Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays

[Extended Abstract]

Zhijin Wu
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
Baltimore, MD 21205, USA
zwu@jhsph.edu

Rafael A. Irizarry*
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
Baltimore, MD 21205, USA
rafa@jhu.edu

ABSTRACT

High density oligonucleotide expression arrays are a widely used tool for the measurement of gene expression on a large scale. Affymetrix GeneChip arrays appear to dominate this market. These arrays use short oligonucleotides to probe for genes in an RNA sample. Due to optical noise, non-specific hybridization, probe-specific effects, and measurement error, ad-hoc measures of expression, that summarize probe intensities, can lead to imprecise and inaccurate results. Various researchers have demonstrated that expression measures based on simple statistical models can provide great improvements over the ad-hoc procedure offered by Affymetrix. Recently, physical models based on molecular hybridization theory, have been proposed as useful tools for prediction of, for example, non-specific hybridization. These physical models show great potential in terms of improving existing expression measures. In this paper we suggest that the system producing the measured intensities is too complex to be fully described with these relatively simple physical models and we propose empirically motivated stochastic models that compliment the above mentioned molecular hybridization theory to provide a comprehensive description of the data. We discuss how the proposed model can be used to obtain improved measures of expression useful for the data analysts.

Categories and Subject Descriptors

J.3 [**Life and medical sciences**]: Biology and genetics

General Terms

Measurement

*To whom correspondence should be addressed.

RECOMB'04, March 27–31, 2004, San Diego, California, USA.

Keywords

Affymetrix probe-level data, background adjustment, microarrays, physical models, stochastic models

1. INTRODUCTION

In the Affymetrix system, a fair amount of further preprocessing and data reduction occur following the image processing step to obtain measures of gene expression. Background adjustments, normalization, and summarization of the probe level data are three typical steps. The model proposed in this paper is especially useful for the background adjustment step, thus we will focus our discussion on this aspect. However, in Section 6 we briefly discuss how it can be useful for normalization and summarization as well.

Affymetrix GeneChip arrays use short oligonucleotides (of length 25 bases) to probe for genes in an RNA sample. 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* (PM) probe and is designed to be specific to transcripts from the intended gene. However, non-specific hybridization and optical noise are unavoidable. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. Affymetrix's approach to adjusting is to pair each perfect match probe with a *mismatch* (MM) probe, that is designed by changing the middle (13th) base, with the intention of measuring only optical background noise and non-specific hybridization (NSB). The default adjustment, provided as part of the Affymetrix system, is based on the difference between perfect match and mismatch probe intensities ($PM - MM$).

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 represents the amount of the corresponding mRNA species. Affymetrix's default algorithm, MAS 5.0, is based on a robust average of $\log(PM - MM^*)$ values (MM^* denotes that some tweaking is performed to avoid logs of non-positives). Various researchers have developed alternative algorithms, motivated by statistical models, that outperform the default algorithm in many applications. For example, Li and Wong [16] notice a strong probe effect in both PM and $PM - MM$ and describe it via a simple multiplicative model. By analyzing various arrays

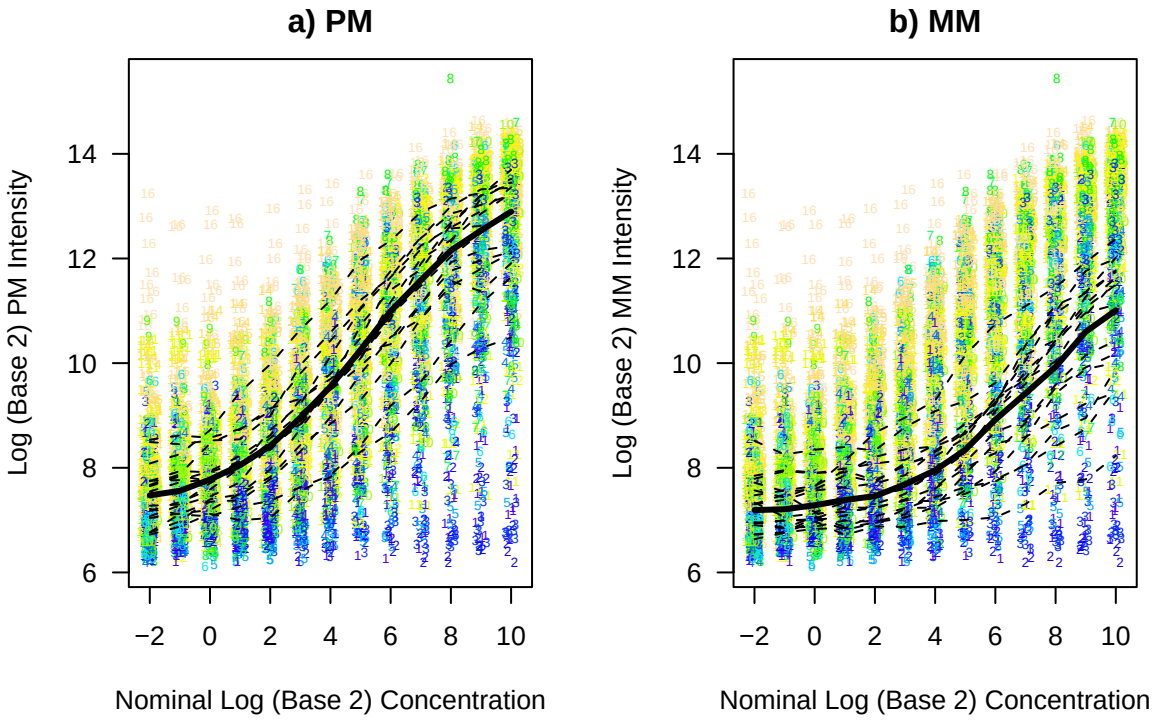

Figure 1: Signal detection by PM and MM probes in the Latin-square spike-in experiment. a). $\log_2(PM)$ **intensities of spike-in genes plotted against concentration. Number indicate the order of each probe within probesets, each number is associated with a color for probesets. The line shows median** $\log_2(PM)$ **for each concentration. The dashed lines is the same median but for each probe-set. b). Same as a) but for MM intensities.**

