The Analysis of Oligonucleotide Microarray Data at the Raw Image Level and the Probe Level

Jeff Palmer

Summer, 2005

Thesis Proposal

1 Introduction

One of the greatest advances in the biological sciences in recent years is arguably the completion of the Human Genome Project. More specifically, scientists have only recently completely enumerated the entire biological signature of the genome for several species, including undoubtedly the most important to mankind, *homo sapiens*. For each species, this entails the determination of the 3^9 base-pair chain (divided into chromosomes) of nucleotides (*deoxyribonucleic acid*, or DNA) that is essentially common to all organisms within a species, along with the identification of the tens of thousands of genes which serve to indirectly control specific traits of the organism.

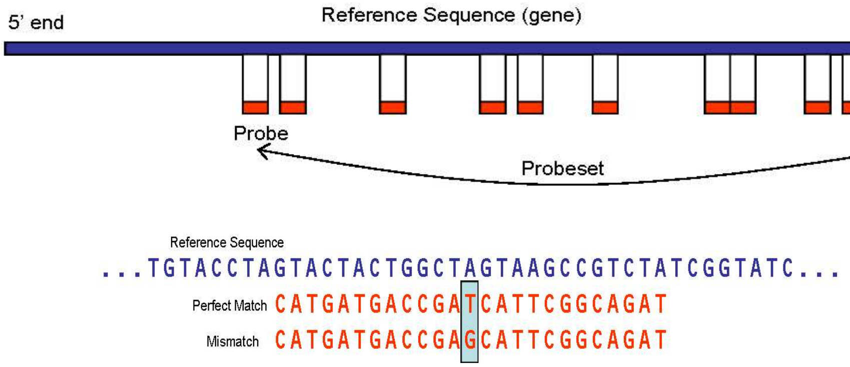
A gene is a subsequence of DNA that codes for a protein, which in turn controls a specific aspect of a cell. Protein encoding occurs in two stages: *transcription* and *translation* (combined, these two processes are what is referred to as the *central dogma of molecular biology*). Transcription is the process of copying the complement of the *antisense* strand of DNA to single-stranded mRNA. An RNA polymerase enzyme travels along the antisense strand of the DNA from the 5' end to the 3' end, encoding the mRNA strand along the way. The 5' end refers to the 5'-hydroxyl group that starts the first nucleotide in the strand, and the 3' end refers ot the 3'-hydroxyl group that ends the last nucleotide. The complement of the nucleotide A (adenine) is T (thymine), and the complement of C (cytosine) is G (guanine). Translation is then the process of translating mRNA into proteins.

While the relationship between protein activation and various biological characteristics is of primary concern, the determination of protein existence in a tissue is very difficult. Therefore, the determination of gene presence in a tissue sample often serves as a proxy in scientific experiments. In the past, the determination of gene presence in a tissue sample was done one gene at a time using methods such as RT-PCR and Northern blotting, however, microarray gene expression methods have taken the forefront in recent years due to their ability to simultaneously measure the expression levels of several thousands of genes in a single tissue sample. The oligonucleotide microarray is one such method, and is the focus of this dissertation. The most popular oligonucleotide microarray design is the GeneChip family of arrays, which are manufactured by the company Affymetrix.

1.1 Oligonucleotide Microarray Technology

The concept of high-density oligonucleotide microarrays was first introduced by Lockhart, *et. al.* (1996). A review of the relevant technology may be found in Lipshutz, *et. al.* (1999). Generally speaking, oligonucleotide microarrays have several hundred thousand *cells*, each approximately 10-20 square microns in size, laid out in a lattice (grid) fashion. Within each cell, millions of copies of a particular *probe* sequence of nucleotides are constructed using photolithographic techniques borrowed from the semiconductor industry. The lengths of these sequences typically range from 10 to 100 nucleotides (thus, *oligo-*), and are selected to be complementary to a particular *target* mRNA sequence of interest.

The current generation of Affymetrix GeneChip microarrays have $\approx 500,000$ probe cells, each 25 basepairs (bp) in length, arranged in *probe pairs*: for each target mRNA sequence, there is a *perfect match* (PM) probe that is perfectly complementary to the target sequence, and a *mismatch* (MM) probe that is equivalent to the PM sequence except for the 13^{th} (middle) nucleotide. The MM probes are designed to control for non-specific binding and cross-hybridization. Non-specific binding refers to the binding of a probe to a gene other than its target, and cross-hybridization occurs when a probe binds to a target sequence that is not completely complementary. Refer to the bottom portion of Display A for a schematic diagram of an example probe pair. The allocation of probes to points on the lattice is for all intents and purposes random, with the only exception being minimal constraints imposed by the manufacturing process.


5' end
Reference Sequence (gene)
3' end
Probe
Probeset
Reference Sequence
...TGTACCTAGTACTACTGGCTAGTAAGCCGTCTATCGGTATC...
Perfect Match CATGATGACCGATCATTCGGCAGAT
Mismatch CATGATGACCGAGCATTCGGCAGAT

Display A: *Schematic diagram of probe structure. Top panel: A reference sequence (blue) and possible locations of probes (red). Bottome panel: An example reference sequence and the corresponding PM and MM sequences for a particular probe pair.*

Due to the fact that the probe sequences are only 25 nucleotides in length, a set of 11-20 probe pairs (termed *probeset*) are typically used to interrogate a particular gene of interest. Gene lengths vary from hundreds to several thousands of nucleotides. The combined hybridization levels of all of the probes in a probeset are used to represent the expression level of the gene in question. The probes were selected for their binding properties, and their locations are typically 3' end biased (the start of a probeset is always within 600 nucleotides of the 3' end of the gene). The specific locations of the probes along the gene vary dramatically from gene to gene, where sometimes probes can overlap, while other times they may be separated by several hundred nucleotides. Refer to the top portion of Display A for a schematic diagram.

The data acquisition process begins with the preparation of a tissue sample by the scientist. A protocol-defined amount of mRNA is extracted from the sample, is tagged with fluorescent markers, and is then subjected to a hybridization period during which the mRNA in the sample binds to the probe sequences on the chip. Following the hybridization period, the fluorescent markers are excited and an image of the chip is produced using a confocal laser scanning microscope. The observed intensities in the image within each cell are intended to measure the relative abundance of the corresponding mRNA sequence in the sample.

1.2 Analysis Issues

The analysis of oligonucleotide microarray data involves a complex sequence of processing steps. The idealized goal is to measure the abundance of particular gene transcripts in a tissue sample, however, several factors at each stage of the analysis process contribute to the variability in these observed abundance levels. The first stage at which error may be introduced is at the tissue sample collection and preparation, and hybridization steps discussed in the previous subsection. Errors at this level may be introduced by inadequate levels of mRNA in the prepared sample, an insufficient period of hybridization to the chip, and in general, any deviation from the protocol set forth by Affymetrix. While these sources of error are important, I will focus my attention in this dissertation to those sources of error that are introduced after the hybridized chip has been imaged.

A typical array image is scanned at a resolution such that each probe cell is represented by approximately 7×7 pixels. So, for an array that has $640 \times 640 = 409,600$ cells, the initial raw image will have approximately 20 million pixels. The first step in image processing is to collapse the dimensionality of the image so that each cell is represented by a single value used to represent the associated *probe intensity.* This process may be loosely referred to as *feature extraction.* The two primary issues involved at this step are the determination of the cell locations in the image and the method used to combine all of the pixel values within each cell into a single summary measure. The overwhelming majority of researchers use the suite of algorithms provided by Affymetrix in their proprietary software to handle this stage of the analysis, the details of which are discussed in Section 2.1.

Hartemink (2001) suggests that the variability in the observed probe intensity values should be decomposed into two types of variability: *interesting variation* and *obscuring variation.* Interesting variation is the variability that is due solely to biological factors, for example, variability between the gene expression levels in two arrays due to a true difference in gene presence between the two samples. Obscuring variation is due to such factors as background noise in the image, optical imaging errors, and non-specific binding and cross hybridization of the probes to the target mRNA. The term *background correction* loosely refers to any method that attempts to adjust for components of variation that are of the obscuring type. Some authors treat background correction as a separate processing step, independent of the other steps in their analysis. Still other authors choose to incorporate a background adjustment as part of another processing step, typically during the stage of the analysis when the multiple probe pairs are combined to arrive at a single gene expression measure. Some of the more widely used background correction methods will be discussed in Section 2.2, and those methods that are incorporated into a subsequent data processing step will be revisited in Section 2.4.

When it is of interest to compare the results from two or more arrays, an essential step in the analysis is to perform *normalization*, which involves applying a transformation to the values on one chip so that they coincide to the greatest degree possible with the values on another chip. The main factor that drives the necessity to normalize is that two prepared samples will typically have different overall levels of mRNA, which causes the observed intensities on the chips to differ. This is discussed further in Section 2.3.

