A Model Based Background Adjustment for Oligonucleotide Expression Arrays

Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez Murillo, and Forrest Spencer *

July 1, 2003

Abstract

High density oligonucleotide expression arrays are widely used in many areas of biomedical research. Affymetrix GeneChip arrays are the most popular. In the Affymetrix system, a fair amount of further pre-processing and data reduction occurs following the image processing step. Statistical procedures developed by academic groups have been successful at improving the default algorithms provided by the Affymetrix system. In this paper we present a solution to one of the pre-processing steps, background adjustment, based on a formal statistical framework. Our solution greatly improves the performance of the technology in various practical applications.

Affymetrix GeneChip arrays use short oligonucleotides to probe for genes in an RNA sample. Typically each gene will be represented by 11-20 pairs of oligonucleotide probes. The first component of these pairs is referred to as a perfect match probe and is designed to hybridize only with transcripts from the intended gene (specific hybridization). However, hybridization by other sequences (non-specific hybridization) is unavoidable. Furthermore, hybridization strengths are measured by a scanner that introduces optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. One approach to adjusting is to pair each perfect match probe with a mismatch probe that is designed with the intention of measuring non-specific hybridization. The default

*Zhijin Wu is graduate student and Rafael A. Irizarry is Assistant Professor of Biostatistics (E-mail: rafa@jhu.edu), Francisco Martinez Murillo is manager of JHMI Microarray Core, Forrest Spencer is Associate Professor of Medicine and Molecular Biology and Genetics and director of JHMI microarray Core, Johns Hopkins University, Baltimore, Maryland 21205. Robert Gentleman is Associate Professor of Biostatistical Science, Dana–Farber Cancer Institute, 44 Binney Street Boston, MA 02115. The authors would like to thank Terry Speed, Ben Bolstad, and Earl Hubbell for insightful discussions that helped develop our ideas. We would also like to thank Wolfgang Huber and Tom Louis for some helpful comments. The work of Rafael Irizarry is partially funded by the Hopkins PGA (www.hopkins-genomics.org) Administrative/Bioinformatics Component (P01 HL 66583). The work of Zhijin Wu is partially funded by the Johnson and Johnson Research Foundation.
adjustment, provided as part of the Affymetrix system, is based on the difference between perfect match and mismatch probe intensities. We have found that this approach can be improved via the use of estimators derived from a statistical model that use probe sequence information. The model is based on simple hybridization theory from molecular biology and experiments specifically designed to help develop it.

A final step in the pre-processing of these arrays is to combine the 11-20 probe pair intensities, after background adjustment and normalization, for a given gene to define a measure of expression that represents the amount of the corresponding mRNA species. In this paper we illustrate the practical consequences of not adjusting appropriately for the presence of non-specific hybridization and provide a solution based on our background adjustment procedure. Software that computes our adjustment is available as part of the Bioconductor project (http://www.bioconductor.org).

1 Introduction

High density oligonucleotide expression arrays are now widely used to measure gene expression in many areas of biomedical research. Affymetrix GeneChip arrays are the most popular and are used by thousands of researchers worldwide. The number of publications in scientific journals based on data produced using this technology is proof of its success. To probe genes, oligonucleotides of length 25 base pairs are used. Typically, an mRNA molecule related to a gene is represented by a probe set composed of 11-20 probe pairs of these oligonucleotides. Each probe pair is composed of a perfect match (PM) probe, a section of the mRNA molecule of interest, and a mismatch (MM) probe, that is created by changing the middle (13th) base of the PM probe with the intention of measuring non-specific binding. See the Affymetrix Microarray Suite User Guide (1999) for details.

After RNA samples are prepared, labeled and hybridized with arrays, these are scanned and images are produced and processed to obtain an intensity value for each probe. These intensities represent the amount of hybridization for each oligonucleotide probe. However, part of the hybridization is non-specific and the intensities are affected by optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. In this paper, we refer to the part of the observed intensity due to optical noise and non-specific hybridization as background noise. The default background noise adjustment, provided as part of the Affymetrix system, is based on the difference PM − MM.

A final step in the pre-processing of these arrays is to combine the 11-20 probe pair intensities, after background adjustment and normalization, for a given gene to define a measure of expression that represent the amount of the corresponding mRNA species. Various alternatives algorithms motivated by statistical models have been proposed that outperform the default algorithm in many applications, see for example Li
and Wong (2001) and Irizarry et al. (2003a). Irizarry et al. (2003a) find that the $PM - MM$ transformation results in expression estimates with exaggerated variance. They propose a background adjustment step that ignores the $MM$ intensities. This approach sacrifices some accuracy for large gains in precision. The resulting algorithm was implemented in the Bioconductor project (http://www.bioconductor.org) affy package (Irizarry et al., 2003b) and has become a popular alternative to the default algorithm provided by Affymetrix. In this paper we demonstrate that the loss of accuracy mentioned above is due to the inappropriate adjustment for the presence of nonspecific hybridization.

Data from our own experiments, molecular biology theory, and publicly available data sets were used to develop a statistical model that describes background noise. An empirical Bayes procedure motivated by this model results in a background adjustment procedure that greatly improves existing approaches. In Section 2 we describe why background adjustment is important. In Section 3 we present the data we generated to motivate the model. In Section 4 we describe the statistical procedures that provide a solution to background adjustment. In Section 5 we show how our method provides an improvement to users of the Affymetrix GeneChip technology. In Section 6 we briefly describe the software we have developed to implement our methodology. Finally, in Section 7 we discuss our findings.