at once they are able to estimate probe effects and use this to improve outliers detection. Li and Wong also propose a non-linear normalization procedure that improves precision of the default re-scaling approach. Irizarry et al. [12] demonstrate that the $\log(PM - MM^*)$ transformation results in gene expression estimates with exaggerated variance. As a practical solution, they propose a global background adjustment step that ignores the MM intensities. This approach sacrifices some accuracy for large gains in precision. After the global background adjustment, arrays are quantile normalized [3] and a log-scale expression effect plus probe effect model is fitted robustly to define the robust multi-array analysis (RMA) expression measure. Irizarry et al. [14] and Cope et al. [5] demonstrate that RMA outperforms MAS 5.0 and the Li and Wong procedure in various practical tasks. RMA has been implemented in the Bioconductor project (`http://www.bioconductor.org`) *affy* package [13], Iobion's Genetraffic (`http://www.iobion.com`), and Insightful's S+ArrayAnalyzer (`http://www.insightful.com`) and has become a popular alternative to the default algorithm provided by Affymetrix. Various other similar algorithms have been proposed [10, 21, 17, 4, 24]. In Section 6 we will argue that the model described in this paper can be used to improve the accuracy of these methods, without much sacrifice in precision.

A simple version of our model can be written as $PM = O + N + S$ with PM the measured intensity of a particular PM probe, O representing optical background noise for this probe, N representing NSB and S represents observed specific signal. Similar models have been proposed by, for example, Hekstra et al. [9] and Zhang, Miles and Aldape [23]. A deterministic model that motivates Affymetrix's approach to background adjustment would be $MM = O + N$ which would imply that $PM - MM = S$. However, in Section 2 we demonstrate that a stochastic model is more appropriate. In this case, $PM = O^{(PM)} + N^{(PM)} + S$ and $MM = O^{(MM)} + N^{(MM)}$ where $O^{(PM)} + N^{(PM)}$ and $O^{(MM)} + N^{(MM)}$ have similar expectations but are not perfectly correlated. In this case the difference $\log(PM - MM^*)$ is approximately unbiased, since $\mathrm{E}[PM - MM] \approx S$, but may have a large variance $\mathrm{var}[\log(PM - MM^*)]$.

In Section 2 we demonstrate that the $O + N$ component of the PM and MM are not perfectly correlated, thus $\mathrm{var}[\log(PM)] << \mathrm{var}[\log(PM - MM^*)]$. In part this explains why PM-only measures, such as RMA, are more precise than measures based on $PM - MM$, such as MAS 5.0. Irizarry et al. [12] empirically show that for low intensity probes the variance of the difference $\log(PM - MM^*)$ can be considerably larger than that of $\log(PM)$. Furthermore, in general, $MM > PM$ for roughly 40% of all probes and this is problematic because we know S is strictly positive. These facts have led some researchers to consider $PM - only$ measures. However, because O and N are strictly positive, not correcting for optical noise and NSB can lead to biased

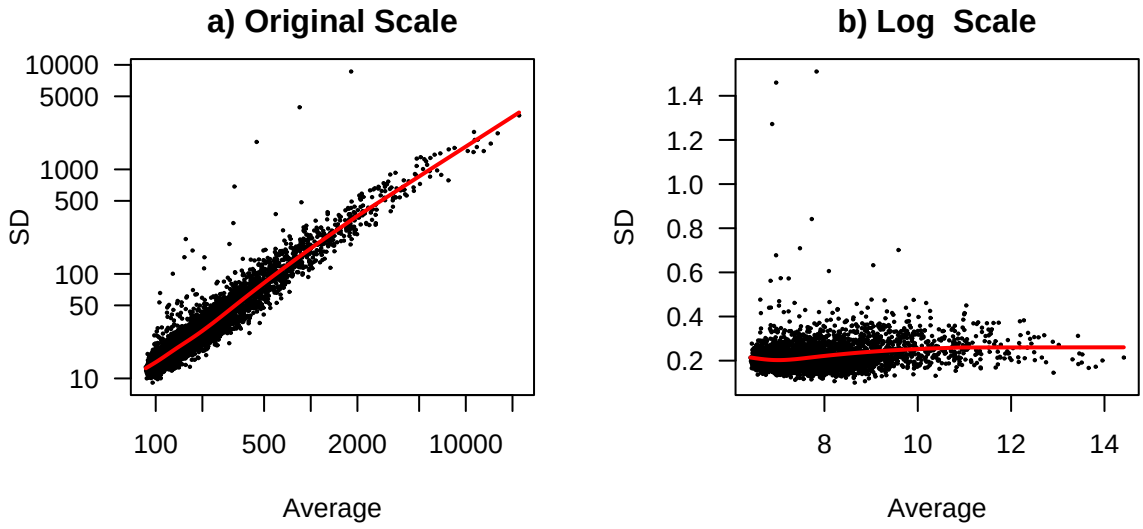

Figure 2: Standard deviation of probe intensity in original scale (a) and log scale (b) plotted a

