addition to adjusting for fluorescence intensity and printing biases. The results also suggest that genes that expected to have a lower level of expression should be spotted near the edges for better detection.

Introduction As the sequencing of the human genome was being completed, Lander et al., 2001; Venter et al., 2001 research interest in genetics and allied fields turned to the study of the genes and structures within the genome that would lead to the big picture of gene expression and its regulation. Microarray and microchip experiments (Fodor, et al., 1993, Chee et al., 1996; Shalon et al., 1996) were developed to identify genes that were being expressed in tissue under specific conditions.
Although details of the experiments vary, and proprietary processes are involved in the preparation and analysis of the results, the basic outline of an idealized cDNA microarray experiment is as follows (Baldi and Hatfield, 2002):

1) **Development of a gene library.** A library consists of thousands of cDNA (complementary DNA) clones, their numbers amplified by PCR (the polymerase chain reaction). These serve as the “probes.”

2) **Printing of the probes on a slide.** Spots of cDNA are printed at high density on a specially prepared glass slide. Each spot contains many copies of a single probe species. Printing is accomplished through a robotic DNA arrayer if a custom library is being used. Commercially available gene libraries can also be purchased pre-printed on a slide or chip.

3) **Collection of mRNA (messenger RNA) under two different conditions.** Using RT (reverse transcriptase) mRNA samples are (reverse) transcribed to complementary DNA (the gene that made it) in the presence of fluorescently labeled nucleotides. The use of different fluorescent dyes (Cy3 and Cy5) for the two samples results in two populations of labeled “target” DNA, the dye incorporated indicating the source of the sample. If a gene is present in the gene library for a particular mRNA, the cDNA probe for that gene will be complementary to its labeled target DNA.

4) **Incubation.** Using the fact that complementary (A with T and G with C) sequences of base pairs will hybridize, the targets from the two samples are incubated with the probes printed on the slide.

5) **Scanning.** After washing, only hybridized probe-target pairs remain on the slide. The slide is scanned after activating each fluorescent dye in turn to recover the level of fluorescence at each spot (a function of the amount of hybridization at that spot) for each color.

6) **Data analysis** to compare level of expression and identify expression patterns of gene networks. There are many monographs and papers on the proper statistical approach to analyze microarray data, for instance Speed et al., 2003 and Parmigiani et al., 2003. These involve experimental design, image analysis, normalization, analysis of variance, and various styles of clustering and classification.

A typical microarray experiment uses several thousand to many tens of thousands of spots, each containing a specific probe. Theoretically, then, the experiment determines the gene expression levels for each represented gene under each of two different conditions. Differential expression in the two samples can be expressed as the ratio of levels of fluores-
cence of the two dyes (properly normalized). In addition to molecular and cellular biological and genomic research, applications of the microarray include pharmacology research and drug discovery (Xiang, et al., 2003), revealing mechanisms in disease (Shaughnessy, 2003), providing a platform for differential cancer diagnostics (Ringner and Peterson, 2003), and forensic and genetic identification (Heller, 2002).

The promise of the microarray has been damped, however, by the variability added at each step of the process and the difficulty in dealing with the quantity and nature of data produced. This has resulted in some skepticism as to the eventual importance of this approach. As noted in Baldi and Hatfield, 2002 (p. xi–xii):

However, while the DNA array will be an important workhorse for the attainment of these goals, it should be emphasized that DNA array technology is still at an early stage of development. It is cluttered with heterogeneous technologies and data formats as well as the basic issues of noise, fidelity, calibration, and statistical significance that are still being sorted out.

In this paper, one source of variability, spatial variability, is addressed through the construction of a dynamic simulator. This simulator tracks the expected capture of targets of each spot through a Markov chain. The ensemble of Markov chains produces a picture of the spatial variation present in the process, and suggests that global normalization of the fluorescence levels introduces errors in the resulting analysis. A test bed of simulated microarray data has also recently appeared, Singhal, et al., 2003, for the purpose of testing analysis techniques. They also deal with distributions of intensities, but without any attention to spatial characteristics.

1 Simplified description of a cDNA microarray experiment

There are many versions of microarray experiments; a simplified description of a representative experiment is given below.

A spot is placed on a slide, the spot containing many copies of cDNA of a gene of interest. The procedure for the printing, handling of the cDNA, and coating on the slides insures that the attached cDNA are single stranded and available for hybridization with a complementary DNA molecule (see Figure 1). Dots typically have a diameter of .075–.1 mm and are placed about .15 mm apart from center-to-center.

A slide, either preprinted or printed by robot, contains such spots placed in an organized rectangular array. Typically, a 25 mm×75 mm
glass slide will have a 22 mm×60 mm maximal printable (coated) area. Other areas on the slide are for experiment identification and handling. For instance, if a 19 mm×19 mm printed area is used, assuming a .15 mm spacing between spots, 127 dots could be placed in each row of 127 rows resulting in over 16,000 printed dots (see Figure 2).