2 Motivation

The Affymetrix spike-in study (data available from: http://www.affymetrix.com/analysis/download_center2.affx) is a subset of the data used to develop and validate the MAS 5.0 expression measure algorithm, Affymetrix’s current default (See http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf). For this experiment, human cRNA fragments matching 16 probe-sets on one of the Affymetrix human chips, the HGU95A GeneChip, were added to a hybridization mixture at concentrations ranging from 0 to 1024 picoMolar in a design similar to a Latin square. Apart from the spiked-in probe-sets, the same RNA mixture was hybridized to 59 arrays. Because we know the spike-in concentrations, it is possible to identify statistical features of the data for which the expected outcome is known in advance. This experiment is described in detail by, for example, Irizarry et al. (2003c) and Cope et al. (2003).

Figure 1 shows a typical histogram of $PM$ intensities obtained from 14 normalized arrays from the Affymetrix spike-in experiment. For these arrays the spike-in concentrations, ranging from 0 to 1024, appear exactly 16 times and exactly once for each spike-in probeset. The arrays were normalized using quantile normalization (Bolstad et al., 2003) and then the geometric averages of $PM$ intensities for each
Figure 1: The solid lines is a density estimator of quantile normalized $PM$ intensities. We only show the (55,800) range in the x-axis for illustrative purposes. The $PM$ values are as high as 20,000. The dotted lines are from a log-normal distribution representing background noise. The arrows denote the geometric means of the $PM$ intensities of probes with the same nominal concentrations. The concentrations (in picoMolar) are shown at the top of the arrows.

A spike-in concentration was computed. These averages are shown with arrows in Figure 1 and plotted against their intended or nominal concentrations in Figure 2. The presence of background noise is clear from the fact that the minimum $PM$ intensity is not 0 and that the geometric mean of the probesets with no spike-in is around 200 units.

By using the log-scale transformation before analyzing microarray data, investigators have, implicitly or explicitly, assumed a multiplicative measurement error model (Dudoit et al., 2002; Newton et al., 2001; Kerr et al., 200; Wolfinger et al., 2001). The fact, seen in Figure 2, that observed intensity increase linearly with concentration in the original scale but not in the log-scale suggests that background noise is additive with non-zero mean. Durbin et al. (2002), Huber et al. (2002), Cui, Kerr, and Churchill (2003), and Irizarry et al. (2003a) have proposed additive-background-multiplicative-measurement-error models for intensities read from microarray scanners. Figures 1 and 2 support this view.
Figure 2: a) Geometric mean of observed concentrations plotted against nominal concentrations. The mean was computed for the concentration groups with 0, 1/4, 1/2, 1, 2, 4, and 8 picoMolar. The solid line represent a regression line fitted to the data. b) Same as Figure a) but the x and y axis are now in the log-scale.

2.1 Previous work

Affymetrix includes $MM$ probes in their arrays with the intention of measuring the nonspecific binding and optical noise components included in the $PM$ intensities. Their first attempt at an expression measure (MAS 4.0) used the transformation $PM - MM$ to adjust for non-specific binding and background noise. In general, $MM \geq PM$ for about 1/3 of the probes on any given array (Irizarry et al., 2003a) which results in negative adjusted intensity values. This results in two obvious problem: 1) we can not use the log transformation to account for the multiplicative measurement error and 2) expression measures based on an averages of the adjusted intensities are negative for about 5% of the probesets. In the most recent version of their software, MAS 5.0, Affymetrix offers a transformation similar to $PM - MM$ but “tweaked” for probes where $MM \geq PM$ to avoid adjusted values less than or equal to 0. A robust average of the log transformed adjusted $PM$s defines the MAS 5.0 expression measure. See http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf for details.

Using replicate array data, Irizarry et al. (2003c) showed that the across-replicate MAS 5.0 log expression measure standard deviation was an order of magnitude higher for probe-sets with low intensities than for probe-sets with large intensities. If we consider $PM$ and $MM$ to be random variables this is likely to be due to the variance introduced by subtracting $MM$s. Non-specific hybridization is likely to make $PM - MM$ a less biased estimate of a quantity proportional to the true amount of RNA than $PM$, but $PM - MM$ will
generally have larger variance than $PM$. The fact that the variance for probe-sets with low intensities is much larger than for probe-sets with large intensities is explained by the fact that for highly expressed genes we expect $PM - MM \approx PM$.

Irizarry et al. (2003a) describe the robust multi-array analysis (RMA) expression measure, an alternative to MAS 5.0. The background adjustment step in this procedure ignores the $MM$s. Instead of subtracting $MM$, an additive background plus specific signal with multiplicative error is proposed. They assume the specific signal follows an exponential distribution and that the background noise follows a normal distribution. After a data-driven ad-hoc procedure is used to estimate the exponential and normal parameters the conditional expectation of the specific signal given the observed intensity is used as a background adjustment. These assumptions result in a closed form transformation for the conditional expectation which provides a practical solution to background adjustment. Irizarry et al. (2003a, 2003c) demonstrate, using data from spike-in and dilution experiments, that RMA outperforms other popular expression measures, including MAS 5.0, in various practical tasks. However, as demonstrated in Section 5, by only performing a global background adjustment RMA does not adjust well for non-specific binding resulting in attenuated correlations between nominal and observed concentrations.