results: $\mathrm{E}(PM) > S$. To see the negative effect this can have in a practical application of, say, estimating expression fold-change in two samples being compared, consider a simple example: Say that the true expression for a particular gene of interest in two samples being compared are μ_1 and μ_2 picoMolar. Ideally we should observe a fold change of μ_1/μ_2. In practice, we observe intensities $PM_1 = O_1^{(PM)} + N_1^{(PM)} + k\mu_1$ and $PM_2 = O_2^{(PM)} + N_2^{(PM)} + k\mu_2$ and an observed fold change $(O_1^{(PM)} + N_1^{(PM)} + k\mu_1)/(O_2^{(PM)} + N_2^{(PM)} + k\mu_2)$. Thus, as the $k\mu_1$ and $k\mu_2$ become smaller, as compared to the the strictly positive mean of the background components O and N, the estimated fold change converges to 1. This results in attenuated fold change estimates. RMA performs a global background adjustment that improves accuracy over non- background adjusted methods. However, as we will discuss later, different probes have different propensities to NSB which implies RMA does not fully account for NSB. In this paper we develop a model that predicts the behavior of optical noise, NSB, and specific binding very well. We use hybridization theory from molecular biology together along with data from carefully designed experiments to motivate the model. We also propose a model for the distribution of the specific signal S intensities within an array. This model can be used to improve existing expression measures and provides theoretical explanations for various facts observed in practice, for example: 1) $MM > PM$ for a considerable amount of probes, 2) $\log(PM - MM^*)$ has much larger variance than $\log(PM)$ when S is small, and 3) $\log(PM - MM^*)$ is more accurate than $\log(PM)$ when S is small.

2. EMPIRICALLY MOTIVATED STOCHASTIC MODELS

In this Section we use publicly available data and data from our own experiments to motivate some of the components of our stochastic models. The first of these data sets is the Affymetrix spike-in experiments. These experiments are described in detail, for example, by Irizarry et al. [12] and Cope et al. [5]. For this experiment, human cRNA fragments matching 16 probe-sets on one of the Affymetrix human chips 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. The second data come from what we call the *empty chip* experiment. For this experiment, sample RNA control from human embryonic kidney derived cells was not labeled, but hybridized following the Affymetrix protocol. Because the RNA was not labeled, the observed intensities for this hybridization will represent optical noise in the presence of biological sample. Finally, the third data come from what we call the *NSB* experiment. For this experiment, yeast control RNA was hybridized to an array probing for human genes. This hybridization will represent the full component of the noise, NSB and optical noise. These two experiments are described in more detail in Wu et al. [22].

2.1 Optical noise

Data from the empty chip experiment (not-shown) appear to follow a normal distribution with mean of roughly 30 and standard deviation (SD) of roughly 2. This motivates modeling the first component of our model, the optical noise component, as normally distributed.

By using a log-scale transformation before analyzing microarray data, a great number of investigators have, implicitly or explicitly, proposed a multiplicative measurement error model [7, 19, 15, 20] for microarray data. A slightly more complicated additive background multiplicative measurement error model has been proposed by, for example, [11, 6, 8]. In Figure 1a we see observed PM log (base 2) intensities from the spike-in data plotted against their nominal log (base 2) concentration. The solid line shows the median value for each concentration. Notice that this line looks very

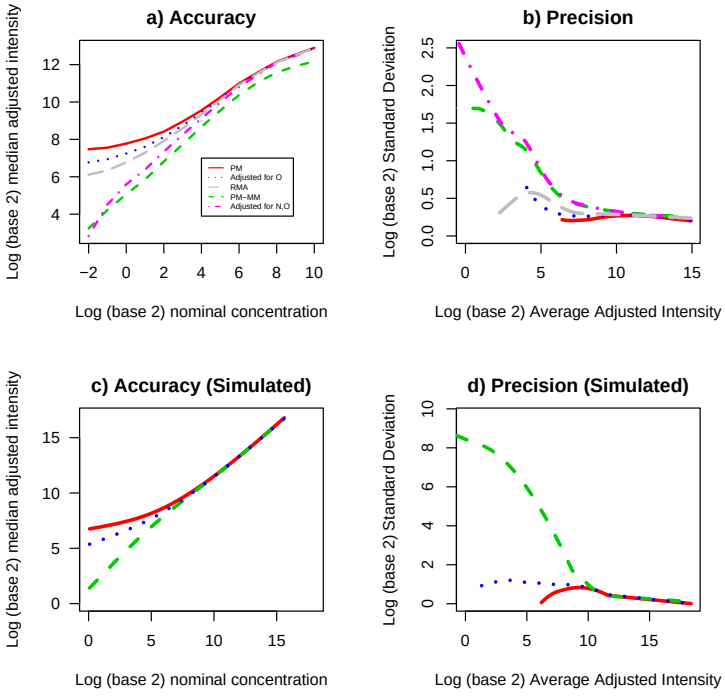

Figure 3: Global accuracy and precision of various background adjustments. a) Log median adjusted intensity plotted against log concentration. b) Standard deviation of log adjusted intensity plotted against log concentration. Adjusted intensities resulting in negative values are ignored. c) As a) but with simulated data. d) as b) but with simulated data.

much like the shape of the function $f(x) = \log_2(x+k)$ with k about 60. The Figure also shows this median value for each probe set with dashed lines. Although the curves are slightly different the general shape is about the same. This confirms that optical noise is additive as opposed to multiplicative.

