What does PLIER really do?

Terry M. Therneau
Karla V. Ballman

Technical Report #75
November 2005

Abstract

Motivation: Our goal was to understand why the PLIER algorithm performs so well given its derivation is based on a biologically implausible assumption.

Results: In spite of a non-intuitive assumption regarding the PM and MM errors made as part of the derivation for PLIER, the resulting probe level error function does capture the key characteristics of the ideal error function, assuming MM probes only measure non-specific binding and no signal.

Contact: ballman@mayo.edu

1 Introduction

The PLIER (Probe Logarithmic Intensity ERror) algorithm was developed by Affymetrix and released in 2004. It is part of several commercially available software packages that analyze Genechip® data such as Strand Genomic's Avadis and Stratagene's ArrayAssist®. The PLIER algorithm produces an improved gene expression value (a summary value for a probe set) for the GeneChip® microarray platform as compared to the Affymetrix MAS5 algorithm. It accomplishes this by incorporating experimental observations of feature behavior. Specifically, it uses a probe affinity parameter, which represents the strength of a signal produced at a specific concentration for a given probe. The probe affinities are calculated using data across arrays. The error model employed by PLIER assumes error is proportional to observed intensity, rather than to background-subtracted intensity. However, the derivation of the method also assumes that the error of the mismatch probe is the reciprocal of the error of the perfect match probe. We find this assumption counter-intuitive.

On the other hand, PLIER definitely performs well. It outperforms MAS5 in terms of the benchmark data and measures [2, 7] for assessing the quality of the summary statistic for a probe set. It also does fairly well compared to other methods that are commonly used to compute gene expression values for GeneChip probeset data. In particular, improvements over MAS5 include a higher reproducibility (lower coefficient of variation) without loss of accuracy and higher differential sensitivity for genes with lower expression values. This inconsistency, good performance of an algorithm derived from a counter-intuitive error model assumption, prompted us to look more closely at the PLIER algorithm. Specifically, we looked at the influence function for the algorithm and compared it to that for a more biologically based one. By examining the behavior of individual probes over a sequence of spiked-in RNA concentrations of a target gene, characterizations of the PLIER error function became clearer. The major finding is that the PLIER error model possesses many of the key characteristics of the ideal error function for fitting individual probe calibration curves.

2 PLIER Description

This description of the PLIER algorithm is based upon two presentations made by Earl Hubbell: one at the 2003 Affymetrix GeneChip Microarray Low-Level Workshop held at the University of California Berkeley campus in August 2003 (a link to the presentation can be found at `www.affymetrix.com/corporate/events/seminar/microarray_workshop.aff`) and the other at the Mathematical Biosciences Institute workshop held at The Ohio State University in October 2004 (`mbi.osu.edu/2004/workshops2004.html`).

Consider a single probeset on an array and a set of $j = 1, 2, \ldots, n$ arrays in the experi-

ment. We assume the probeset contains $i = 1, 2, \ldots, m$ probe pairs; a probe pair i consists of a perfect match (PM_{ij}) and mismatch probe (MM_{ij}). Let pm_{ij} and mm_{ij} represent the observed binding intensity for the perfect match and mismatch probe i on array j, respectively. The expected value for the observed binding for the perfect match and mismatch probes is assumed to be

$$\begin{aligned} E(pm_{ij}) &= \mu_{ij} = a_i c_j + B_{ij} \\ E(mm_{ij}) &= B_{ij} \end{aligned} \tag{1}$$

where

- B_{ij} is background binding for probe pair i on array j (background is assumed to be the same for the PM and MM probes within a pair),
- μ_{ij} is the binding level of probe i on array j,
- a_i is the binding affinity of probe i,
- c_j is the concentration of RNA in sample j, which is hybridized to array j.

The quantities B_{ij}, μ_{ij}, a_i, and c_j represent the (unknown) true values of the background binding, probe binding, affinity, and concentration, respectively, whereas pm_{ij} and mm_{ij} are the observed intensity values.