Fluorescent target DNA produced from mRNA as described in step 3) above is incubated with the array of cDNA “probes” printed on the slide. Generally, a small amount of this material is placed on glass coverslip that is large enough to just cover the printed area. The slide is inverted and placed on the coverslip. The resulting covered slide is then again flipped and placed in an incubator for about 12 hours (the time could be shorter depending on the conditions of incubation and the specifics of the DNA being studied). The coverslip can either be fixed in place with rubber cement, or the covered slide can be placed in a moist chamber without sealing. Both methods prevent drying of the slide that would end the hybridization process prematurely.

During the hybridization, target molecules diffuse under the coverslip until they hybridize with a complementary cDNA probe. Not all targets will find a probe, either because one was not available or the hybridization time was not long enough to enable all targets to find their complementary molecules.
The basic quantity of interest is the ratio of the level of hybridization at each spot between the two samples. That quantity indicates the relative level of expression of the associated gene—seen as the abundances in the associated mRNA in the original tissue sample. This quantity is measured through detecting the level of fluorescence at that spot from each dye. Two passes of a specialized scanner, each pass exciting a different dye, results in an array of many thousands of colored spots, false colored to reflect the level of intensity found in each pass. Experimental variation makes necessary the introduction of normalization methods to correct for the number of cells in the sample, the total RNA isolation efficiency, mRNA isolation and labeling efficiency, hybridization efficiency, fluorescent signal variation, and printing biases (Baldi and Hatfield, 2002; Speed, 2003). In this paper, a model will be constructed to explore the role of the location of the spot on the degree of hybridization and its variation.

2 The distribution of initial distances
Consider a single spot. When incubation begins, an unknown number of targets specific for the probes in that spot are distributed uniformly under the coverslip. The distribution of distances of targets to the spot depends on the location of that spot. To describe this distribution, consider a rectangular grid, 254 \times 254 representing positions of the printable area. To adjust the spacing appropriately, assume that alternate grid coordinates are printed with a spot. This assumes the following geometry as shown in Figure 3. A grid of the type shown in Figure 3 can accommodate 254 rows and 254 columns in a 19 mm square. Using the “taxicab” metric (row distance + column distance), the distribution of target distances to a given spot location can be described. For a given \((i^*, j^*)\) location, 1 \leq i^*, j^* \leq 254, for each entry in a 254 \times 254-matrix, record the taxicab distance. A relative frequency histogram of these distances gives the distribution of distances to \((i^*, j^*)\). As seen in Figures 4a and 4b, both the mean distance and the shape of the distributions depend on position. When targets are initially present under the coverslip, the distances they must diffuse to reach a complementary probe are chosen from these distributions.

3 A Markov chain model of hybridization at a single spot
If all spots have unique specificity and there is no cross-reactivity with targets, then each spot experiences hybridization independently of all the others. It is as if there is only one spot and only one species of
FIGURE 3: Illustrates the location of spots with space between. The spots have a diameter of .075 mm and the spots are .15 mm apart on center.

FIGURE 4a: The exact distribution of distances to (125, 125) in a 254 × 254 grid.
target DNA under the coverslip. In addition, if there are many more probes in a spot than targets for that specificity, one can also assume that there is little competition between targets for probes. We propose that the kinetics of the hybridization at a spot can be modeled by a distance time Markov chain (see Sobel and Heyman, 2003 for a good general discussion of Markov Chains).

For each spot location \((i, j)\) consider a 1-d discrete time Markov chain with states 0, 1, \ldots, \(N_{ij}\). The states correspond to distances using the taxicab metric from a specific spot. The time interval \(\Delta t\) is defined to be the time necessary for a DNA target to have the probability of 1/2 to move out of grid square as defined in Section 2. A move will be defined as a move horizontally or vertically in the grid. It may be the case that \(\Delta t\) varies with time and conditions of the incubation—making the chain time-homogeneous. The chain is then specified by describing the \((n + 1) \times (n + 1)\) constant transition matrix for \((i, j)\). The matrix has a similar structure for any spot location, but specifics of the location determine the manner in which the distances change over time.

In a given array, define relative locations in reference to a specified spot: corners, edges and interior locations. In addition, an edge loca-
FIGURE 5: With a specified spot at the 4th row and 3rd column of a $5 \times 4$ grid, the relative locations: corner (diagonal stripe), edge not same row/column (vertical stripe), edge same row/column (dark gray), interior same row/column (light gray), interior not same row/column (white).

The importance of the relative locations is that transitions in distance from each type of point depend on the relative location. For instance, if a target is in a corner and the spot is in the interior of the grid (as in Figure 5), then every move from the corner moves the target closer to the spot location. On the other hand, if a target is at an interior position that is on the same row or column as the spot, then 3 of 4 adjacent grid squares are farther away in the taxicab metric and only one move would bring the target closer to the spot.