Figure 2a shows SD of probe intensities, computed across 28 replicate arrays, plotted against the respective average intensity. Figure 2b shows the same plot for log intensities. The mean-variance dependence that is removed by applying a log transformation is a strong argument for a multiplicative error model. We therefore propose using an additive background multiplicative measurement error model.

Because the standard deviation of the optical noise is so small, as compared, for example, to the range of intensities, we will assume it is constant and correct for it by simply subtracting the minimum probe intensity (and adding 1 to avoid logs of 0). In Figure 3a we see the median intensities for each nominal concentration, as in Figure 1, for the PM and the PM adjusted for optical noise (along with other adjustments described later). We expect the curves in Figure 3a to be lines with slope 1, since every time the nominal concentration doubles observed concentration should double. We fit a line to the curve in this Figure and the slope for the PM intensities is 0.51. For the adjusted PM we have a slope of 0.59. The background adjustment slightly improves accuracy.

2.2 Non-specific binding

Molecular hybridization theory predicts that short oligonucleotides will hybridize to non-complementary transcripts. Our data from the NSB experiment support this. Figure 5a demonstrate a log-scale scatter plot of optical noise adjusted PMs versus optical noise adjusted MMs. This plot demonstrates intensities due to NSB are larger (by orders of magnitude) than those obtained just from optical noise. Because in this data there is no specific signal, if in fact the MM are an exact measure of the NSB captured by the PM then the predictive power of the MM should be 1 and this plot should have no scatter. However, as expected, we do see scatter. The relative predictive power or R^2 for this scatter plot is 0.71. Although not perfect, the large R^2 suggest that there is information on NSB to be extracted from the MM. Notice also that Figure 5a seems to suggest that after adjustment for optical noise the NSB component of the PM, MM pairs appear to follow a bivariate normal distribution.

To see that NSB is an additive effect more than it is a multiplicative effect, we adjusted the PM by subtracting and by dividing the MM. The resulting median intensity of $PM - MM$ is shown in Figure 3a. The estimated slope is 0.90 which is a good improvement over the non-adjusted PM. The PM/MM adjustment is very inaccurate (not shown in Figure 3a). The slope is only about 0.14. This suggest that NSB is an additive effect more than it is a multiplicative effect.

In Figure 3b we show a smooth curve demonstrating the over-all log-scale SD, across 28 replicate arrays, as a function of average log intensity, for the different adjustments. Notice that the $PM - MM$ adjustment is very noisy, especially at the low end. The loss of precision is quite significant; the median SD grows from 0.20 for non-adjusted PM, to 0.36

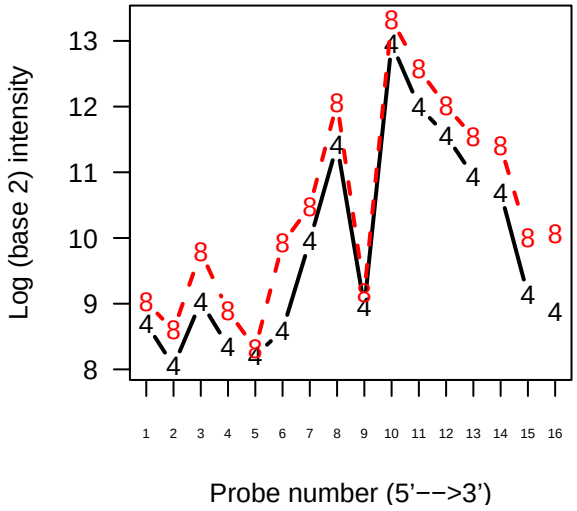

Figure 4: Probe log (base 2) intensities for the same probe set on two arrays spiked in at 4 and 8 pico-Molar respectively.

for optical noise adjusted PM, to 0.91 for $PM - MM$. The loss in accuracy of ignoring the MM is not as drastic.

2.3 Specific Signal

Li and Wong [16] demonstrate that, even after subtracting MM, there is a strong probe effect. Notice in Figure 1 that the range of probe intensities measuring the same nominal amount of RNA cover various orders of magnitude. In Figure 1 we use color and numbers to denote the same probes. The probes that have, on average bigger effects, are shown in yellowish colors, those with lower values in blue colors. The fact that the blue are always at the bottom, the yellow at the top demonstrate the strong and consistent probe effects. The fact that Figure 1 is a log-scale plot, suggests that their exists a multiplicative probe effect as well as the measurement error.

Figure 4 better illustrate the size of the probe effects compared to the differential expression effect sizes. This Figure shows probe intensity on the y-axis (logged) against probe number on the x-axis (ordered by proximity to the 5' end of the target transcript), for one probeset. The two lines represent the intensities read on two arrays where this particular probe-set was spiked-in at 4 and 8 picoMolar respectively. This plot clearly shows that probe effects are huge compared to array effects. Other probe-sets behave similarly (data not shown).