It is fairly well established from empirical data that the logarithm (log) of the observed binding intensities is approximately equivariant; in other words, the error is multiplicative. This implies the following model

$$\begin{aligned} pm_{ij} &= \mu_{ij} \epsilon_{ij}^P \\ mm_{ij} &= B_{ij} \epsilon_{ij}^M \end{aligned}$$

where ϵ^P and ϵ^M are random terms for the PM and MM probes, respectively, from an appropriate distribution, a log-normal for instance. Subtracting the observed MM probe binding intensity from its corresponding PM partner yields,

$$pm_{ij} - mm_{ij} = (a_i c_j + B_{ij})\epsilon_{ij}^P - B_{ij}\epsilon_{ij}^M.$$

The assumption that the perfect match and mismatch error for probe pair i are equal, i.e. $\epsilon_{ij}^P = \epsilon_{ij}^M = \epsilon_{ij}$, produces

$$pm_{ij} - mm_{ij} = (a_i c_j)\epsilon_{ij},$$

which is the original MAS5 equation. The issues and limitations associated with this error model, especially for low intensities (low binding), are well known [5, 6]. PLIER does not

assume that the perfect match and mismatch errors within a probe pair are equal, but rather assumes that $\epsilon_{ij}^P = 1/\epsilon_{ij}^M$. This seems counter-intuitive biologically; especially since the PM and MM probes within a given probe pair are physically adjacent to each other on the array. Any local artifact would be expected to affect both probes in the same direction rather than causing the error of one to increase when the error of the other decreases. Under the PLIER error assumption, equations (1) can be rearranged as,

$$\begin{aligned} \epsilon_{ij} &= \frac{a_i c_j + \sqrt{(a_i c_j)^2 + 4 pm_{ij} mm_{ij}}}{2 pm_{ij}} \\ &= \frac{\hat{\mu}_{ij}/pm_{ij} + \sqrt{(\hat{\mu}_{ij}/pm_{ij})^2 + 4(mm_{ij}/pm_{ij})}}{2} \end{aligned} \tag{2}$$

The PLIER algorithm selects a and c such that the average "residual" $r = \log(\epsilon)$ equals zero. Specifically, this is accomplished by minimizing a robust average of the r^2 values. The particular robust M-estimator used (Geman-McClure) is not of particular interest here. If the mismatch binding MM is zero, then $\log_2(pm_{ij}) = \log_2(\hat{\mu}_{ij}) + r_{ij}$, which shows that the estimate $\hat{\mu}$ is closely related to the geometric mean of the PM probes. The presence of MM binding increases the estimate for μ.

To more concretely understand how this algorithm works, consider a case of a single probeset on a single array. The goal is to obtain an estimate of the gene expression value for the probeset. For simplicity, assume there are only 3 probe pairs in the probeset. In this example, we use the first three probes of the U95A probeset 37777_at where the corresponding gene, protein tyrosine phosphatase receptor B (PTPRB), was spiked into a background of human pancreas RNA (at a concentration of 32 pM). The observed (pm, mm) intensity pairs were: (1801,627), (542, 132), and (229, 111). Figure 1 displays the r^2 curves for these probes as a function of the estimate for the true intensity ($\mu_{ij} = a_i c_j$), as well as the average error across all three probesets. Average probeset error is minimized by an estimate of 220 as the the true expression level of this gene. The argument is similar for the complete probeset of 16 probe pairs; the plot would just be more crowded.

3 The direct argument

3.1 Spike-in data

To better understand why PLIER does well, we begin by examining characteristics of the Affymetrix data. A spike-in experiment dataset was created by Affymetrix and is publicly available at their web site `www.affymetrix.com`; search on the phrase "Latin square data" to find the link to the page containing a description of the experiment and the downloadable files of data. In this experiment, mixtures of a common RNA background, in which 16 probesets were spiked in according to 14 different concentrations (0, 0.25, 0.5, 1, 2, 4, ...,