If the spots are assumed to always be in the interior of the grid, then the number of each relative location is the same but considering the distance distribution, the number of each type at each distance differs with the spot location. The transition matrix corresponding to a spot at a given location is created by considering the distance and type distribution, and the probability of moving closer or farther away at each step. For $\Delta t$ defined as above, the following rules were used to create
the transition matrices.

1) At distance \( j \) from the spot location, the probability is .5 that a target stays in that same grid square (and staying at the same distance) and probability is .5 that a target moves to an adjacent square.

2) If moving to an adjacent square, the distribution of types at the current distance will determine the probability that that move would bring a target closer, or farther away, from the spot.

3) Only moves to adjacent grid squares by horizontal or vertical moves are allowed in \( \Delta t \) of time.

4) It is assumed that the spot location is absorbing, that is, once a target enters that spot, it stays forever in that spot.

Given the above rules, transition matrices \( P_{ij} \) were constructed for each interior \((i, j)\) location of a spot on a \( 254 \times 254 \) grid. These matrices describe a large family \( \{ Y_{ij}^n \} \) of Markov chains.

4 Simulations of the expected hybridized fraction Both the number of states and the transition matrix entries vary with the position of a spot. To describe the spatial variation in the expected hybridized fraction at each spot, for each location, the initial distribution \( a_{ij}^0 \) and the transition Matrix \( P_{ij} \), were calculated. Then the chain was iterated

\[
a_{ij}^{n+1} = a_{ij}^n P_{ij}, \quad n = 0, 1, \ldots, n_{\text{final}}.
\]

As the hybridization process in a microarray is stopped after a finite time, it is the state after a finite number of iterations that is of interest here. The final number of iterations depends on the length of the incubation and \( \Delta t \). Our estimate is that \( \Delta t \approx .11 \) sec. for this size grid, meaning that 400,000 iterations correspond to roughly 12 hours. This estimate depends on the size of the targets (not uniform from experiment to experiment), the height of the coverslip above the slide (not uniform across experiments), and the temperature and conditions of the hybridization (not uniform). From the simulations shown in Figure 6, the basic shape of the spatial dependence does not depend on this estimate. The vertical axes, the fraction hybridized, is the probability of absorption by the “zero” state (the spot which is at zero distance from itself) does depend on the number of iterations. Due to the assumption that once within zero distance the probability of hybridization is 1, the probability of absorption in one of the chains and the probability of hybridization are the same.
Figure 6: Results from 50,000 and 400,000 iterations of 252 \times 252 different Markov chains simulating the fraction of possible hybridization for spot positions in that grid after a simulated 1.5 hour and 12 hour incubation, respectively. Note the “bright edges” and the lower hybridization in the center. Simulations and graphs from Matlab 6.5 (Natick, MA USA).
The mean fraction absorbed by the final time and the variance at each spot location can also be computed by the following argument. For a given spot location \((i, j)\), define \(y_k(n) = P(Y_{ij}^n = 0 | Y_{ij}^0 = k)\). This is the probability of being absorbed in \(n\) iterations (time units) if the chain started at distance \(k\) at time 0. Then for \(k > 0\) we can condition on the first transition,

\[
y_k(n + 1 | Y_1 = \ell) = \begin{cases} 
1 & \text{if } \ell = 0, \\
y_k(n) & \text{if } \ell > 0.
\end{cases}
\]

Using the law of total probability, for each \(k > 0\),

\[
y_k(n + 1) = p_{k0} \cdot 1 + \sum_{i=1}^{N_{ij}} p_{k\ell} y_{\ell}(n).
\]

In this expression, \(p_{k\ell}\), the elements of the matrix \(P_{ij}\), is the transition probability of making a transition from distance \(k\) to \(\ell\) in one time unit when the spot is at position \((i, j)\). The initial condition for this system of difference equations is \(y_k(0) = 0\) for all \(0 < k \leq N_{ij}\) and \(y_0(0) = 1\). Define the random variable \(X_{ij}\) by \(X_{ij} = 1\) if a target complementary to spot at \((i, j)\) is hybridized by the final hybridization time \(n_{\text{final}}\), and \(X_{ij} = 0\) otherwise. Then

\[
P(X_{ij} = 1) = \sum_{k=0}^{N_{ij}} P(X_{ij} = 1 | Y_0 = k) P(Y_0 = k)
\]

\[= a_0^0 + \sum_{k=1}^{N_{ij}} y_k(n_f) a_k^0.
\]

The expected value is

\[
E[X_{ij}] = P(X_{ij} = 1) \cdot 1 + P(X_{ij} = 0) \cdot 0 = P(X_{ij} = 1),
\]

and the variance is

\[
Var(X_{ij}) = P(X_{ij} = 1) - (P(X_{ij} = 1))^2.
\]

The quantities plotted in Figure 6 are \(E[X_{ij}]\). This means that the variance as a function of position can also be easily displayed. The variance as a function of position is shown for 400,000 iterates in Figure 7.
FIGURE 7: The variance in the hybridized fraction at 400,000 iterations as a function of position.