3. PHYSICAL MODELS

Zhang et al. [23] propose a stacking energy, positional-dependent-nearest-neighbor (PDNN) model for RNA/DNA duplex formed on microarrays. Their energy model takes into account the sequence of nearest-neighbors (adjacent two bases) and the position of these nucleotide pairs. It has been suggested that the effect of nearest-neighbor nucleotide pairs is the most important factor in determining RNA/DNA duplex stability. Zhang et al. add a positional weight factor to reflect the different contributions from different parts of the probe.

The energies for gene-specific binding (signal specific probe effect) and NSB of the j-th probe in i-th probe-set is thus calculated as,

$$E_{ij} = \sum_{k=1}^{24} \omega_k \epsilon(b_k, b_{k+1})$$
$$E^*_{ij} = \sum_{k=1}^{24} \omega^*_k \epsilon^*(b_k, b_{k+1}),$$

respectively, where ω_k, ω^*_k are weights, $\epsilon(b_k, b_{k+1}), \epsilon^*(b_k, b_{k+1})$ are nearest-neighbor stacking energies, and b_k is the base (A, T, G, or C) at position k. Zhang et al. [23] then proceed to describe the PM intensity of the j-th probe in i-th probe-set as

$$PM_{ij} = S_i/\{1 + \exp(E_{ij})\} + N^*/\{1 + \exp(E^*_{ij})\} + O,$$

where S_i is the number of expressed mRNA molecules of gene i, and $S_i/\{1 + \exp(E_{ij})\}$ is the contribution from gene specific binding. N^* is population of RNA molecules contributing to NSB for the entire array, and $N^*/\{1+\exp(E^*_{ij})\}$ is the contribution to intensity of j-th probe in i-th probe-set. The weights (ω_ks) are estimated empirically (see [23] for more details).

Naef and Magnasco [18] propose a simpler model to describe the probe effect, that considers only the sequence composition of the probes. Affinity of a probe is described as the sum of position-dependent base affinities:

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

where j is the base letter index, $k = 1, \ldots, 25$ indicates the position along the probe, b_k represents the base at position k as before, $1_{b_k=j}$ is an indicator function, and $\mu_{j,k}$ represents the effect of having base j in position k. Naef and Magnasco [18] make the model more parsimonious by assuming that the $\mu_{j,k}$ follow a polynomial of degree 3 as a function of position k. Their model is fitted to many arrays at once to obtain an affinity value for each sequence. We adapt this model to describe non-specific binding by fitting the model to our NSB experiment data and by modeling the $\mu_{j,k}$ as spline functions with 5 degrees of freedom. Notice that this model does not take into accout the interactions between nearest neighbors. Naef and Magnasco [18] demonstrate the these interactions for add much predictive power for specific signal probe effects. We find the same is true when predicting NSB. Notice also this model predicts that for certain probes the hybridization strength of the MM ill be stronger than the PM which implies $PM - MM$ is only an approximately unbiased estimate of S.

In Figure 5b and 5c we plot the optical noise adjusted $\log_2(PM)$s from the NSB data set against Naef and Magnasco's affinities and Zhang's PDNN $\log_2(N^*) - \log_2(1 + \exp(E^*_{ij}))$. Notice that Naef and Magnasco's affinities predict the NSB almost as well as the MM. The R^2 is 0.62. Zhang's PDNN also does relatively well with an R^2 of 0.28. However, notice that the slope of the PDNN model scatter plot is not 1.

Figure 5 demonstrates that these physical models can not predict NSB perfectly. However, they motivate a simple

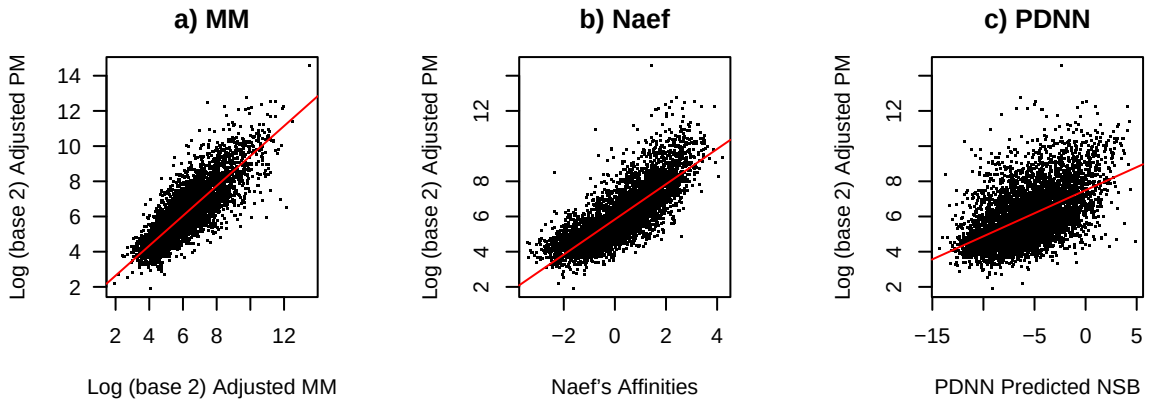

Figure 5: Intensity prediction ability comparison. a) Optical noise adjusted $\log_2(PM)$ intensity plotted against optical noise adjusted $\log_2(MM)$ intensity. b) Optical noise adjusted $\log_2(PM)$ against Naef's predicted affinity. c) Optical noise adjusted $\log_2(PM)$ against PDNN predicted non-specific binding.

stochastic model. In Section 4 we propose a model that describes the NSB contribution as log-normal distributed with log-scale mean proportional to Naef and Magnasco's affinities. The two parameters describing this relationship is estimated from the data and not predicted using physical models. The model works similarly when using Zhang's $-\log_2(1+\exp(E^*_{ij}))$ as the affinity measure.