Ultimately, one of the primary endpoints of a microarray expression study is the assessment of *differential gene expression* among two or more experimental conditions. Differential expression may be quantified in one of many ways, however, the use of a fold change estimate is typically preferred. The observed values on an array have no units and thus their meaning is arbitrary, though it is hypothesized that the values are related to the true gene expression levels on a relative scale. To facilitate the assessment of differential expression, it is typical to combine the probe values for each array under study to form a single *gene expression summary measure.* Fold change between two arrays can then be directly calculated from these summary measures. Some of the current approaches to combining the probe values are discussed in Section 2.4.

2 From Raw Data to Differential Expression: Current Approaches

This section attempts to review a few of the current approaches to each of the analysis issues introduced in section 1.2. In section 2.1, I discuss the current feature extraction method used by the Affymetrix suite of software and feature extraction algorithms proposed by independent researchers. Several proposed methods that deal with background correction (removing optical noise, non-specific binding, and cross-hybridization) are reviewed in section 2.2. The need for normalization of multiple chips has received much attention in the literature and several methods are discussed in section 2.3. In section 2.4, I discuss the current competing approaches to computing gene-specific expression measures.

2.1 Dealing With the Raw Scanned Image: Feature Extraction

As alluded to in section 1.2, the proprietary software provided by Affymetrix (Affymetrix, 1999, 2001) is the current standard for feature extraction from the raw scanned image. Due to the proprietary nature of the software, documentation of their algorithms is limited. Each chip is equipped with a checkerboard pattern of bright features occuring at each of the four corners to facilitate the determination of the four corner pixels of the array. These features, also occurring along the borders of the array, are control *B2 oligo* probe cells. During tissue sample preparation, a protocol-defined amount of complementary mRNA is added to the hybridization cocktail to ensure strong binding levels to these probes, thus yielding the bright features. Based upon the checkerboard patterns, the software outputs a list of the four corner pixels. Figure 1 presents four 84×84 pixel subimages of an example MGU74A_V2 array (mouse genome), each representing the four corners of the array. Each panel in the figure has as its origin the respective corner *landmark* pixel output by the software (coordinates in parentheses). The checkerboard pattern of features is clearly evident in all four panels.

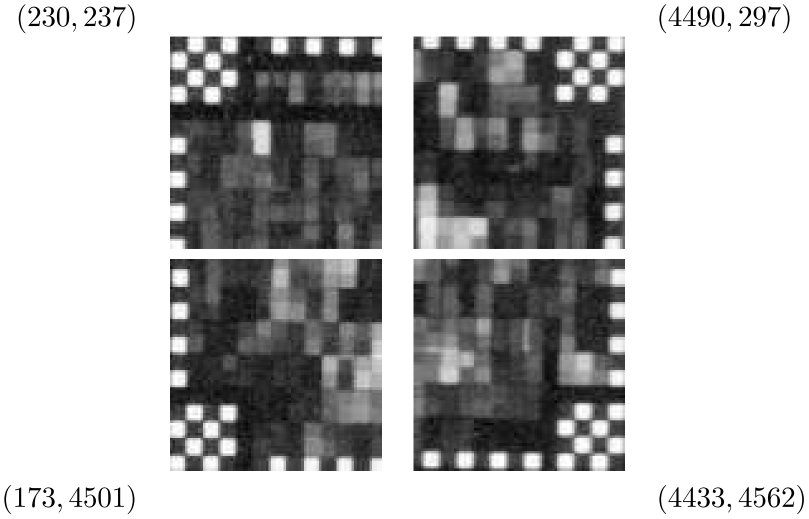


Figure 1: *Four corner* 84×84 *pixel subimages from the raw image. Respective landmark coordinates appear in parentheses.*

Once the four landmark pixels are determined, the software proceeds by mapping the locations of each cell using a weighted average of the landmarks coordinates. For each cell, the border pixels are removed, and then the feature value is reported as the 75^{th} percentile of the remaining pixels. Schadt, *et al.* (2001) noticed that this global interpolation procedure often incorrectly estimates the true feature locations. They proposed a coarse-to-fine procedure where once a candidate cell location is determined, they impose an iterative edge removal algorithm to adjust the estimated feature position. At each iteration they remove the edge whose removal results in the smallest variance among the remaining pixels. This is repeated until a predetermined feature size is reached, typically 4×4 pixels. They then use the mean pixel value as their feature value.

Affymetrix (Affymetrix, 2001) has since refined their algorithm to address this issue. Once a candidate feature center is determined, the algorithm considers all 25 possible feature centers by moving the candidate center ± 2 pixels in both directions. The final feature region is the one that results in the smallest coefficient of variation in the 5×5 pixels centered at each of the 25 possibilities.

2.2 Background Correction

In this section we review a few of the current methods found in the literature for performing background correction. Each method can be categorized as attempting to adjust for one of two types of background: overall background noise, or non-specific binding and cross hybridization.

One method for adjusting for overall background noise is the *location specific correction* method introduced by Affymetrix (2002). This algorithm begins by dividing the array into a $k \times k$ (default is 4×4) grid and computing the mean of the lowest 2% intensity values within each square zone. Each mean value, $Z_{ij}, i, j = 1, .., k$, is referred to as the "background" for that respective zone. Each pixel in the image is then adjusted by these background values, with weights inversely proportional to the distance from the pixel to the centroid of the region in question. More specifically, if (x, y) denotes the coordinates of a pixel, then the euclidian distance from the pixel to the centroid of zone (i, j) is denoted $d_{ij}(x, y)$. Weights are defined as $w_{ij}(x, y) = \frac{1}{d^2_{ij}(x,y)}$, and the background level for the pixel is computed as $b(x, y) = \frac{\sum_i^k \sum_j^k w_{ij}(x,y) Z_{ij}}{\sum_i^k \sum_j^k w_{ij}(x,y)}$.

The mismatch (MM) probe cells were built into the design of the Affymetrix array for the purpose of measuring the amount of non-specific binding to the target sample. Intuitively, if a PM value is large relative to its associated MM value, then it seems reasonable to attribute the observed expression level of that PM probe primarily to the binding of that probe to its intended target mRNA sequence (specific binding). In light of this, Affymetrix introduced a background correction method, termed *mismatch correction*, that simply used the differences, $PM - MM$, as the observed intensity values. However, Irizarry, Bolstad, *et. al.* (2003) pointed out that mathematical subtraction is not necessarily equivalent to biological subtraction. Wu and Irizarry (2004) reported that it is typical to have $var(PM) < var(PM - MM)$, and so an undesirable component of variance is introduced when the MM values are subtracted from the PM values. Also, Naef, *et. al.* (2001) noticed that it is typical to observe a large proportion (30% in their experience) of probe pairs with $MM > PM$. In response to this latter finding, Affymetrix subsequently introduced *ideal mismatch correction*, where the MM value is set to some number arbitrarily close to, but less than, the associated PM value in the cases where $MM > PM$. The aforementioned problems with background correction using the MM values has led several researchers to ignore the MM values entirely and to consider alternative means by which to correct for non-specific binding.

Irizarry, Hobbs, *et. al.* (2003) proposed a convolution model to adjust for background non-specific binding. They assume that the observed PM values may be modeled as $PM = N + S$ where N denotes background and S denotes signal. They consider the background adjustment transformation $B(PM) = \mathrm{E}(S|PM)$, for which they obtain a closed form representation by assuming $N \sim N(\mu, \sigma^2)$ and a strictly positive distribution for S, *i.e.*, $S \sim Exp(\alpha)$.

Naef and Magnasco (2003) and then later Wu, *et. al.* (2004) developed a model for predicting PM intensities from their respective 25-mer nucleotide sequences. Specifically, they use least squares to fit the model

$$PM = \sum_{k=1}^{25} \sum_{j \in \{A,T,C,G\}} \mu_{j,k} I_{b_k = j} \qquad \text{with} \qquad \mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l \tag{1}$$

where $k = 1, .., 25$ indicates position along the probe, j indicates base letter, and b_k indicates the base at position k. The effects $\mu_{j,k}$ are treated as polynomials of degree 3. Note that the MM values may be modeled in a similar fashion. Both sets of authors noted that the effects for A and C were larger than those for G and T, and that the closer the base was to the center of the probe, the stronger the effects of A and C were. Zhang, *et. al.* (2003) had proposed a similar model based on hybridization theory that accounted for interactions between neighboring nucleotides in the probe sequence, however, Naef and Magnasco (2003) noted that these interactions provide little predictive power.

Wu, *et. al.* (2004) noted that the convolution model of Irizarry, Hobbs, *et. al.* (2003) yields background adjusted estimates with low accuracy, most likely due to an inappropriate adjustment for the presence of non-specific hybridization. They therefore proposed to adjust the PM values by a modified MM value that has been shrunk towards the mean of the MM probes with similar fitted values from (1) (they call these fitted values *affinities*).