Also of interest is the time it takes to hybridize a target complementary to a spot at $(i,j)$. The probability of absorption in the $n$th transition starting at time 0 at $k$ are the differences

$$z_k(n) = y_k(n) - y_k(n-1).$$

Using these differences and the fact that for these finite chains a sample path will be absorbed with probability 1, the expected time to absorption for a chain starting at state $k$ is

$$E_k = \sum_{n=1}^{\infty} nz_k(n).$$

The variance of the time to absorption for a chain starting at state $k$ is

$$V_k = \left( \sum_{n=1}^{\infty} n^2 z_k(n) \right) - E_k.$$
As the hybridization of each target is assumed to proceed independently of that of other targets, the variance associated with a particular location can be computed as the weighted sum of the variances, the weighting being the initial distribution of distances $a_{ij}^0$.

What is different in this application is that as the hybridization process does not run forever, not all sample paths will be absorbed by the final time. Define

$$E^m_k = \sum_{n=1}^{m} n z_k(n) \quad \text{and} \quad V^m_k = \left( \sum_{n=1}^{m} n^2 z_k(n) \right) - (E^m_k)^2.$$  

These quantities in the limit are the expected value and the variance, respectively. The weighted sum

$$V^m_{ij} = \sum_{k=1}^{N_{ij}} V^m_k (a_{ij}^0)_k$$

for large $m$ approximates the variance in the time to absorption at position $(i, j)$.

## 5 Evidence for spatial dependence in microarray data

To find evidence of spatial dependence in microarray data, we randomly selected experiments from several public databases containing microarray data, such as the Stanford Microarray Database (Collub, et al., 2003) and the ArrayExpress (Brazma, et al., 2003), and generated scatter plots of spot intensity vs. spot distance from the slide center. A curve is then fit through the scatter plots, which should be approximately flat if spot intensity is not dependent on the distance from the slide center.

Obtained data include coordinates for each spot within the scanned image. The slide center for each hybridization was estimated from these coordinates by calculating the midpoint between the extreme upper left and lower right coordinates found in a data file. The spot distance from the center was calculated as the Euclidean distance between the spot center and the estimated slide center. A curve is then fit through the scatter plot of spot intensity vs. spot distance from the center using robust locally weighted regression (Cleveland, 1979) implemented by the `lowess` function in R version 1.7.1 (Ihaka and Gentleman, 1996).

The results from applying this approach to one data set are shown in (Whitfield, et al., 2002) in Figure 8. (This data set is experiment ID 7271 in the Stanford Microarray Database.) In the two pictures,
FIGURE 8: A plot of LOWESS curve fit through distance vs. intensity data derived from the two foreground channels of experiment ID 7271 in the Stanford Microarray Database (Whitfield, et al., 2002).
intensity is increasing as the distance from the slide center increases, as is predicted by our model. These results are typical of the results we see for foreground intensities across the experiments we investigated.

Bright edges in microarray and in the background have been frequently noticed (see Figure 7 of Smyth, et al., 2002 and Schaldt, et al., 2000) and are usually labeled as artifact or a result of protocol issues.

6 Discussion Systematic bias in data obtained from microarray experiments has long been identified as an important problem. In this paper we have developed a model of hybridizations for microarray experiments that shows a spatial dependence on the fraction of target DNA hybridized to their complementary probes. As spot intensity is related to the amount of hybridized target DNA, a systematic increase in intensity in spots that are farther from the slide center is predicted by our model. Evidence from microarray data supports our prediction.

Our results suggest that normalization of microarray data should take into account the location of the spot on the slide in addition to other factors. Currently, two general approaches to normalization are employed: global normalization, which multiplies each ratio of intensities by a single scalar normalization factor; and local normalization, which applies a normalization factor to spots grouped in some fashion, such as by average intensity or subarray (i.e., same pin used for printing). Neither of the two general normalization procedures consider a spatial dependence (Workman, et al., 2002, Edwards, 2003, Speed, 2003).

The design of microarray slides is also affected by our results. Genes known to have expression at low abundance should be spotted closer to the edges of the microarray slide rather than in the center, so as to capture a higher proportion of target DNA. Slide scanners are known to have difficulty in scanning low intensity spots, and using this method would improve the chances of obtaining data for those low abundance genes. The fact that the variance also varies with position raises further questions concerning how the ratios of intensities of the two fluorescent dyes vary. This will be examined at a later time.
REFERENCES


Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI 53201-1881 USA.