3.1 Specific Signal

Irizarry et al. [14] postulate a log-scale additive model for the specific signal component of the probe level data:

$$\log(S_{ij}) = s_i + \alpha_j + \epsilon_{ij}, i = 1, \ldots, I, \text{ and } j = 1, \ldots . J$$

Here s_i represents log scale expression (the quantity of interest), α_j represents the probe effect, and ϵ_{ij} is the multiplicative measurement error. This model fits very well in practice and if one has enough arrays the probe effects α_j can be efficiently estimated [14]. As described above, Zhang et al. [23] and Naef and Magnasco [18] describe physical models useful for predicting α_j from the probe sequence. For example, we could add the further assumption that $alpha_j = \text{Affinity}_j\beta$. However, as described in section 6, it is also useful to describe the within array distribution of the s_i. This distribution will vary somewhat among RNA sources, and varies greatly among chip designs and other processing and hybridization factors. In this section we describe a parametric distribution useful for describing the empirical distribution of s_i.

Wu et al. [22] notice that for a wide variety of arrays (including arrays hybridized to samples of various type from humans, mice, and rats) the distribution of the specific signal for probe intensities can be well approximated with a power-law distribution. Observed intensities contain components of both the background noise and the signal. However, for higher intensities, where background has little effect, we observe that log frequency versus log rank plots fall on a line with slope near 1, as predicted by a power law, specifically by Zipf's Law [25]. This distribution has also been observed empirically in SAGE data [2].

Wu et al. [22] demonstrate that a log-exponential distribution (a special case of Zipf's law) appropriately predicts the signal. If this assumption holds then we can write:

$$\Pr(s_i < s) = 1 - \exp(s/\alpha), \alpha \approx 1.$$

Although we do not expect this assumption to hold true for all hybridization, we do find it useful in many instances.

4. UNIFIED PHYSICAL/STOCHASTIC MODEL

The described physical models perform relatively well at predicting NSB and, as will become apparent, the distribution of the specific signal. However, the predictions are not perfect and are complimented well with stochastic versions. The system producing intensities is very complicated and we argue that one can use physical models to approximate the process relatively well, but the lack-of-fit is best described with a stochastic model.

Our model for the PM intensity contains NSB and specific signal components that on the probe sequence composition as described by the physical models. The model can be written as

$$PM_j = O_j + N_j + S_j$$

where $O_j \sim \text{Normal}(b_o, \sigma_o^2)$, $\log(N_j) \sim \text{Normal}(b_N(A_j), \sigma_N^2)$, and $\log(S_j) = s + A_j\beta + \epsilon_j$. Here $b_N(A_j)$ is a smooth function of A_j, and $s \sim \text{Exponential}(1)$. We assume independence across probes. For the MM we assume the same model except for the lack of the S_j:

$$MM_j = O_j + N_j.$$

Furthermore, we assume that the PM and MM $\log(N_j)$ have a correlation of 0.7 (what we observe in the NSB experiment).

Notice that this model is defined by only few parameters and that we have over 200,000 probe intensities to fit them. One can use maximum likelihood estimation to do this. However, writing down the likelihood for this model is complicated as it involved a convolution of 3 densities.

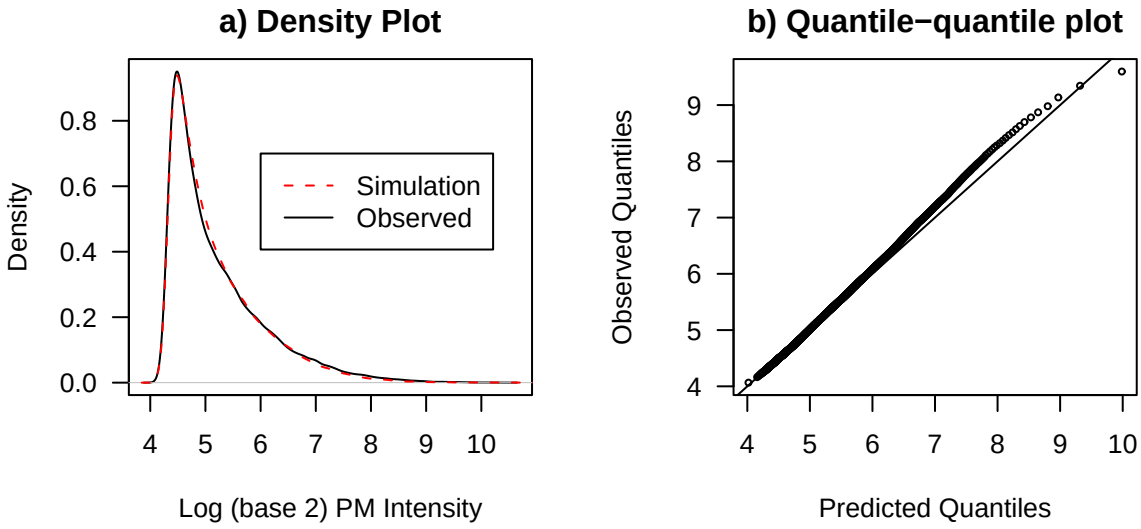

Figure 6: Simulation result. a). Kernel density estimates of simulated PM intensity and PM intensity from Latin-square spike-in experiment. b). Quantile quantile plot of PM intensity from Latin-square spike-in experiment against simulated PM intensity.