2.3 Comparing Multiple Chips: Normalization

To motivate the current discussion on the need for normalization, refer to Figure 2 for a scatterplot of the PM probe intensities from two example MGU74A_V2 arrays, with a unit line superimposed. The intensities were determined using the Affymetrix feature extraction algorithm described in Section 2.1. The values have both been log transformed to reduce the inherent skewness in the data.

It is evident from this figure that there is an overall *chip effect* causing the distributions of the PM intensities on the two chips to differ. There is a general disagreement among most authors as to whether normalization should be based on PM only, both PM and MM, or some combination of the two measures, such as PM-MM. There is also a lack of agreement about whether to perform normalization at the probe, probe set, or expression level. Normalization at the probe set level requires that the same transformation be applied to all probes within a given probe set. Expression level normalization means normalizing after gene-specific expression measures have been computed (current approaches to computing these measures are discussed in the next subsection). For the

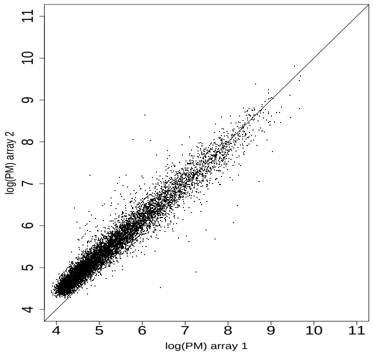


Figure 2: *Scatterplot of log-transformed PM values from two MGU74A_V2 arrays with unit line superimposed showing the need for normalization.*

following discussion, let X_{ij} be the value for the i^{th} array and the j^{th} probe (for probe- or probeset-level normalization) or gene (for expression level-normalization).

The earliest suggested normalization procedure is one termed *scaling normalization*, where an affine transformation is applied to the values on one array so that the mean value on the transformed array equals the mean value on some baseline array. The most commonly implemented scaling method is that supplied in the Affymetrix software, where only a scaling factor is used. If X_{1j} are the values from a baseline array and X_{2j} are values on another array, the scaling factor is computed as $c = \frac{\tilde{X_1}}{\tilde{X_2}}$, where $\tilde{X_1}$ and $\tilde{X_2}$ are the trimmed means from the baseline array and the target array, respectively. The target chip is then transformed via $X'_{2j} = cX_{2j}$. Another naive approach would be to fit a regression equation to the points in Figure 2 and use this to adjust the values, however this method suffers from a dependence on the choice of the baseline chip (*i.e.*, you get different regression lines depending on which chip appears on the x-axis). Alternatively, one could use a principal components axis through the data as a normalizing line.

Dudoit, *et al.* (2002) proposed an alternative graphical representation in the context of two-channel cDNA microarrays in an effort to normalize hybridization intensities between the two color channels. This graphical technique, termed MA-plot, has recently gained popularity in the oligonucleotide microarray literature. For two arrays i and i', the transformation is $M_j = \log\left(\frac{X_{ij}}{X_{i'j}}\right)$ and $A_j = \frac{1}{2}\log(X_{ij}X_{i'j})$, which when the M_j are plotted against the A_j, yields a 45-degree rotation of the scatterplot in Figure 2. Although their application was to cDNA microarrays, Sellers, *et al.* (2004) proposed fitting a normalizing line to the log-scaled data by finding values a, b, and c such that the non-linear objective function $\sum_j \left[\log\left(\frac{X_{ij}-a}{X_{i'j}-b}\right) + c\right]^2$ is minimized.

Several authors have recently observed that the intensities on two chips may often exhibit a non-linear relationship. To circumvent this issue, Bolstad (2003) introduces *quantile normalization*, where multiple arrays are simultansously transformed so that the distributions of their intensities are the same. Astrand (2002) introduced *contrast normalization*, where for I chips, the logged intensity values are transformed using an $I \times I$ orthonormal transformation matrix. Normalization is performed in this new basis using localized regression (*loess*), and then transformed back to the original basis. A few authors (Kepler, *et al.* (2002), Schadt, *et al.* (2001), and Li and Wong

(2001a)) have proposed non-linear normalization methods that are based on a subset of the probes. Each of these authors determine a subset of probes that are approximately rank-invariant between two arrays. Rank invariance means that the relative ranks of an intensity measure in two chips are essentially the same. Kepler, *et al.* (2002) use loess curves, while Schadt, *et al.* (2001) use smoothing splines, and Li and Wong (2001a) use piecewise running median lines.

The concept of performing normalization on a self-selected subset of intensity values described in the previous paragraph is closely related to normalization based on a set of *housekeeping* probes, or a set of spiked-in cRNA control probes. These procedures select the values to be used in normalization *before* the data are collected. Housekeeping probes are typically selected to be non-complementary to the target mRNA sample, and so their relative true hybridization levels are expected to be constant across arrays. In the case of spiked-in control probes, specific concentrations of complementary mRNA are spiked-in to the hybridization cocktail, and so the relative true hybridization levels of these probes are assumed to be known. Hill *et al.* (2001) uses a non-linear normalization technique based on predetermined spiked-in control probes. Hartemink *et al.* (2001) proposed an additive error model for I chips and J spiked-in controls:

$$\log(X_{ij}) = \mu_j + \rho_i + \epsilon_{ij},$$

where μ_j is the true log-transformed expression level for spiked control j, ρ_i is a log-transformed scaling factor for array i, and $\epsilon_{ij} \sim N(0, \sigma_i^2)$. Note that this model implicitly assumes a multiplicative error model for the untransformed data. They compute optimal scaling factors, ρ_i, by imposing a flat prior on the σ_i^2 and using maximum a posteriori estimation (MAP).

A few authors (Bolstad, *et al.* (2003), Quackenbush (2002), and Hoffmann, *et al.* (2002)) have performed empirical studies in an effort to compare various proposed normalization methods. In particular, Bolstad, *et al.* (2003) compared the Affymetrix scaling algorithm, the non-linear methods of Li and Wong (2001a) and Schadt, *et al.* (2001), the contrast normalization method in Astrand (2002), and the quantile normalization method proposed in their current paper. They reported that the quantile normalization method outperformed the others based on bias and variance across several arrays in a spiked-in control probe experiment. In their paper, they also introduce a classification of normalization methods into ”complete data” methods and ”baseline” methods, where those falling into the latter category suffer from their dependence on the choice of a baseline array when normalization is performed on more than two arrays.

2.4 Computing Gene-Specific Expression Measures: Summarization

Recall that probes on the array are grouped together into probesets that correspond to particular genes. Each probe pair (PM and MM) in a probeset (typically there are 11-20 such pairs on an Affymetrix GeneChip) interrogates a different region of the gene of interest. Bolstad (2004) refers to the process of combining the probe values into a gene-specific expression measure as *summarization*, or *probe-level analysis*. The objective of summarization is to estimate the unobserved gene presence in the sample. There have been several proposed summarization methods in recent years, and below we review a few of them. All of the methods discussed in this section assume that the probe values have undergone a normalization preprocessing step. Some of the methods also assume that the data have been previously background corrected while others do not.

2.4.1 Affymetrix HGU133A Latin-Square Spike-In Study

The major difficulty with summarization is that we do not know the true amount of gene species in the sample that we are trying to estimate, therefore it is difficult to validate any proposed summarization method. In an effort to foster collaborative efforts with the research community, Affymetrix has made publicly available a database consisting of 42 HGU133A (human genome) arrays that can be used for this purpose. A total of 42 Human cRNA fragments matching specific probesets (genes) on the array were added (spiked-in) to the hybridization mixtures for different arrays at 14 different levels of concentration ranging from 0 to 512 pM (the 14 concentrations were tapered as 0,0.125,0.25,0.5,...,512 pM). Each concentration level was replicated 3 times, yielding 42 arrays. The assignment of concentration level to probeset and to array was performed using a cyclic latin square design. Refer to Figure 3 below for profile plots of the log-transformed PM values from three of the genes (207641_at, AFFX-ThrX-5_at, and AFFX-HUMISGF3A/M97935_MB_at) from the HGU133A array. The first two probesets were spiked-in and the third was not, so we have a reasonable understanding as to what the signal should look like for the first two, whereas we should expect no signal from the third. The x-axes in the plots are an ordinal representation of the probes in order from the 5' end to the 3' end of the gene. In the left two panels of Figure 3 we can clearly see a separation between the probe hybridization levels on arrays with different spike-in concentrations. The three replicate arrays per concentration level are also evident in these two plots. In the right panel, it appears that there is no difference in hybridization levels among the arrays, at least relative to the level of variability across probes. It appears that in all three plots, the probe profiles for each array follow virtually identical patterns. Note that the first array has 11 probe pairs, while the second two have 20 probe pairs the number of probe pairs used to interrogate a gene varies across genes within a chip and across chip type).