### 2.2 Practical Considerations

One of the most popular applications of microarray technology is the identification of genes that are differentially expressed. A common parameter of interest is fold-change. Notice in Figure 2b that, on average, the observed fold-change between probes spiked-in at 0.25 and 0.50 picoMolar is approximately 1 when it should be 2. Comparisons between the higher concentrations, 4 and 8 picoMolar, for example, have observed fold changes closer to what is expected. To see how this is related to background, say that true expression values in two samples being compared are $\mu_1$ and $\mu_2$ picoMolar. Ideally we should observe a fold change of $\mu_1 / \mu_2$. In practice, we observe intensities $PM_1 \approx k\mu_1 + B_1$ and $PM_2 \approx k\mu_2 + B_2$, with $B_1$ and $B_2$ representing background noise and $k$ a constant to account for the change in units. Because the $B$’s are strictly positive, the observed fold changes of the non-background-corrected raw expression values are biased (smaller than expected). This bias will be more pronounced for smaller values of $\mu_1$ and $\mu_2$ as observed in Figure 2b.

Receiver operator characteristic curves presented in Irizarry et al. (2003c) demonstrate that RMA outperforms MAS 5.0 when using large absolute value of observed log fold change to define differential expression. However, as pointed out by Irizarry et al. (2003c), fold-change estimates obtained with RMA
are less accurate (attenuated) than those obtained with MAS 5.0. Cope et al. (2003) compute the slope estimates from observed and nominal log fold change values to be 0.69 and 0.61 for MAS 5.0 and RMA respectively. For probesets with nominal concentrations of 4 picoMolar or less these slopes drop to 0.65 and 0.35 respectively. Reports from researchers that use RMA in practice related to this issue can be found here: https://www.stat.math.ethz.ch/pipermail/bioconductor/. In Section 2.3 we will argue that this because the global background correction performed by RMA does not take into account non-specific binding. In Section 5 we show some data that support this claim and demonstrate that our proposed background adjustment results in an expression as precise as RMA and as accurate as MAS 5.0.

2.3 G-C content

The physical system producing probe intensities is a complicated one. Theoretical predictions of non-specific bindings based solely on sequence are unlikely to be as successful as empirical ones. Affymetrix technology implements an empirical approach by including MM probes and using the observed intensities to adjust for background noise. However, as mentioned above, empirical evidence suggests that simply subtracting MM is a sub-optimal strategy. Due to the associated variance with the measured MM intensities we argue that one data point is not enough to obtain a useful adjustment. Physically adding other probes that can serve as MM to the arrays is not practical for logistic and financial reasons. In this paper we propose using probe sequence information to select other probes that can serve the same purpose as the MM pair. We do this by defining subsets of the existing MM probes with similar hybridization properties. We refer to these probes as pseudo-mismatches (pseudo-MM).

The specificity of a nucleic acid hybridization experiment depends on the stringency of the conditions for hybridization and for washing. Conditions such as low temperature, high salt, high nucleic acid concentration, will favor the formation of double stranded molecules. An optimal hybridization condition intended to identify only perfect match between two nucleic acid strands will specify these parameters, and will vary for different nucleic acid sequences (Watson et al., 1987).

In traditional hybridizations (such as Southern blots) performed at moderate or low stringency, non-specific hybridization background is often observed to be due to partial nucleic acid homology between two single strands with imperfect complementarity. This problem is pronounced when the G-C content is high because each G-C pair forms three hydrogen bonds whereas each A-T pair forms two. Microarray hybridizations typically query complex mixtures of labeled single stranded nucleotide in solution against a large number of immobilized single stranded nucleotide features on the array. Because of the number of
nucleotide sequences present in a single hybridization reaction, and because the stringency used cannot be
individually optimized for all array features, the opportunity for cross-hybridization is great (Cutler et al.,
2001).

We therefore propose, as a first attempt, to use subsets of probes with the same G-C content as a pop-
ulation of MM probes that can be considered pseudo-MM for all PM with the same G-C content. Figure
3 shows a box-plot of the observed log intensities for these populations of probes. This plot shows a clear
relationship between intensity and G-C content. Although the theory predicts $G$ and $C$ should behave sim-
ilarly we have found that $C$ has a bit of a smaller effect than $G$. Cutler et al. (2001) have also seen this
phenomenon. Although in principle we should subset by $G + C$ content, for the Affymetrix technology, we
have found that sub-setting by $G, C$ (ordered pair instead of sum) content yields better results.

Figure 3: Boxplots of log intensities from the array hybridized to Yeast DNA for strata of probes defined
by their G-C content. Probes with 6 or less G-C are grouped together. Probes with 20 or more are grouped
together as well. Smooth density plots are shown for the strata with G-C contents of 6, 10, 14, and 18.

We note that in our discussion we have focused the discussion on GC content of the probes. We are
aware that there are other potential causes of differences in intensity. These could be related to differential
labeling of the mRNAs (this phenomenon is well known for cDNA arrays). However, these sources are
typically based on the sequence and hence completely confounded with any possible GC content effect (at
least for the measurements we have available), thus it does not matter if we attribute the effect to GC content
or to some other cause, as we cannot tell the two apart. It would, perhaps, be useful to identify all potential
causes of the differences in intensity but that is well beyond the scope of the present paper.
3 Data