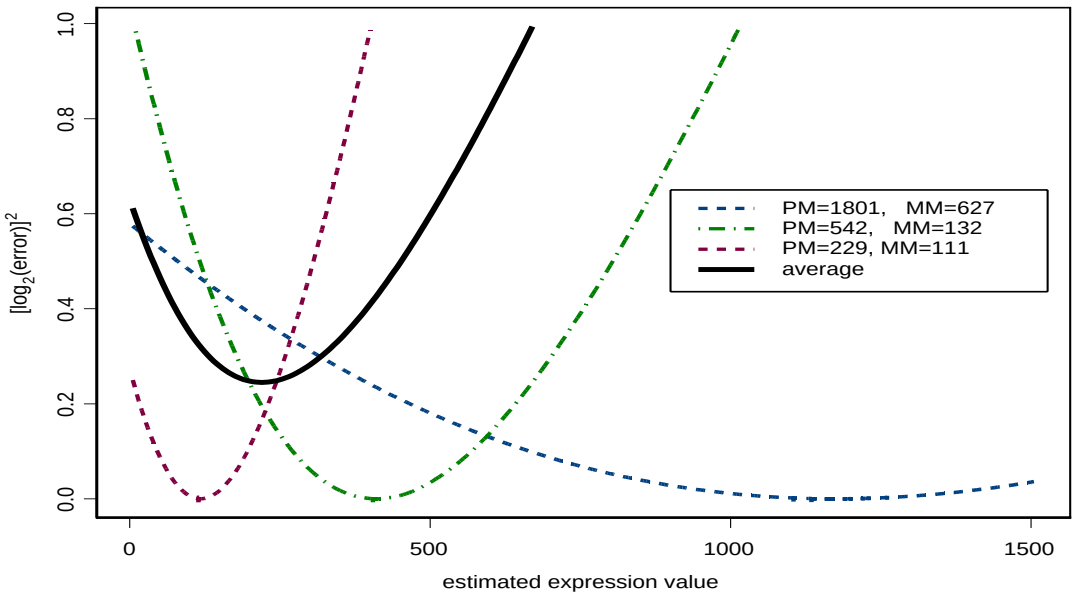

Figure 1: Error curves for various combinations

512, 1024 pM), were hybridized to a set of Affymetrix U95Av2 arrays. In most cases, each pattern of the 16 probeset concentrations was replicated three times. A cyclic latin square design was used for the spike-in pattern of the target RNAs. Irizarry et al. [6] provide a more detailed description of this experiment. Figure 2 contains the plots of the probes within the first spiked-in probeset 37777_at. Each plot contains the observations of the perfect match and mismatch probes. The probeset 37777_at was spiked in at 14 different concentrations (0, 0.5, 1, , . . . , 1024 pM) across a total of 59 arrays. The observed expression values were plotted on the y-axis and the spike-in concentrations were plotted on the x-axis; both on a $\log_2$ scale. A panel is shown for each probe in the probeset; the perfect match (PM) and mismatch (MM) values were plotted using different symbols. Fitted S-shaped curves were superimposed on the data, where the PM function differed from the MM function only in the location of its inflection point; a paper by Ballman and Therneau [1] contains the complete set of plots for this and other spike-in experiments. As can be seen in Figure 2, the S-shaped curves appear to fit the data well.

3.2 Models of the data

From the literature, there are at least two data models appropriate for the Affymetrix data. One model utilizes Langmuir isotherms. Its appropriateness for modeling Affymetrix data is described nicely by Hekstra et al. [4]. If the binding to the surface does not change