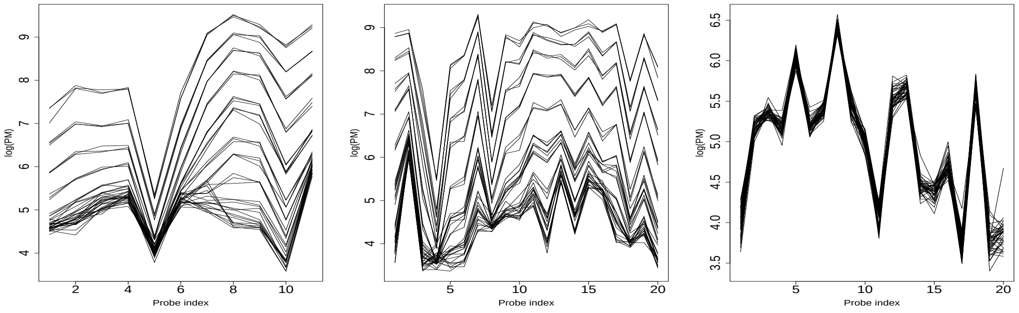


Figure 3: *Probe profile plots for three probesets from the Affymetrix HGU133A spike-in study. The left two panels are spike-in genes, while the right panel is not.*

2.4.2 Summarization Methods Proposed by Affymetrix

Affymetrix (1999) had initially proposed using a log-transformed trimmed mean estimate, termed AvDiff, for gene expression. In the initial implementation of the measure, the values were first normalized using scaling normalization (discussed in Section 2.3) and background corrected using the location-specific correction and the mismatch adjustment discussed in Section 2.2. Let Y_{ij} be the normalized and background corrected values for array i and probe pair j from a particular gene.

Then the expression estimate is

$$\hat{\theta}_i = \log\left(\frac{\sum_{j=2}^{J-1} Y_{i(j)}}{J-2}\right)$$

where J is the number of probe pairs in the probeset and $Y_{i(j)}$ is the j^{th} order statistic in the sample.

In subsequent iterations of their software, Affymetrix introduced another robust average estimator that utilizes a 1-step Tukey biweight (see Affymetrix (2002) and Hubbell (2002)). This estimate is

$$\hat{\theta}_i = \frac{\sum_{j=1}^{J} w(Z_{ij}) \log(Y_{ij})}{\sum_{j=1}^{J} w(Z_{ij})}, \tag{2}$$

where $w(z)$ is the bisquare function ($w(z) = (1-z)^2$ when $|z| \leq 1$; $w(z) = 0$ otherwise), and $Z_{ij} = (\log(Y_{ij}) - M)/(cS + \epsilon)$.

Here M is the median of the $\log(Y_{ij})$ and S is the associated median absolute deviation (MAD). The constants c and ϵ are adjustable tuning parameters. Note that the current implementation of the software considers the Y_{ij} to be normalized and background corrected using location-specific correction and the ideal-mismatch adjustment. Bolstad (2004) pointed out that this estimator is a particular case of the larger class of M-estimators.

2.4.3 Model-Based Expression Index (MBEI)

Using their own experimental database consisting of 21 Affymetrix HuGeneFL arrays, Li and Wong (2001a, 2001b) noticed that the variation in the probe values (PM, MM, or $PM - MM$) within a probeset was several orders of magnitude greater than the variation of a particular probe value across arrays. This feature is evident in the right panel of Figure 3. They concluded that the probe values are highly "reproducible" and "predictable", and so to capture the apparent probe-specific effects they introduced the following multiplicative multi-chip model for a particular gene:

$$\begin{aligned} MM_{ij} &= \nu_j + \theta_i \alpha_j + \epsilon_{ij} \\ PM_{ij} &= \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \epsilon_{ij} \end{aligned}$$

where PM_{ij} and MM_{ij} are the perfect match and mismatch values, respectively, for the i^{th} array and j^{th} probe pair in the probeset. In this model, θ_i acts as an expression index for array i, ν_j is the baseline value for the j^{th} probe pair due to non-specific hybridization, α_j is the rate of increase of the MM value in response to changes in the expression index level, and ϕ_j is the corresponding rate of increase for the PM value. The ϵ_{ij} are assumed to be independent normal errors, $\epsilon_{ij} \sim N(0, \sigma^2)$. Li and Wong (2001a) also report fitting an additive model to the data, however, this invariably yielded nonlinear residuals, suggesting non-additivity.

They proceeded by combining the two components of the above model into the following simplified model:

$$Y_{ij} = T(PM_{ij}) - T(MM_{ij}) = \theta_i \phi_j + \epsilon_{ij} \tag{3}$$

where $T()$ represents their piecewise running median line normalization transformation based on a rank-invariant subset of the probes discussed in Section 2.3. For identifiability of the parameters, they impose the constraint $\sum_j \phi_j^2 = p$, where p is the number of probe pairs in the probeset. Estimation is performed using least squares by first assuming the θ_i are known and estimating the ϕ_j. Then the ϕ_j are assumed known and the θ_i are estimated. This process is iterated until convergence. Li and Wong (2001a, 2001b) provide standard error calculations for the estimates, $\hat{\theta}_i$, and report that the results from their model can be used for the automatic determination of outliers and contaminated regions of the array. They also report that they obtain reasonable estimates by simply ignoring the MM values and fitting $Y_{ij} = T(PM_{ij}) = \theta_i \phi_j + \epsilon_{ij}$.

2.4.4 Robust Multi-Array Analysis (RMA)

Holder *et. al.* (2001) proposed a similar model to that of MBEI to estimate array-specific gene expression levels, however, they proposed an additive model on data that is suitably transformed, and they also allow for replicate chips within each experimental condition. Their model is

$$Y_{ijk} = T(PM_{ijk} - MM_{ijk}) = \theta_{ik} + \phi_j + \epsilon_{ijk} \tag{4a}$$

where now j indexes the probe-pair, i indexes the experimental condition, and k indexes replicate arrays within experimental condition, yielding ik total arrays. The ϵ_{ijk} are independent normal random errors. Here $T()$ denotes a scaling normalization on the $PM - MM$ values plus a linear-log hybrid transformation that they claim homogenizes the variability across a wide range of expression levels. They use a robust median polish algorithm to estimate the parameters (see Tukey (1949), for example), with the parameter constraint $\sum_j \phi_j = 0$. One possible drawback to this procedure is that it is not possible to perform standard error calculations on the estimates.

Irizarry, Bolstad *et. al.* (2003) and Irizarry, Hobbs *et. al.* (2003) proposed the following additive model that they termed Robust Multi-Array Analysis (RMA):

$$Y_{ij} = \log(B(T(PM_{ij}))) = \theta_i + \phi_j + \epsilon_{ij} \tag{4b}$$

where $T()$ represents the quantile normalization transformation discussed in Section 2.3 and $B()$ represents the convolution model background adjustment described in Section 2.2. They also use the robust median polish technique for fitting the parameters. As noted in Section 2.2, the convolution model for background adjustment yields imprecise estimates, and so Wu, *et. al.* (2004) recently modified the RMA model (now called GC-RMA) so that the transformation $B()$ is the background adjustment that utilizes the fitted probe affinities from eq. (1).

3 Preliminary Results

3.1 Raw Data

Through my personal empirical experience, I have found several issues of concern related to the Affymetrix feature extraction algorithm. Firstly, it rarely, if not never, occurs that the true edge of a probe cell on the physical medium gets matched exactly to the edge of a pixel in the scanned image. Take, for example, the MGU74A_V2 chip displayed in Figure 1. The Affymetrix MGU74A_V2 line of chips have $640 \times 640 = 409,600$ square cells. However, the resolution of the resultant scanned image is such that this array is represented by approximately 4260 pixels in the horizontal direction and 4264 pixels in the vertical direction. This means that each cell is approximately 6.66×6.66 pixels in the scanned image. This also implies that the landmark corners of the chip should be reported at the sub-pixel level, and not at the pixel level as is currently the case (the true corners of the array are most likely not aligned with any specific discrete pixel location). Another implication is that the Affymetrix feature finding algorithm incorrectly assumes that the center of each cell can be represented by a single pixel location. The cell centers should in fact also be reported at the subpixel level.

In addition to the cells being represented by a square grid with non-integer number of pixels as their dimensions, it is typical for an array to be rotated slightly in the image. This is evident in the chip displayed in Figure 1. Take for instance the six *B2 oligo* probe cells (white spots) along the top edge of the upper left panel of the figure. The top edges of the two left-most cells appear to coincide reasonably well with the border of the image, while the top edge of the right-most cell appears to be shifted down in the image by about one pixel. This array is rotated clockwise in the image by approximately 0.81 degrees. The implication here is that the true edges of the cells are not parallel to the edges of the pixels in the image. Fianlly, the array in the image is not even a parallelogram.

In the following I propose an alternative feature finding algorithm to the current Affymetrix standard that deals with the issues discussed in the preceding paragraphs. Generally speaking we wish to apply a transformation to the raw image that maps the four landmark points to four points that define a perfect square (*i.e.*, force the x and y dimensions of the chip to be of equal length, and to have the angles formed by the four sides of the square be 90 degrees). In addition, we want the dimensions of the cells to be an odd integer number of pixels (so that the cell centers correspond to a single pixel). To preserve the structure of the image in the interior of the polygon defined by the landmark points, we restrict our attention to transformations that are affine.