There have been attempts at using the sequence information to predict non-specific binding (Zhang et al., 2003). These are based on deterministic models. To build appropriate stochastic models to describe how mismatch probes measure non-specific binding we have carried out three experiments. First, we obtained RNA from human embryonic kidney derived cells to create a control sample. We also used genomic DNA from yeast to create a hybridization mixture with DNA molecules non-specific to transcripts synthesized on human arrays. The processing of the sample was done following Affymetrix specifications with specific variations depending on the sample content. Briefly, after extraction/synthesis of total RNA, 5 micrograms of total RNA were used to synthesize first strand cDNA using oligonucleotide probes with 24 oligo-dT plus T7 promoter as primer (Proligo LLC), and the SuperScript Choice System (Invitrogen). Following the double stranded cDNA synthesis, the product was purified by phenol-chloroform extraction, and biotinylated anti-sense cRNA was generated through in vitro transcription using the BioArray RNA High Yield Transcript Labelling kit. 15 µg of the biotinylated cRNA were fragmented at 94 C for 35 min, and 10µg of total fragmented cRNA were hybridized to the Affymetrix human genome U95Av2 GeneChip for 16hr at 45 C with constant rotation (60 rpm). Affymetrix Fluidics Station 400 was then used to wash and stain the arrays. Fluorescence was detected using the Hewlett-Packard G2500 GeneArray Scanner and image analysis of each GeneChip was done through the Micro Array Suite 5.0 software from Affymetrix, using the standard default settings. Three samples were prepared to study different aspects of background noise:

Unlabeled - In this sample RNA control from human embryonic kidney derived cells was not labeled, but the cocktail was created as described above, using 15 µg of amplified cRNA. Because the RNA was not labeled, the observed intensities for this hybridization will represent optical noise in the presence of biological sample.

No RNA - In this sample Hybridization cocktail was hybridized with no RNA. As for the previous sample the observed intensities for this hybridization will represent optical noise in the absence of biological sample.

Non-specific Yeast control RNA was hybridized to an array probing for human genes. This hybridization will represent the full component of the noise, non-specific binding and optical noise.

Figure 4a shows a quantile-quantile plot comparing the log intensities read from the array hybridized to the unlabeled sample to a normal distribution. The no RNA sample data looks almost identical. Notice that the distribution of optical noise is well approximated by a log normal distribution. Notice also that
log scale mean translates to intensity of about 32. Figure 3 shows the distributions of 4 G-C strata in the non-specific sample. These also seem to be well approximated by a log-normal distribution. As expected, the means of these distributions increase with G-C content and are larger than the intensities obtained from unlabeled sample. Figure 4b shows a log frequency versus log rank plot for the intensities read from an array hybridized with labeled human RNA. This figure shows that when specific binding occurs the distribution has a “fat-tail” and appears to follow a power law. In the next section we describe a model motivated in part by the data obtained from these experiments.

Figure 4: a) Normal quantile-quantile plot of observed log intensities form an array hybridized to the mixture with no label. b) Log-frequency versus log-rank plot for the intensities from a typical array hybridized to human RNA. The vertical dotted lines shows the estimated background noise log-scale mean and mean plus 3 standard deviations. The solid line is the result of a regression fitted to the left of the right-most dotted line.

4 Background Model and Adjustment

Any given probe will have some propensity to non-specific binding. As described in Section 2.3 and demonstrated in Figure 3, this tends to be directly related to its G-C content. The pseudo-MM will contain: 1) information about the probe’s propensity to non-specific binding and 2) as all probes on the array, information on optical noise. Because the probe-pair MM is constructed to be identical except for one base, it is likely to contain more information on non-specific binding than any pseudo-MM. However, this does not mean we should ignore the information available from the thousands of probes with the same G-C content. In
this section we propose a statistical model that describes the relationship between the \( PM, MM \), and probes of the same G-C content. Our mathematical formulation results in a background adjustment procedure that appropriately incorporates all the available information.

For any given probe intensity we assume \( PM = B_{PM} + S \), with \( B_{PM} \) a random variable representing background noise and \( S \) a quantity proportional to RNA expression (the quantity of interest). Let \( f_B \) represent the distribution of \( B_{PM} \). We can think of \( f_B \) as the distribution of background noise for strata of probes with similar hybridization properties, specifically with the same G-C content. For simplicity we will assume \( B_{PM} \) does not depend on \( S \). We consider \( f_B \) to describe two levels of variation: 1) probe specific variation and 2) measurement error. If we assume the \( MM \) is measuring only non-specific binding then we can write \( MM = B_{MM} \) with \( B_{MM} \) an observation from the \( f_B \) distribution. Because of the way the \( MM \) is constructed, changing just one base, \( B_{PM} \) and \( B_{MM} \) are considered correlated outcomes from the \( f_B \) distribution.

The data described in Section 3 support the assumption that \( f_B \) is log-normal. We therefore write the model formally in the following way:

\[
PM = B_{PM} + S \\
MM = B_{MM} \\
\log(B_{PM}) = B'_{PM} + \epsilon_{PM} \\
\log(B_{MM}) = B'_{MM} + \epsilon_{MM}
\]

where \( B_{PM} \) and \( B_{MM} \) don’t depend on \( S \). Here the \( B' \)'s are independent of the \( \epsilon \)'s and the \( \epsilon \)'s are independent from each other. As mentioned above, we assume \( B_{PM} \) and \( B_{MM} \) are correlated:

\[
(B'_{PM}, B'_{MM})^T \sim N \left( \beta \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \sigma_0^2 \begin{bmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{bmatrix} \right).
\]

Here the \( \epsilon \)'s can be thought to represent measurement error due to, for example, optical noise (the optical noise mean level is absorbed into \( \beta \)):

\[
(\epsilon_{PM}, \epsilon_{MM})^T \sim N \left( 0, \tau^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).
\]

We believe this model is an appropriate description of the natural process involved in measuring probe intensities. For estimation purposes it can be simplified to:

\[
(\log(B_{PM}), \log(B_{MM}))^T \sim N \left( \beta, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).
\]
The parameters $\beta$, $\rho = \rho_0 \sigma_0^2 / (\sigma_0^2 + \tau^2)$, and $\sigma^2 = \sigma_0^2 + \tau^2$ will depend on the G-C content of $PM$ and can be estimated from data. The large amount of data results in very precise estimate of these parameters. A background adjustment procedure can then be formalized as the statistical problem of predicting $S$ given that we observed $PM$ and $MM$ and assuming we know $b$, $\rho$, and $\sigma^2$.