We have developed some ad-hoc procedures to estimate the parameters that yield very good fits.

To estimate the parameters for the background model, we notice that $\sigma_o^2 << \sigma_N^2$, thus we can assume that optical noise is approximately constant ($\sigma_o \approx 0$) which implies that b_o is approximately equal to the minimum observed intensity. Our estimate is thus $\hat{b}_o = \min\{\min_j PM_j, \min_j MM_j\} - 1$ (we subtract 1 to avoid nonpositive adjusted values). We then construct the intensities adjusted for optical noise: $PM'_j = PM_j - \hat{b}_o, MM'_j = MM_j - \hat{b}_o$ on an array. If we assume that the MMs do not measure specific signal, then $\Pr(PM_j < MM_j) \approx \Pr\{PM'_j < b_N(A_j)\}$. Because $b_N(A_j)$ is assumed to be a smooth function of A_j, we can estimate $\hat{b}_N(A_j)$ by considering a *neighborhood*, $\{PM'_k, k \in \Lambda_j\}$, with $k \in \Lambda_j$ if A_k is "close" to A_j, define the quantile

$$q = \Pr(PM_j < MM_j) \approx \frac{1}{\#\{\Lambda_j\}} \sum_{k \in \Lambda} 1_{PM_k < MM_k},$$

and then define the estimate $\hat{b}_N(A_j)$ as the q-th quantile of the $\log(PM'_k), k \in \Lambda_k$ probe intensity empirical distribution. Because we know that the $N(A_j)$ is an increasing function we force it to be monotonic using the pool-adjacent-violators (PAV) algorithm [1]. σ_N^2 can be estimated using the negative residuals $PM'_k - \hat{b}_N(A_j)$ (the positive residuals will contain part of the signal S_j). Finally, the β parameter can be estimated by simply regressing the PM'_j on A_j. We estimate the $N(A_j)$ for the MM_j similarly but because we assume the MM_j do not detect signal we use $q = 0.5$.

5. RESULTS

We fitted the model as described in the previous Section. The model fits extremely well. Figure 6a shows kernel density estimates of the PM intensities for one of the spike-in arrays along with the predicted distribution from the model. Notice that this model has less than 10 parameters (the smooth function fit uses about 4 degrees of freedom) and 200,000 data points so over-fitting is not a concern. Furthermore, the model is based on molecular hybridization theory. Figure 6b shows a quantile-quantile plot that confirms the good fit.

In Figure 3c and 3d we present the results shown in Figure 3a and 3b but instead of real data we use data simulated from our model. Notice the similarity between the real and simulated results. Our model predicts that, for low intensity probes, $\log(PM - MM^*)$ is a low precision transformation and that $\log(PM)$ is a low accuracy transformation. This suggests that our proposed model can be used for simulations related to statistical procedures based on Affymetrix data. For example, one could use it decide among different test statistics (Wilcoxon, t-test, SAM, etc...)

Finally, we point out that under this model, as fitted to this array, predicts the probability of a $MM > PM$ to be 0.40 which is exactly what we see empirically. Thus having many $MM > PM$ is not necessarily a bad thing. It is just a consequence of the noisy character of the system and the differenct in affinities for different sequences. Both these issues can be dealt with statistically.

6. DISCUSSION

We have presented a stochastic model motivated by molecular hybridization theory that fits Affymetrix GeneChip probe level data very well. Apart from giving a theoretical explanation for various facts observed in practice, this model can also be used to improve expression measures. For example, once we have fitted the model, we could correct for optical noise and NSB by computing the expectation of S given that we have observed a PM and MM. An approach such as this has been used by Wu et al. [22] with very encouraging results. Wu et al. describe an expression measure algorithm similar to RMA but using a model such as the one described here to adjust for background. Their expression measure is

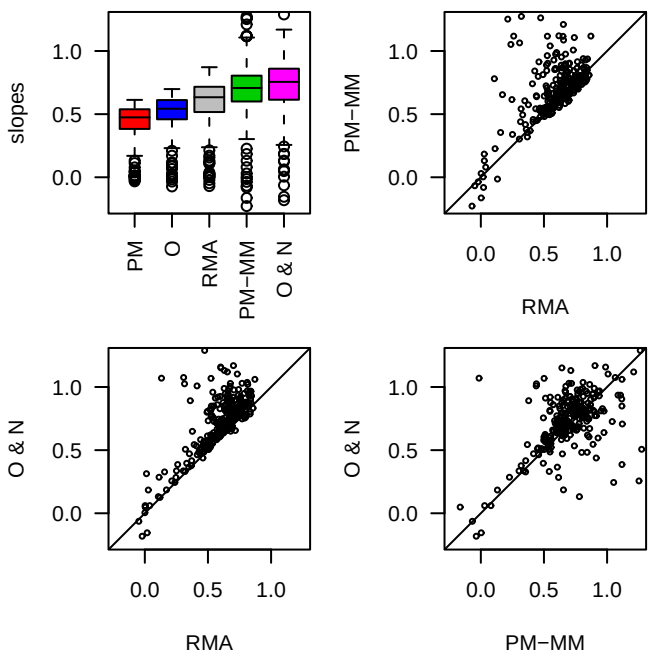

Figure 7: Boxplots show distribution of probe-specific fitted observed versus nominal log-scale slopes for different background adjustments. The scatterlots compare these slope for PM-MM, RMA, and the procedure described in this paper. Each point represents a probe.

about as precise as RMA but more accurate. In fact, it is more accurate than MAS 5.0.