An affine transformation in two dimensions is not possible using four landmark points; it in fact requires exactly three. It is therefore proposed that we split the array within the image into two disjoint triangles and apply an affine transformation independently to each triangle. At this point we will regard the choice of how to split the array as arbitrary (along the diagonal or off-diagonal). To proceed, we perform each affine transformation within the constructs of a homogeneous coordinate system. This provides more flexibility than using a cartesian coordinate system, where we are restricted to rotation, translation, and scaling. In a homogeneous system, we can perform shearing, which will allow us to rectify the fact that the array is not a perfect rectangle in the image. A homogeneous coordinate system in 2D involves the addition of a third dimension, w, and given a triple (u, v, w) in homogeneous coordinate space, the corresponding point in cartesian coordinate space is $(u/w, v/w)$. The point $(u/w, v/w)$ is a projection through the origin of the point (u, v, w)

onto the $w = 1$ plane. Given a point in cartesian coordinate space (x, y), one possible homogeneous coordinate is $(x, y, 1)$, which we will use hereinafter.

Let the cartesian coordinates of the landmark points of a particular triangle be (x_i, y_i), $i = 1, 2, 3$, and let the corresponding target points for the landmarks be (x'_i, y'_i), $i = 1, 2, 3$. If we let

$$A = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} x'_1 & x'_2 & x'_3 \\ y'_1 & y'_2 & y'_3 \\ 1 & 1 & 1 \end{pmatrix}$$

then the homogeneous transformation matrix, T, is computed via

$$T = BA^{-1}.$$

The image is transformed separately for the two homogeneous transformation matrices, call them T_{11} and T_{21}, computed for the two triangles using the landmark points output by the Affymetrix software. The mapped pixels from the interiors of the two triangles are then combined to form a final transformed image. One result from this process is that the transformed pixels that lie along the diagonal (or off-diagonal) are equivalent in both of the transformed images. In other words, we have preserved C^0 continuity in the image along the diagonal (or off-diagonal). The choice of which transformed triangle the pixels along the diagonal (or off-diagonal) should come from is therefore arbitrary.

A plethora of methods exist to perform an affine transformation on an image, but the common denominator is that they must in some way approximate the pixel values using a 2-dimensional interpolation function (note that the pixel values themselves may be interpreted as a 2-dimensional step function). The transformed image is then created by sampling from the interpolation function at the desired locations. I have chosen to work with the family of Catmull-Rom (Catmull and Rom (1974)) interpolating splines in the following discussion. One attractive feature of these interpolants is that they are constrained to pass through the control points (pixels) in the original image, so that if a point is sampled from the function at a location corresponding to an original pixel location, the pixel value from the original image will be recovered.

As an example, I applied the aformentioned transformation to the chip displayed in Figure 1 using two triangles formed by drawing a line from the upper right corner to the bottom left corner of the array. The landmark points in the original image stated that the array was approximately 4264 pixels by 4260 pixels. I applied transformations that mapped the landmarks so that they defined a 4480×4480 square grid, meaning that both dimensions of each cell are $\frac{4480}{640} = 7$ pixels.

Li and Wong (2001a) noted that in their experience, it it possible for the reported landmark points to be off by as much as a full feature (several pixels). Due to the fact that the landmark points supplied by the Affymetrix software may not be precise, the corresponding landmarks in the transformed image described thus far will be imprecise as well. The top panel of Figure 4 displays a contour plot of the 28×28 (pixels) upper left subimage of the transformed image, which should include the checkerboard pattern of bright *B2 oligo* features. Overlayed in the graph are grid lines that should define the boundaries of the probes. It is evident from this display that there is a slight amount of misalignment, which is attributable to the original landmark point for this corner of the array being incorrectly determined.

I therefore propose a two-stage procedure to correctly align the array in the transformed image. The first stage is to perform the piecewise transformation as described above. Next, I use the eight bright checkerboard features in each of the four corners of the array to estimate the displacement vector resulting from the imprecise determination of the original landmarks. Each vector (one for each corner) is estimated by minimizing the Mean Square Error (MSE) of the difference between a transformed version of the subimage (28 × 28 pixels in the example) containing the checkerboard pattern of features and an idealized reference image of the same size. The reference image is constructed by imputing the maximum pixel value from the original 28 × 28 subimage into the regions where we expect the bright probes to occur, and the minimum pixel value everywhere else. See the bottom right panel of Figure 4 for the subimage from the upper left corner of the array (same as top panel), and the bottom left panel for the associated reference image. Both of these images have been padded with zero values to allow for displacements outside the boundary of the original image.

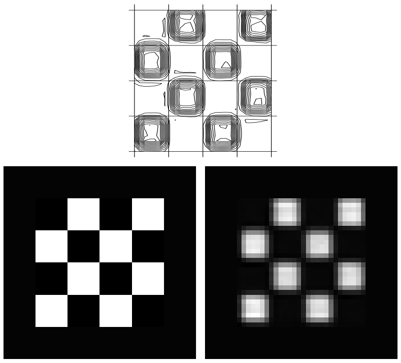


Figure 4: *Top panel: contour plot of upper left* 28 × 28 *pixel subimage of the array following first stage of image warping. Bottom right panel: image plot of same data as in top panel. Bottom left panel: target reference image.*

The minimization is performed iteratively by applying shift transformations using Fourier interpolation until convergence of the MSE (see Eddy, *et. al.* (1996)). More specifically, let x and y index the two dimensions of the image, and let $I'(x, y)$ be the value of the pixel in the reference image at (x, y). Let $u, v \in \mathbb{R}$ and let $I(x + u, y + v)$ be the value of the sampled point on the Fourier interpolation function defined at $(x + u, y + v)$. Then the idea is to find the displacement vector (u, v) that satisfies

$$\text{argmin}_{u,v} \sum_x \sum_y \left(I'(x, y) - I(x + u, y + v)\right)^2 .$$

Once the displacement vectors for the four corners are determined, we create two new transformation matrices (one for each of the triangles), T_{12} and T_{22}, that map the transformed landmarks from the first stage to points defined by the estimated displacement vectors. Instead of introducing additional errors from mapping each of the triangles twice (and hence applying two approximating functions), we apply a single transformation to each triangle using the composite homogeneous transformation matrices

$$T_1 = T_{12}T_{11} \qquad \text{and} \qquad T_2 = T_{22}T_{21}.$$

The four left panels in Figure 5 below show the same 84 × 84 pixel subimages from the array displayed in Figure 1, and the four panels on the right display the 84 × 84 pixel subimages after applying the aforementioned two-stage transformation. Note that in the four right panels of this figure, the *B2 oligo* probe cells are aligned well with the edges of the images. Also note that each probe cell is now 7 × 7 pixels, and so each image contains exactly $\frac{84}{7} = 12$ cells in both directions. Here we have addressed each of the concerns described at the beginning of this section.

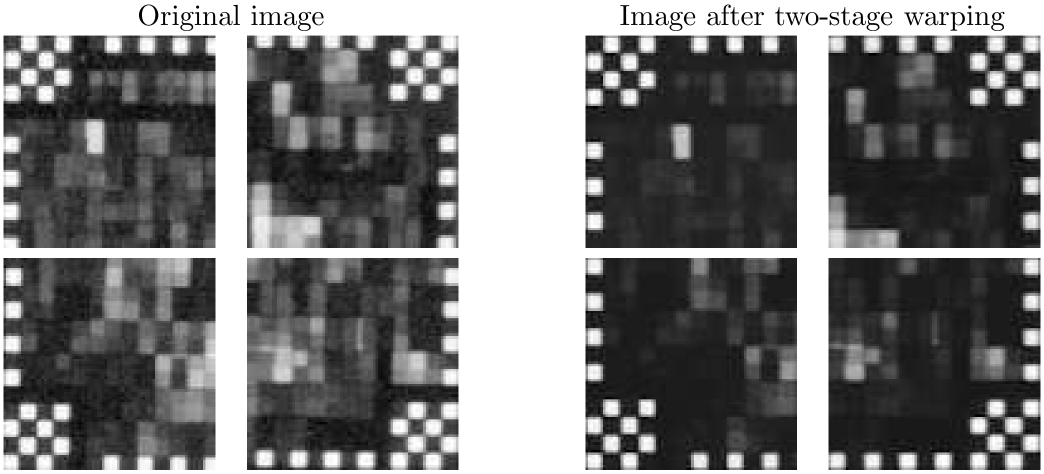


Figure 5: *Comparison of the four corner* 84 × 84 *pixel subimages in the raw image (left four panels) and in the image following the two-stage warping procedure (right four panels).*

Intuitively, this proposed transformation process will facilitate subsequent computation by allowing one to easily locate each probe of interest in the image. In an effort to quantify the added benefit of this process, I computed the coefficient of variation (CV) among the pixel values in a 5 × 5 pixel region centered at the center pixel for each probe in the transformed image. I then computed the CV for each probe in the untransformed raw image using the Affymetrix approach of using a weighted average of the landmark points to determine the probe center pixels. For the latter computation, I used a weighted CV within a 4.75 × 4.75 pixel region centered at the probe centers. A window size of 4.75 × 4.75 was used here since the probes are approximately 6.66 × 6.66 pixels in size. Figure 6 below shows a scatterplot of the CV values from both approaches, with a unit line superimposed. It is evident that the current proposed transformation statistically significantly reduces the variability in the pixel values comprising a probe (t-statistic of 121.3 testing for equality of means of the two distributions on the log scale). One possible explanation for this observation is that the array in the original image is rotated slightly, which means that pixels used in a CV computation for a specific probe may be contaminated by a contiguous probe. We have assumed in this process that pixels within a probe region posess less variability than do pixels across probes, and so our two-stage warping procedure has improved the alignment of the array within the image.