### 4.1 Maximum Likelihood Estimate

The maximum likelihood estimate (MLE) of $S$ can be easily shown to be:

$$\hat{S} = \begin{cases} PM - \hat{B}, & PM > \hat{B} \\ 0, & \text{otherwise} \end{cases}$$

with $\hat{B} \equiv \exp\{\rho \log MM + (1 - \rho)\beta - (1 - \rho^2)\sigma^2\}$. To get a numeric value of $\hat{B}$ we can use plug-in estimators of $\rho$, $\beta$ and $\sigma^2$ described in Section 4.3.

The MLE provides an intuitive result. Instead of subtracting $MM$ as an adjustment, we subtract a shrunken $MM$ quantity. The $MM$ is shrunken toward the mean of the log-intensity distribution of the G-C strata, with the amount of shrinkage depending on the correlation of the background component of the $PM$ and $MM$ intensities. The shrinking is performed in the log scale which, in practice, helps protect against extremely large values of $MM$ resulting from the multiplicative error structure.

Although the MLE provides a transformation with desirable properties, it is not necessarily a practical solution if one is interested in a task such as detecting genes that are differentially expressed. For example, the fact that the log of the MLE is undefined whenever $PM \leq \hat{B}$ makes it unpractical to use the MLE background adjustment in conjunction with any summary method that uses log transformation. As mentioned, one of the most important application of microarray technology is estimating fold-change of expression measured for the same probeset on two arrays. In this case, instead of maximizing a marginal likelihood, a more appropriate estimate is obtained from minimizing the mean squared error:

$$E[\{\log(\bar{S}/S)\}^2 1_{\{S > 0\}} | PM, MM].$$

(2)

Not all genes are expressed in a typical cell, thus we expect $S = 0$ for some probesets. Therefore, if one is eventually going to consider fold-change, via log-ratios, as the quantity of practical interest, excluded cases for which $S = 0$ from the loss function becomes important.

There are many strategies that could be devised to define appropriate estimates based on minimizing (2). An empirical-Bayes-type approach is described in the next section.
4.2 Empirical Bayes Estimate

A practical and simple solution is provided by an empirical-Bayes-type approach. In this case we treat $S$ as a random variable which implies minimizing (2) is equivalent to minimizing

$$E[\{\log(S/S)\}^2 | S > 0, PM, MM].$$

Thus the solution is the posterior mean estimate: $\tilde{s} = E[s | S > 0, PM, MM]$, with $s = \log(S)$. Here, the random variable $S$ represents a quantity proportional to the number of transcripts, in the hybridization mixture, that are compliments to a randomly chosen $PM$. We will denote the number of transcripts with $N$. Blades et al. (2003) carefully studied the distribution of $N$ in various RNA samples, using serial analysis of gene expression (SAGE) technology (Velculescu et al., 1995). Figure 1 in Blades et al. (2003) shows that this distribution follows a power law. This is consistent with the heavy tails observed for $PM$ intensities. Because a scanner is used to measure our intensities, in our model $S$ represents a quantity proportional, not equivalent, to the number of transcripts $N$. Because we assume $S$ is proportional to $N$, $S$ should also follow a power law. The data in 4b contains components of both the background noise $B$ and the signal $S$. As we move right from the vertical line, representing the log-scale mean of $B$, the log frequency versus log rank plot falls on a line with slope near 1 as predicted by a power law, specifically by Zipf’s Law (Zipf, 1949).

Because $S$ is proportional to the number of transcripts we can think of it as discrete distribution. Due to the conditioning, to compute $\tilde{s}$ we need to know the smallest positive value of $S$, which we will refer to as $m$. However, because we do not know the transformation introduced by the scanner it is hard, if not impossible, to determine $m$. We therefore consider $m$ a tuning parameter in our approach.

For the purposes of obtaining a practical background adjustment transformation we will assume that, conditioned on $S \geq m$, $s$ follows an exponential distribution with rate $\lambda$. This implies $S = \exp(s)$ follows a power law. This assumption permits us to write out the empirical Bayes estimate as

$$\tilde{s} = \frac{\int_{-\infty}^{a} \phi\left( \frac{b-\beta^*}{\sigma} \right) \{PM - \exp(b)\}^{-\lambda} \log\{PM - \exp(b)\} \, db}{\int_{-\infty}^{a} \phi\left( \frac{b-\beta^*}{\sigma} \right) \{PM - \exp(b)\}^{-\lambda} \, db},$$

where $\phi(\cdot)$ is the standard normal density, $a = \log(PM - m)$, and $\beta^* = \log\{\rho MM + (1 - \rho)\beta\}$. This empirical Bayes estimate equation, although not as simple as the one for MLE, does provide an intuitive result: $\tilde{S}$ is a weighted average, over all possible background values $b$, of $\log\{PM - \exp(b)\}$. For most values of $PM$ and $\lambda$, the closer $b$ is to $\beta^*$ the larger the weight. We use the plug-in estimators of $\rho$, $\beta$, $\lambda$ and $\sigma^2$ described in Section 4.3. Because the scanner produces 16-bit images (pixels are integers between 0 and $2^{16} - 1$) the smallest positive pixel intensity value is 1. We therefore recommend using $m = 1$. However, through our
software the user can choose any $m > 0$. We have found that decreasing and increasing $m$ yields estimates with better accuracy and precision respectively.