Notice that this model can also be used for normalization and summarization. The fact that we have a prior distribution for the specific signal component suggest that one could use the log-exponential as the reference distribution used in quantile normalization. Furthermore, by incorporating information about probe-sets in the model (i.e. which probes represent which genes) one could directly obtain MLE estimates of expression measures from the model. One aspect that is not described by our model is the existence of outliers probes. This is subject of future work.

Finally, we point out that the described model motivates a PM-only expression measure that can be as accurate as those that use MM (such as MAS 5.0). In Figure 3a we show the global accuracy of the background adjustment defined by

$$PM_j - \hat{E}[O_j + N_j] \equiv PM_j - \left\{ \hat{b}_o - \exp\left(\hat{b}_N(A_j) + \frac{1}{2}\hat{\sigma}_N^2 \right) \right\}.$$

This adjustment has similar precision to $PM - MM$ but slightly better accuracy. Notice this adjustment does not depend on MM_j, except for that fact that we used to them to obtain the quantile q used to estimate $\hat{b}_N(A_j)$. Although there is complimentary information in the MM and in the affinities, PM-only measures are attractive for various reasons, for example: 1) We can have twice as many probes on the chips and 2) the MM seems to detect signal as demonstrated by Figure 1b. Using our model one could obtain roughly the same accuracy without the need for MM probes.

Figure 7 demonstrates that the gains in accuracy presented by the use of our model is not only a global result. The figure demonstrates that for almost all probes accuracy is improved by subtracting out the optical and NSB effects.

7. ACKNOWLEDGMENTS

We would like to thank Ben Bolstad, Felix Naef, Terry Speed, Li Zhang, and the reviewers for helpful comments and ideas that motivated and improved this extended abstract. We would also like to thank Forrest Spencer and Francisco Martinez-Murillo who helped us design and ran the experiments for us. Finally we would like to thank the R and Bioconductor developers for providing great software, specifically Robert Gentleman and Wolfgang Huber for developing the software to deal with sequence information.

8. REFERENCES

[1] R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67:140–147, 1972.

[2] N. Blades, J. Jones, S. Kern, and G. Parmigiani. Denoising of data from serial analysis of gene expression. *Bioinformatics*, page To appear., 2003.

[3] B. Bolstad, R. A. Irizarry, M. Astrand, , and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19(2), 2003.

[4] T. Chu, W. B., and R. Wolfinger. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math Biosci*, 176:35–51, 2002.

[5] L. Cope, R. Irizarry, H. Jaffee, Z. Wu, and T. . Speed. A benchmark for affymetrix genechip expression measures. *Bioinformatics*, page http://www.biostat.jhsph.edu/ ri-rizarr/papers/index.html, 2003. In press.

[6] X. Cui, M. K. Kerr, and G. A. Churchill. Data transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2:Article 4, 2003.

[7] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.

[8] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(Suppl. 1):S105–S110, 2002.

[9] D. Hekstra, A. R. Taussig, M. Magnasco, and F. Naef. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide array. *Nucleic Acids Research*, 31(7):1962–1968, 2003.

[10] D. Holder, R. F. Raubertas, B. V. Pikounis, V. Svetnik, and K. Soper. Statistical analysis of high density oligonucleotide arrays: a SAFER approach. In *Proceedings of the ASA Annual Meeting, Atlanta, GA 2001*, 2001.

[11] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 1:1:9, 2002.

[12] R. Irizarry, F. C. B. Hobbs, Y. Beaxer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[13] R. Irizarry, L. Gautier, and L. Cope. An R package for analyses of affymetrix oligonucleotide arrays. In G. Parmigiani, E. Garrett, R. Irizarry, and S. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*. Springer, Berlin, 2003.

[14] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31, 2003.

[15] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.

[16] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A*, 98:31–36, 2001.

[17] F. Naef, D. A. Lim, N. Patil, and M. O. Magnasco. From features to expression: High density oligonucleotide array analysis revisited. *Tech Report*, 1:1–9, 2001.

[18] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E*, 68:011906, 2003.

[19] M. Newton, C. Kendziorski, C. Richmond, and F. Blattner. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37:52, 2001.

[20] R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001.

[21] C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielsen, H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3, 2002.

[22] Z. Wu, R. A. Irizarry, R. Gentleman, F. M. Murillo, , and F. Spencer. A model based background adjustment for oligonucleotide expression arrays. Technical report, Johns Hopkins University, Dept. of Biostatistics Working Papers, 2003. http://www.bepress.com/jhubiostat/paper1.

[23] L. Zhang, M. F. Miles, and K. D. Aldape. A model of molecular interactions on short oligonucleotide microarrays: implications for probe design and data analysis. *Nature Biotechnology*, 21:818–821, 2003.

[24] L. Zhang, L. Wang, A. Ravindranathan, and M. Miles. A new algorithm for analysis of oligonucleotide arrays: application to expression profiling in mouse brain regions. *J Mol Bio*, 317:227–235, 2002.

[25] G. Zipf. *Human Behaviour and the principle of least effort*. Addison-Wesley, Cambridge, MA, 1949.