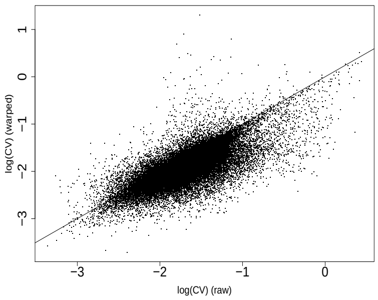


Figure 6: *Scatterplot of log-transformed coefficient of variation (CV) values comparing variability among pixels in probe cells in the raw image (x-axis) and in the image following the two-stage warping procedure (y-axis). Notice that the majority of the points fall below the unit line.*

3.2 Summarization

Several methods for estimating gene expression measures were reviewed in Section 2.4. The general consensus among researchers is that when possible, the use of multiple chips in constructing gene expression estimates is highly advisable. Although the Affymetrix estimator (eq. 2) is designed to be robust to outlying probe values, estimation is performed on each array individually, and therefore does not borrow strength from the possibly multiple arrays comprising an experiment. In the present discussion we follow a multi-array modeling approach similar to that of the MBEI model (eq. 3) of Section 2.4.3 and the RMA model (eqs. 4a and 4b) of Section 2.4.4. In the remainder of this section, I will show that the MBEI model can be described as a factor analysis model whose solution can be obtained via the singular value decomposition (SVD) of the data matrix, instead of the iterative approach adopted by Li and Wong (2001a). I introduce one possible extension to their model within this factor analysis framework. I also introduce a simple extension to the RMA model, and in both cases show that the extensions provide a better fit to the data. The better fits come at the cost of needing to estimate more parameters, however, it will also be shown that the newly proposed extensions can overcome possible violations of error structure assumptions made by the MBEI and RMA models. The data used for this comparison study is the Affymetrix HGU133A Spike-In study described in Section 2.4.1.

To show the equivalence of the MBEI model to a factor analysis model, consider the following theorem due to Eckart and Young (1936):

Theorem (Eckart-Young) *Let Y be an $n \times p$ matrix of rank k. Let the singular value decomposition (SVD) of Y be*

$$Y = U\Sigma V^T$$

where U and V are $n \times k$ and $p \times k$ orthonormal matrices, respectively, and Σ is a diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$. Then for $r < k$, the solution to

$$\min_{Z \in \mathbb{R}^{n \times p} | \text{rank}(Z) = r} \sum_{i=1}^{n} \sum_{j=1}^{p} (Y_{ij} - Z_{ij})^2$$

is $Z = U_r \Sigma_r V_r^T$, where $\Sigma_r = diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and U_r and V_r are comprised of the first r columns of U and V respectively. Furthermore, the minimized value is $\sum_{i=r+1}^{k} \sigma_i^2$.

For a proof of this theorem, refer to Golub and Van Loan (1996). Gabriel (1978) provides alternative formulations of the objective function. The theorem states that the best rank-r approximation to Y, in a least-squares sense, is obtained via the rank-r version of the SVD of Y. It also provides a measure of distance (or error) from the original matrix to its approximation.

The Eckart-Young theorem is typically used in factor analysis where it is of interest to approximate an $n \times p$ matrix Y by $\hat{F}\hat{\Lambda}^T$, where $\hat{F}$ is $n \times r$ and $\hat{\Lambda}$ is $p \times r$. Here the rows of $\hat{F}$ are estimated scores, and the columns of $\hat{\Lambda}^T$ are estimated loadings. Croux *et. al.* (2003) note that, without constraints imposed on the matrices, $\hat{F}$ and $\hat{\Lambda}$ are not unique in that $\hat{F}\hat{\Lambda}^T = (\hat{F}T^T)(\hat{\Lambda}T^{-1})^T$ for any nonsingular $r \times r$ matrix T. Note though that the fitted matrix $\hat{F}\hat{\Lambda}^T$ is unique.

In terms of the MBEI model, the goal is to find the rank-1 matrix, $\hat{\theta}\hat{\phi}^T$, that minimizes $\|Y - \theta\phi^T\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} (Y_{ij} - \theta_i \phi_j)^2$. Given the decomposition $U\Sigma V^T$ of Y, the solution is $\hat{Y} = \hat{\theta}\hat{\phi}^T = \sigma_1 U_1 V_1^T$. We still have the problem of estimating $\hat{\theta}$ and $\hat{\phi}$ due to the identifiability issue raised in the previous paragraph. However, given the constraint $\sum_{j=1}^{p} \hat{\phi}_j^2 = p$, and the fact that $\sqrt{\sum_{j=1}^{p} v_{1,j}^2} = 1$ ($v_{1,j}$ is the j^{th} element of V_1), we have $\sum_j \hat{\phi}_j^2 = \sum_j (\sqrt{p} v_{1,j})^2 = p \sum_j v_{1,j}^2 = p$. So, pairing the elements, we have $\hat{\phi}^T = \sqrt{p} V_1^T$. Therefore, the vector of array-specific effects is $\hat{\theta} = \frac{1}{\sqrt{p}} \sigma_1 U_1$, which is equivalent to the estimates obtained using their iterative algorithm.

It is not uncommon for subsequences of two or more probes to interrogate the same section of a gene, and in some extreme cases two probes may overlap by 24 nucleotides. When probe sequences overlap, we have observed that this introduces correlation, and the level of correlation is related to the level of overlap. For instance, all sets of probes that overlap by 24 nucleotides in one particular MGU74A_V2 array posess a correlation coefficient of 0.96. All sets that overlap by 15 nucleotides in the same array exhibit a correlation of 0.32. Since principal components analysis is a standard method for modeling correlation, I propose to extend the MBEI model by considering alternative decompositions of the form $Z = U_r \Sigma_r V_r^T, \ r \geq 2$.

The RMA model (eq. 4) has a simple ANOVA structure in which it is implicitly assumed that the effects are additive. The ϵ_{ij} act as residuals, though consider the case where there is structure in the residuals and we can model them through a factor model akin to the discussion earlier in this section. Then we have the simple additive plus multiplicative model extension to the RMA model:

$$Y_{ij} = \theta_i + \phi_j + \sum_{k=1}^{r} \sigma_k f_{ik} \lambda_{jk} + \tau_{ij}, \tag{5}$$

where we assume the τ_{ij} are normally distributed random errors. Croux and Filzmoser (1998) developed a robust fitting algorithm for this model that is a generalization of the median polish technique for the purely additive model. Their algorithm fits the model by applying alternating weighted L_1 regressions, weighted so that sensitivity to leverage values (influential factors or scores) is minimized. According to Croux and Filzmoser (2003), the median polish technique can be shown to approximate an L_1 fit of a purely additive model. So, we can view equation (5) as a direct extension of the RMA model if we use the aforementioned robust fitting procedure.

In the following, I present results from the MBEI model (PM-only version), the decomposition $Z = U_r \Sigma_r V_r^T$ (with $r = 2$), a simple additive ANOVA model, the RMA model, and the RMA extension model (eq. 5) with $r = 2$. In particular, I chose a representative spiked-in gene (207641_at, same gene as displayed in the left panel of Figure 3) to fit these models to. Values across replicate arrays have been averaged for the current exercise. As a comparative exercise, all models will be fit to the $\log_2$ transformed data after normalizing the arrays via the rank-invariant procedure developed by Li and Wong (2001a). Although the MBEI model was designed to be fit to the untransformed data, empirical evidence suggests that a multiplicative model may be appropriate on the log scale for some genes. Refer to Figure 7 for profile plots of the fitted values from each model, and an image plot of the corresponding residuals. The residual matrices have probes as their columns and arrays as the their rows (first row is the array with 0 pM concentration, second with 0.125 pM, and so on down to the last row which represents the array with concentration 512 pM).