### 4.3 Plug-in Estimates for Model Parameters

Affymetrix arrays typically have over 100,000 probe pair intensities. These can be used to get stable and precise estimates of the parameters $\rho$, $\beta$, $\lambda$ and $\sigma^2$.

For estimating $\beta$ and $\sigma^2$ we could simply use the log-scale observed mean and variance of the $\log(MM)$ values. However, as discussed in Section 7, in practice we observe that the $MM$ probes detect some specific signal. This results in overestimation of both $\beta$ and $\sigma^2$. Using the mode of an empirical estimate of the $\log(MM)$ distribution as an estimate of $\beta$ and data provides better results in practice. To obtain an estimate of $\sigma^2$ we use a statistic based on the data to the left of the estimated $\beta$. In Figure 1 we show, with dotted lines, the log-normal distribution obtained with these estimates.

Treating $\beta$ and $\sigma^2$ as known we can obtain stable MLE estimates of $\rho$ and $\lambda$ using, for example, the EM algorithm. However, in practice we have to perform the background adjustment calculations for various arrays. We sometimes analyze hundreds of arrays at once. To make our method useful to practitioners we need procedures that are less computer intensive. A quick estimate of $\lambda$ can be obtained by fitting a line to the larger ranks in a log frequency versus log rank plot, as done in Figure 4. We find that the estimate obtained from this procedure yields very similar results to the MLE estimate.

The correlation between $\log(B_{PM})$ and $\log(B_{MM})$ should not change from array to array. For this reason we obtain an estimate from the non-specific data described in Section 3 where there is no specific signal so $PM = B_{PM}$ and $MM = B_{MM}$. We also looked at the correlation between $\log(PM)$ and $\log(MM)$ values obtained from genes known or assumed not to have signal. For all these data the estimates where close to $\hat{\rho} = 0.8$. We use this plug-in estimator throughout the paper.

### 4.4 Simulation

To assess the performance of the four discussed background adjustment we performed a simulation. We generated data using model (1). We chose values for the parameters based on typical estimated values from the data. Specifically, we used $\beta = 4.6$, $\tau = .08$, $\rho_0 = 0.88$, $\sigma_0 = .25$. For ease of computation we selected log$_2$ expression levels $s = 0, 1, \ldots, 12$ instead of generating them at random. We simulated 320 pairs of $\{\log(B_{PM}), \log(B_{MM})\}$ and these same 320 pairs were used as the non-specific binding components in the 13 levels of $s$. By generating $250 \times 320 \times 13$ independent $\varepsilon_{PM}$s and $\varepsilon_{MM}$s, 250 simulated arrays were created.
with \( PM = B_{PM} \varepsilon_{PM} + 2 \) and \( MM = B_{MM} \varepsilon_{MM} \). We then background adjusted each simulated array using the MAS 5.0 adjustment, the RMA adjustment, and the MLE and empirical Bayes adjustments described here. Because, MAS 5.0 does background correction within probesets, we divided 320 probes at each expression level into 20 probesets of 16 probes. All other methods correct background at individual probe level.

Figure 5 shows assessments of accuracy and precision. To assess accuracy we show the average adjusted log intensity for each value of \( s \) plotted against the true \( s \). Notice that, as expected, the RMA adjustment is the most biased for small values of \( s \). MAS 5.0 and the MLE adjustment perform similarly. The empirical Bayes adjustment is less biased than the RMA adjustment but appears to over correct a bit in the \( 4 < s < 6 \) range. In terms of precision the RMA adjustment is the best, which is not surprising since it avoids the variance introduced by the \( MM \) intensities. The empirical Bayes adjustment has much better precision than MAS 5.0 and the MLE for \( s < 6 \). For \( s > 6 \) all procedures have roughly the same precision. The results presented in Figure 5 are consistent with what is observed with real data, as shown in in Section 5.

![Graphs showing accuracy and precision](image)

Figure 5: a) Average, over 250 simulated replicate arrays, adjusted probe log intensity plotted against “true” \( s \) values. The identity line is also shown. b) Standard deviation across 250 simulated replicate arrays of log-scale intensities plotted against “true” \( s \) values.

Note that the model predicts that variance dependence on \( s \) which is commonly seen in practice. As mentioned, fold change is typically used to decide which genes are differentially expressed. This simulation shows that genes with less expression will produced more variable estimates of fold changes estimates and that appropriate inferential approaches, based on model (1) could greatly improve the specificity and sensitivity of differential expression calls. This is discussed further in Section 7.
5 Results

In this section we assess our background adjustment procedure using bottom line results of special scientific interest obtained on real data. Both accuracy and precision are discussed. We use 28 arrays from the spike-in data described in Section 2. The 28 arrays were chosen so that array and probe-set effects were balanced with respect to concentration. We computed expression measures for these 28 arrays using MAS 5.0, RMA as described by Irizarry et al. (2003a), and RMA using the background adjustment procedures described in this paper. We use the acronym GC-RMA to denote the expression measures that uses the background adjustment described here. We will sometimes refer to the expression values obtained as observed RNA concentrations.