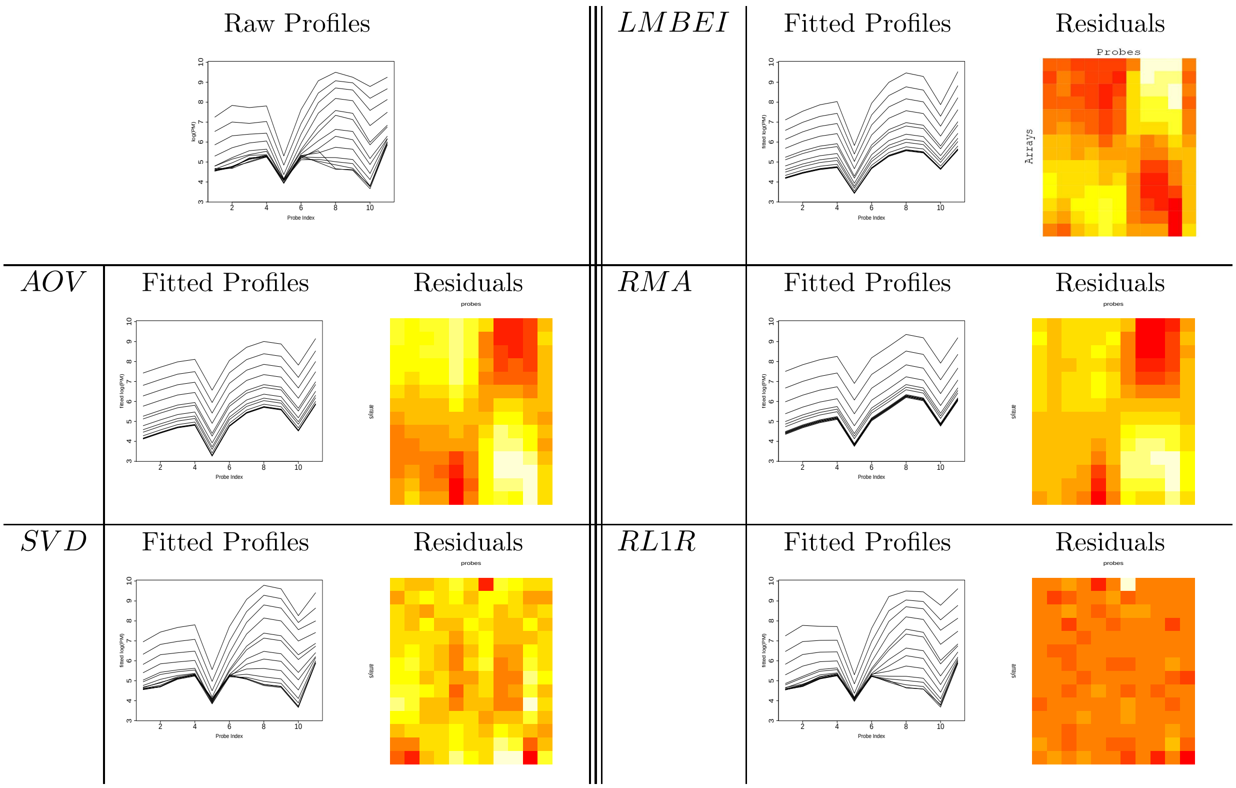


Figure 7: *Top left panel: original probe profiles for probeset 207641_at on the log scale. Remaining panels show fitted values and residuals from five models: MBEI on log scale (LMBEI), ANOVA (AOV), RMA, SVD model with $r = 2$ (SVD), and robust L_1 regression of additive plus multiplicative model with $r = 2$ (RL1R). Residual matrices have probes as columns and arrays as rows, where the first row is array with 0 pM concentration, and subsequent rows have increasing concentrations.*

Paying particular attention to the residual matrices, the most striking feature of the MBEI, ANOVA, and RMA models is the clearly evident block structure in the errors. This implies that estimates from these models may be biased and that the assumption of uncorrelated errors is clearly violated. The two extension models are more effective at capturing the correlation structure of the data, however, this comes at the cost of estimating additional parameters. An additional concern is interpretability of the model parameters. For instance, with the $Y = U_2 \Sigma_2 V_2^T$ model, we no longer have a single value, θ_i, as an array-specific expression estimate, we now have a two-dimensional projection as the array-specific estimate. I also fit the MBEI model to the untransformed data and observed the same block error structure.

4 Proposed Work

The goal of this dissertation is two-fold, the first being to conduct a thorough investigation into several aspects of the image processing stage of the analysis. While it has been shown that it is possible to reduce the variability in the pixel values that comprise a probe cell by applying an appropriate transformation of the raw input image, other aspects of image processing warrant further consideration and research. Firstly, related to the current proposed array alignment algorithm is the choice of sampling function to use in the transformations. The Catmull-Rom interpolating spline has been used thus far, but other methods exist such as bezier curves, additional spline based methods, and even Fourier-based methods. I propose to investigate to the greatest degree possible how the choice of interpolating function impacts the final warped image. With any interpolating function comes a degree of inaccuracy, as the function serves only as an approximation to the underlying source being imaged (note though that the pixels comprising the image are themselves an approximation to the true nature of the imaged object).

As mentioned in section 2.1, the Affymetrix software by default outputs the 75^{th} percentile of the pixel values comprising a probe cell as an estimate of the hybridization level of the probe to the target mRNA. Although a rigorous justification of this choice is unkown to me, it is apparent that this summary measure was chosen to be robust to outlying pixel values in the probe region. Other researchers, such as Schadt *et. al.* (2001), have proposed using the mean pixel value from the probe region. It is apparent that a thorough study of which measure is most appropriate is lacking in the literature. One way to proceed with an assessment of such univariate summary measures is to utilize replicate arrays and investigate how the measures behave across arrays in terms of accuracy and precision.

An implicit assumption in using a univariate statistic on the pixel values to measure the hybridization level of a probe is that the pixel values are independent realizations of the true hybridization level on the chip. We have noted that this may not entirely be true, and that the pixel values may actually exhibit spatial correlations. Take for instance the contour plot displayed in Figure 4. The regions corresponding to the bright control (*B2 oligo*) probes posess a mound-shaped characteristic, and it is clear that the pixel values are spatially correlated. It is hypothesized that probe regions with different true hybridization levels will exhibit different spatial structures. It is my intention to invesitigate the appropriateness of taking these spatial correlations into account when computing a summary measure.

The choice of representing the hybridization level of a probe with a univariate statistic seems itself arbitrary. It seems to me appropriate to at least consider alternative ways of representing the data. An alternative could include carrying all of the pixel values for a probe forward through all subsequent analyses of the data and treating the observed bybridization level of a probe as a multi-dimensional object. Another alternative could be to use a univariate statistic where the pixel values are weighted according to their location within the probe region. Still yet another alternative would be to project the pixel values onto a lower dimensional space using principal components analysis, which would preserve the spatial correlation structure among the pixels.

The second component of this dissertation is to focus on the assessment of differential gene expression between two or more experimental conditions. I have broken this down into several subcategories: experimental design (how many samples/arrays do we need to collect so that we may obtain reliable and precise estimates of differential expression?), background correction, summarization, and choice of differential expression metric. I have ommitted normalization from this list because I do not intend to perform any research on this topic other than to gauge what impact the choice of normalization procedure has on the other components of the analysis.

For the background correction component, I propose to investigate the utility of the binding affinity model, proposed by Naef and Magnasco (2003) and Wu, *et. al.* (2004), in adjusting for non-specific binding and cross-hybridization. I have verified the results reported by these authors on our own set of MGU74A_V2 arrays, though this topic appears to warrant further research. As stated in Section 3.2, it often happens that probes that interrogate the same gene overlap and are correlated. For pairs of probes that maximally overlap (*i.e.*, share 24 out of 25 nucleotides), their correlation is profound and we can to some degree view these probes as replicate interrogation sequences.

Another area of interest is that we have observed a nominal relationship between probe hybridization level and its distance (in nucleotides) to the 3' end of the gene. We have observed that the set of probes that are closest, among all probes in their respective probeset, to the 3' end of the gene yield stiatistically significantly lower hybridization levels than the set of all other probes. In any event, it is clear that non-specific binding and cross-hybridization are components of variation that must be accounted for when assessing differential gene expression.

For the summarization component, I propose to develop a gene expression summary measure that has desirable statistical properties, namely that the measure should be unbiased and posess as little variability as possible. The Affymetrix latin-square spike-in study discussed in Section 2.4.1 provides a suitable reference database on which to test various expression measures in terms of bias and variance. Cope, *et. al.* (2003) have developed a webtool where researchers can submit their expression measure estimates and have results reported back to them. Some of the assessment results include variability within replicate arrays, accurate estimate of fold change for spike-in genes (compared to expected fold change computed from the respective mRNA concentrations), and a correct determination of no fold-change for genes that were not-spiked in.

The aforementioned webtool unfortunately requires that you compute a single-valued gene expression estimate for each array. I have proposed a more complex gene expression measure that is based on a principal component projection of order $r \geq 2$ (see Section 3.2). This type of measure is not applicable to the webtool, so if I can determine that it is superior to other single-valued measures, I will construct an alternative metric for the quantification of differential expression. If, however, a single-valued expression estimate exists that posesses desirable statistical properties, then the fold-change metric for the assessment of differential expression will most likely suffice. With my proposed model, I have shown that currently existing modeling approaches have serious assumption violations in terms of their error structure. It is clear that more work needs to be done in the area of a proper handling of the probe correlation structure. It is rather straightforward to estimate the correlation structure directly from the data, and so this may be a useful component that can be built into a model.

My proposed additive plus multiplicative model also handles the correlation structure better than current modeling techniques, though it requires the fitting of additional parameters. When appropriate, I intend to use the block structure of the residuals from say the simple additive fit to determine a small number of additional parameters necessary to handle the error structure, instead of the full $2(n+p)$ additional parameters required by the current proposition. Starting with an additive fit, I can also look at the traditional One Degree of Freedom Test for Nonadditivity (see Tukey (1949)) to test for possible multiplicative effects. It must be noted here that the applicability of either an additive or multiplicative model is dependent on the particular data transformation employed. It is quite possible that the logarithmic transformation is not always appropriate, and so I intend to find a transformation for the data that yields homogeneous variances across different levels of the data.

Though the models proposed in Section 3.2 show some promise, I will not restrict my search to those models since a thorough investigation of their statistical properties has yet to be carried out. While I have fitted my proposed additive plus multiplicative model robustly, the SVD based model suffers from its sensitivity to outliers. I will therefore investigate additional fitting techniques based on robust methods (see for instance, Baccini, *et. al.* (1996), Pison, *et. al.* (2003), and Daigle and Rivest (1992).)

5 References

Affymetrix. (1999). *Affymetrix Microarray Suite Users Guide, version 4.0.* Affymetrix, Santa Clara, CA.

Affymetrix. (2001). *Affymetrix Microarray Suite Users Guide, version 5.0.* Affymetrix, Santa Clara, CA.

Affymetrix. (2002). *Statistical Algorithms Description Document.* Affymetrix, Santa Clara, CA.

Astrand, M. (2002). Contrast Normalization of Oligonucleotide Arrays. *J. Comp. Biology*, 10(1). pp. 95-102.

Baccini, A., Besse, P., de Falguerolles, A. (1996). A L_1-norm PCA and a Heuristic Approach. In Diday, E., Lechevalier, Y., Opitz, O., eds., *Ordinal and Symbolic Data Analysis.* Springer. pp. 359-368.

Bolstad, B., Irizarry, R., Astrand, M., Speed, T. (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19(2). pp 185-193.

Bolstad. B. (2004). *Low-level Analysis of High-density Oligonucleotide Array Data: Backgroud, Normalization, and Summarization.* University of California, Berkeley, Phd thesis.

Catmull, E., Rom, R. (1974). A Class of Local Interpolating Splines. In Barnhill, R., Reisenfeld, R., eds., *Computer Aided Geometric Design.* Academic Press, San Francisco. pp. 317-326.

Cope, L., Irizarry, R., Jaffee, H., Wu, Z., Speed, T. (2003). A Benchmark for Affymetrix GeneChip Expression Measures. *Bioinformatics*, 20(3). pp. 323-331.

Croux, C., Filzmoser, P. (1998). Robust factorization of a data matrix. In Payne, R., Green, P., eds., *COMPSTAT, Proceedings in Computational Statistics.* Physica-Verlag, Heidelberg. pp. 245-249.

Croux, C., Filzmoser, P., Pison, G., Rousseeuw, P. (2003). Fitting Mulitiplicative Models by Robust Alternating Regressions. *Statistics and Computing*, 13, pp. 23-36

Daigle, G., Rivest, L. (1992). A Robust Biplot. *Canadian Journal of Statistics*, 20 (3). pp. 241-255.

Eddy, W., Fitzgerald, M., Noll, D.C. (1996). Improved Image Registration by Using Fourier Interpolation. *Magnetic Resonance in Medicine*, 36-6. pp. 923-931.

Gabriel, K. (1978). Least Squares Approximation of Matrices by Additive and Multiplicative Models. *J. R. Statistical Society B*, 40 (2), pp. 186-196.

Gautier, L., Cope, L., Bolstad, B., Irizarry, R. (2004). affy - Analysis of Affymetrix GeneChip Data at the Probe Level. *Bioinformatics*, 203(3). pp. 307-315.

Golub, G., Van Loan, C. *Matrix Computations.* Johns Hopkins University Press. Baltimore, MD.

Hartemink, A., Gifford, D., Jaakkola, T., Young, R. (2001). Maximum Likelihood Estimation of Optimal Scaling Factors for Expression Array Normalization. In *SPIE BIOS 2001.*

Hill, A., Brown, E., Whitley, M., Tucker-Kellogg, G., Hunter, C., Slonim, D. (2001). Evaluation of Normalization Procedures for Oligonucleotide Array Data Based on Spiked cRNA Controls. *Genomebiology*, 2. pp. 1-13.

Hoffman, R., Seidl, T., Dugas, M. (2002). Profound Effect of Normalization on Detection of Differentially Expressed Genes of Oligonucleotide Microarray Data Analysis. *Genome Biol*, 3(7). RESEARCH0033

Holder, D., Raubertas, R., Pikounis, V., Svetnik, V., Soper, K. (2001) Statistical Anslysis of High Density Oligonucleotide Arrays: a SAFER Approach. In *Proceedings of the ASA Annual Meeting*, Atlanta, GA.

Hubbell, E., Liu, W., Mei, R. (2002). Robust Estimators for Expression Analysis. *Bioinformatics*, 18(12). pp. 1585-1592.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T. (2003). Exploration, Normalization, and Summarization of High Density Oligonucleotide Array Probe Level Data. *Biostat*, 4(2). pp. 249-264.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs. B., Speed, T. (2003). Summaries of Affymetrix GeneChip Probe Level Data. *Nucleic Acids Res*, 31(4):e15.

Kepler, T., Crosby, L., Morgan, K. (2002). Normalization and Analysis of DNA Microarray Data by Self-Consistency and Local Regression. *Genome Biol*, 3(7). RESEARCH0037.

Li, C., Wong, W. (2001a). Model Based Analysis of Oligonucleotide Arrays: Expression, Index Computation and Outlier Detection. *Proc Natl Acad Sci USA*, 98(1). pp. 31-36.

Li, C., Wong, W. (2001b). Model Based Analysis of Oligonucleotide Arrays: Model Validation, Design Issues, and Standard Error Application. *Genome Biol*, 2(8). RESEARCH0032.

Lipshutz, R., Fodor, S., Gingeras, T., Lockhart, D. (1999). High Density Synthetic Oligonucleotide Arrays. *Nat Genetics*, 21(1 Suppl). pp. 20-24.

Lockhart, D., Dong, M., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E. (1996). Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays. *Nature Biotechnology*, 14, pp. 1675-1680.

Naef, F., Lim, D., Patil, N., Magnasco, M. (2001). From Features to Expression: High Density Oligonucleotide Array Analysis Revisited. *Technical Report*.

Naef, F., Magnasco, M. (2003). Solving the Riddle of the Bright Mismatches: Labeling and Effective Binding in Oligonucleotide Arrays. *Physical Review E.*, 68. 011906.

Pison, G, Rousseeuw, P, Filzmoser, P., Croux, C. (2003). Robust Factor Analysis. *Journal of Multivariate Analysis*, 84. pp. 145-172.

Quackenbush, J. (2002). Microarray Data Normalization and Transformation. *Nat Genetics*, 32(Suppl). pp. 496-501.

Schadt, E., Li, C., Su, C., Wong, W. (2000). Analyzing High-Density Oligonucleotide Gene Expression Array Data. *J Cell Biochem*, 80(2). pp. 192-202.

Schadt, E., Li, C., Ellis, B., Wong, W. (2001). Feature Extraction and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data. *J Cell Biochem Suppl*, Suppl 37. pp. 120-125.

Sellers, K., Miecznikowski, J., Eddy, W. (2004). Systematic Variation in Genetic Microarray Data. *Biostatistics*, 1(1). pp. 1-47.

Sidorov, I., Hosack, D., Gee, D., Yang, J., Cam, M., Lempicki, R., Dimitrov, D. (2002). Oligonucleotide Microarray Data Distribution and Normalization. *Information Science*, 146(1-4). pp. 67-73.

Tukey, J. (1949). One Degree of Freedom Test for Non-Additivity. *Biometrics*, 5. pp. 232-242.

Wu, Z., Irizarry, R. (2004). Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. In *Proceedings of RECOMB 2004*.

Wu, Z., Irizarry, R., Gentleman, R., Murillo, F., Spencer, F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Technical Report, JHU, Dept. of Biostatostics.

Zhang, L., Miles, M., Aldape, K. (2003). A Model of Molecular Interactions on Short Oligonucleotide Arrays. *Nature Biotechnology*, 21. pp. 818-821.