For Figure 6 the observed log (base 2) concentration for each concentration group were averaged and plotted against their respective nominal log (base 2) concentration. Ideally, if the nominal concentration doubles, so should the observed concentration, and we should see a line with slope 1. We can see in Figure 6 that both MLE and empirical Bayes versions of GC-RMA outperform MAS 5.0 which in turn outperforms RMA.

![Figure 6: a) Average observed log\(_2\) intensity plotted against nominal log\(_2\) concentration for each spiked-in probeset for all 28 arrays from the Affymetrix spike-in experiment. b) For each non-spiked-in probeset we calculate the mean log expression and the observed log-scale standard deviation across all 28 arrays. The resulting scatterplot is smoothed to generate a single curve representing mean standard deviation as a function of mean log expression.](image)

Figure 6 also shows smooth functions fitted to the scatter plot of the standard deviation (SD) versus the
Table 1: Slope and $R^2$ from the log observed concentration versus log nominal concentration regression are displayed in the first two columns. The median standard deviation (SD) across all non-spiked-in probesets obtained from the 28 arrays is in the third. The last two columns are the percentile and rank (out of 12626) of the slope value within the distribution of log fold changes obtained from all replicate array combinations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Slope</th>
<th>$R^2$</th>
<th>Median SD</th>
<th>Expected percentile</th>
<th>Expected rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAS 5.0</td>
<td>0.69</td>
<td>0.85</td>
<td>0.63</td>
<td>82.43</td>
<td>2188</td>
</tr>
<tr>
<td>RMA</td>
<td>0.61</td>
<td>0.80</td>
<td>0.11</td>
<td>99.97</td>
<td>4</td>
</tr>
<tr>
<td>GC-RMA (MLE)</td>
<td>0.71</td>
<td>0.80</td>
<td>0.14</td>
<td>98.01</td>
<td>248</td>
</tr>
<tr>
<td>GC-RMA (EB)</td>
<td>0.85</td>
<td>0.85</td>
<td>0.08</td>
<td>99.98</td>
<td>2</td>
</tr>
</tbody>
</table>

average of the observed log concentration for each gene across the 28 replicate arrays. These plots show that in terms of precision, RMA and the empirical Bayes GC-RMA outperform the MLE version which outperforms MAS 5.0. For genes with over-all low expression RMA and the empirical Bayes GC-RMA are considerably better than MAS 5.0.

The slope of these curves shown in Figure 6 can be interpreted as the expected value of log fold change for a probe-set that should be truly differentially expressed with a fold-change of 2. Table 1 shows fitted slope values obtained from the data shown in Figure 6 and the median across replicate SD obtained from all genes. These two numbers give an over-all idea of accuracy and precision. Table 1 also shows the expected percentile of a gene with a true-fold change of 2 when compared to genes with true-fold change of 1 (non-differentially expressed). Notice that the ideal is a percentile of 100%. To put this in the context of microarray applications we translate the percentiles to the rank among 12625 genes (the number of genes in the arrays used in this example). The distribution of fold-changes for non-differentially expressed genes was obtained empirically, i.e. using the 28 replicate arrays.

In Figure 6 we can see that the RMA curve appears to be flatter for lower nominal concentrations than for higher nominal concentrations. This is in agreement with the fact the the background adjustment used by RMA does not account for non-specific binding appropriately. This is corroborated by Figure 7 which shows the median expression values obtained for probesets stratified by their average (across probes) G-C content.

Table 2 shows the local slopes observed in Figure 6. As the slopes shown in Table 1, the local slopes represent the expected observed log fold-change for probesets with true fold-change of 2. However, unlike the Table 1 slopes, the local slopes are a function of the total nominal probeset concentration in the two
Figure 7: For each probe-set the average G-C content is computed and rounded off to nearest integer. Then for each G-C strata the median log expression level is computed for each expression measure. The median log expression is plotted against average G-C content.

Table 1 also shows the $R^2$ values obtained in the observed versus nominal concentration regression. Methods that correct for non-specific binding have higher $R^2$ values. This is the scale for local $R^2$ values (not shown). It is no surprise that the consistency of observed concentrations between probes with the same nominal concentration is greater if we adjust for non-specific binding. This will improve the validity of within array comparisons of different probesets.
Table 2: For concentration groups that differ by a multiple of 2 the local slope is computed. Notice that local slope is simply the difference between groups in average log observed concentration. The column headers define what two concentration groups are being compared.

<table>
<thead>
<tr>
<th>Method</th>
<th>0.50</th>
<th>0.75</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAS 5.0</td>
<td>0.90</td>
<td>0.54</td>
<td>0.67</td>
<td>0.78</td>
<td>0.82</td>
<td>0.80</td>
<td>0.75</td>
<td>0.65</td>
<td>0.57</td>
<td>0.31</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMA</td>
<td>0.29</td>
<td>0.30</td>
<td>0.52</td>
<td>0.62</td>
<td>0.71</td>
<td>0.73</td>
<td>0.78</td>
<td>0.74</td>
<td>0.65</td>
<td>0.55</td>
<td>0.40</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>GC-RMA (MLE)</td>
<td>0.74</td>
<td>0.58</td>
<td>0.74</td>
<td>0.79</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>0.75</td>
<td>0.66</td>
<td>0.56</td>
<td>0.39</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>GC-RMA (EB)</td>
<td>0.68</td>
<td>0.71</td>
<td>0.92</td>
<td>1.20</td>
<td>1.20</td>
<td>0.92</td>
<td>0.93</td>
<td>0.80</td>
<td>0.69</td>
<td>0.60</td>
<td>0.41</td>
<td>0.33</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: As Table 2 but the local slopes are converted to local ranks out of 12626.

<table>
<thead>
<tr>
<th>Method</th>
<th>0.50</th>
<th>0.75</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAS 5.0</td>
<td>1690</td>
<td>2696</td>
<td>2249</td>
<td>1968</td>
<td>1870</td>
<td>1906</td>
<td>1859</td>
<td>2021</td>
<td>2317</td>
<td>2595</td>
<td>3920</td>
<td>4070</td>
<td></td>
</tr>
<tr>
<td>RMA</td>
<td>311</td>
<td>279</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>64</td>
<td>287</td>
<td></td>
</tr>
<tr>
<td>GC-RMA (MLE)</td>
<td>229</td>
<td>366</td>
<td>227</td>
<td>200</td>
<td>182</td>
<td>171</td>
<td>172</td>
<td>224</td>
<td>292</td>
<td>396</td>
<td>733</td>
<td>1130</td>
<td></td>
</tr>
<tr>
<td>GC-RMA (EB)</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>11</td>
<td>58</td>
<td>143</td>
<td></td>
</tr>
</tbody>
</table>

6 Software

Clearly the technology being proposed here relies on efficient access to probe-level sequence data. Since there are many different types of chips the software should be modular with respect to chip type. It is also evident that there are many other potential uses for probe-level sequence data so a second guiding software design principle is to ensure that the data are readily available for other uses.

We have used a simple design. There is a software package (matchprobes) that performs two functions. First, given the chip manufacturers data it produces a new software package that contains the sequence data stored in a standard format. matchprobes also provides software tools for determining the base content of a sequence, for finding the complementary sequence and for various other quantities of interest. The probe sequence data comes in separate software packages (one per chip-type) which interface directly with matchprobes. Users can add functionality such as new string matching algorithms to matchprobes and thereby make those available to all users independently of the chip that they are using.

To produce a data package matchprobes takes the manufacturers data files indicating the sequences and installs them into an R data object that is appropriately indexed to allow for fast retrieval and querying. These secondary data packages are specialized to each chip. Commonly used data sets are available for
downloading from the Bioconductor web site. Users with custom chips can create their own data packages using `matchprobes`.

A package specifically for the computation of the empirical Bayes and MLE background procedures can also be found on the Bioconductor web page. This package together with the `affy` package (Irizarry et al., 2003b) can be used to compute expression measures using these background adjustment. All these components are available as part of the Bioconductor project (http://www.bioconductor.org).

7 Discussion

We have presented a statistical model for background noise in Affymetrix GeneChip arrays. Our model takes advantage of sequence information to appropriately describe non-specific binding variation. Estimation procedures motivated by the model result in practical adjustment that improve the over-all sensitivity and specificity of the current default methods. Figure 8 presents log fold change versus average log concentration (MA) plots demonstrating this.

The empirical Bayes approach performs best at the task of detecting differential expression. This method is designed for this purpose via the loss function being minimized (3). However, this criteria can be changed depending on the application. For example, using a procedure as that described in Shen and Louis (1998) or Huber et al. (2002) we could obtain variance stabilizing transformations.

Although the background adjustment procedures presented here outperform existing ones, there are three known problems with our model. We do not expect the improvements gained from solving these problems to be large. However, we plan to address all three in future work. The known problems are: 1) Irizarry et al. (2003a) demonstrate the the `MM` probes detect a bit of signal. This implies that a more correct assumption in model (1) is $MM = B_{MM} + \alpha S$ with $\alpha < 1$. 2) We expect the strength of specific binding to also depend on the probe sequence. This may explain why in Figure 7 we still see some dependence of the MAS 5.0 and GC-RMA expression measures values on $G-C$ content. 3) Because the dynamic range of the scanner is finite the higher intensities are attenuated. The fact that in Table 2 and Figure 6 the local slopes decrease as we get close to the highest concentration we believe is a consequence of this attenuation. We plan to run an experiment to study the attenuation function and incorporate this appropriately into our model.

In this paper we have proposed a model that described probe intensities. We have only used this model to perform background adjustment. To obtain expression measures from probe level data two more steps are involved: normalization and summarizing the data. In Section 4.4 we saw how the model can correctly predict the relationship between variance and expression values $s$. This suggests that inference on, for
Figure 8: The MA plot shows log fold change as a function of mean log expression level. A set of 14 arrays representing a single experiment from the Affymetrix spike-in data are used for this plot. A total of 13 sets of fold changes are generated by comparing the first array in the set to each of the others. Genes are symbolized by numbers representing the nominal log₂ fold change for the gene. Non-differentially expressed genes with observed fold changes larger than 2 are plotted in red. All other probesets are represented with black dots. The smooth lines are 3SDs away with SD depending on log expression.

example, differential expression can greatly improve by using uncertainty information derived from the model. In Figure 8 we include lines that are 3 standard errors, as a function of mean intensity, away from log fold change 0. If we require points to be outside these “confidence bands” to call the associate genes differentially expressed we greatly improve the specificity obtained when calling all genes with fold changes larger than 1 differentially expressed. Future work is to use our model to develop inferential statements about, for example, differential expression. Furthermore, we believe this model could be applied to other microarray technologies.
References