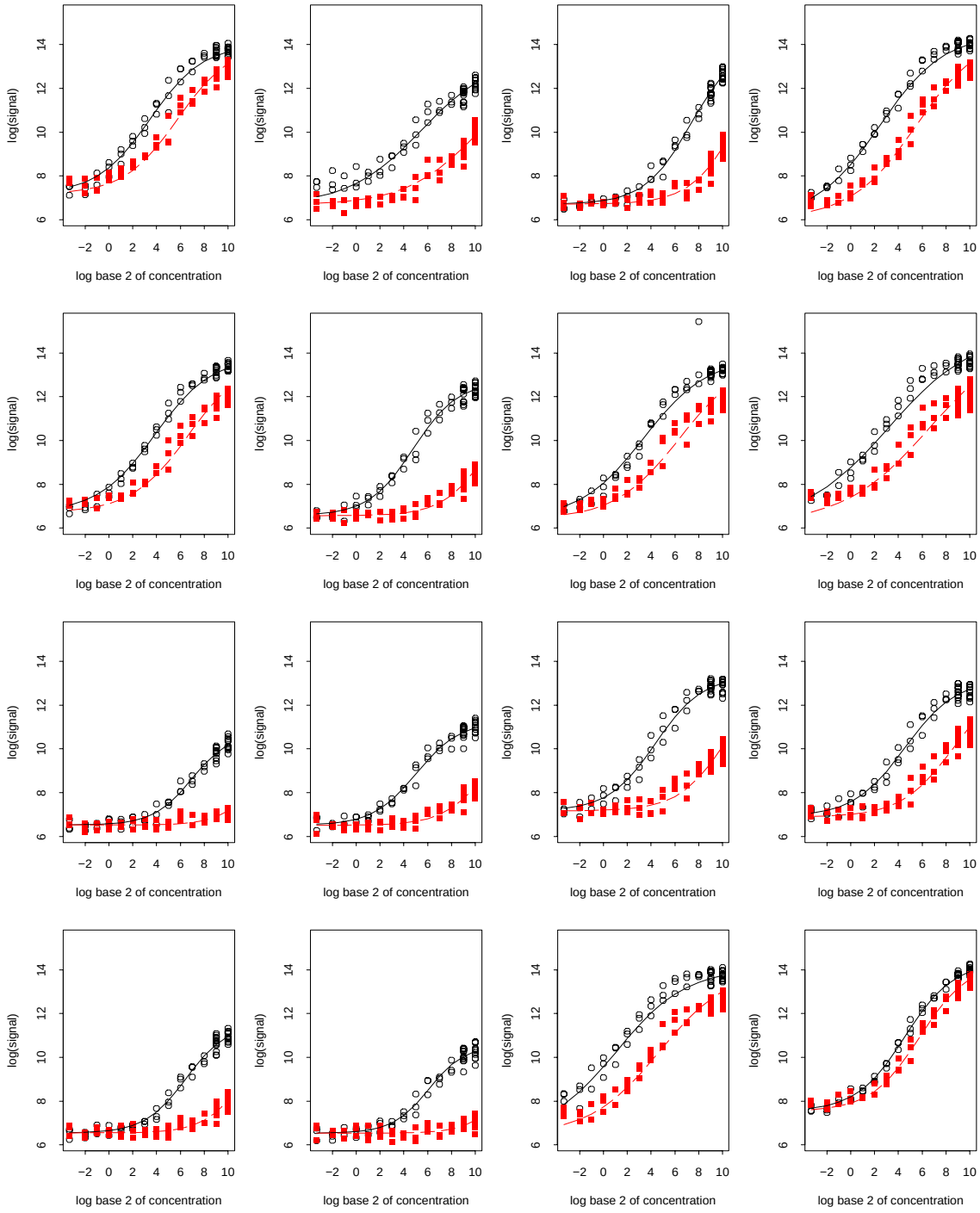

Figure 2: The probe pair intensity values versus the spiked concentration of a spiked gene, probeset 37777_at, from the Affymetrix U95A spike-in experiment. There are 16 probe pairs. The open circles are the values of the PM probes and the filled squares are the values of the MM probes. The solid and dashed lines correspond to the fitted logistic calibration curves for the PM probes and MM probes, respectively.

the concentration of the target in solution, then the Langmuir adsorption isotherm is an elementary model of surface adsorption. Let x be the specific target RNA concentration, then the fraction of occupied probe sites θ is given by ($\log_2$ scale)

$$\theta = \frac{2^x}{2^x + K}$$

where K is the concentration at which half the surface sites are occupied, if there were no non-specific hybridization. Assuming the measured fluorescence intensity to be linearly dependent on the amount of complementary RNA bound to the probe, we get the following model for the intensity y

$$y = b + d\theta = b + d\frac{2^x}{2^x + K} \tag{3}$$

where b and d have units of intensity. The model predicts chemical saturation at $b + d$ for high concentrations of RNA. It can also be shown that competitive cross-hybridization by non-specific RNAs in the target solution does not change the functional form of equation 3 but only affects the parameter values.

The second model was described by Finney [3] for behavior of calibration curves of radioligand assays where x is the log of the (known) dose and y the log of the observed intensity from the assay. Finney suggested using a logistic function for this type of data. A logistic function as a model for gene expression data also appears reasonable since the intensity values span a wide range. As seen in Figure 2, an S-shaped curve such as a logistic appears appropriate because it captures the effect of background binding and/or lower limits of detection (i.e. the flat lower portion of the lefthand part of the curve) and the effect of biochemical saturation and/or the instrumentation (i.e. the upper portion of the righthand part of the curve).

Is one of these models more appropriate than the other? Figure 3 shows a logistic curve and Langmuir isotherm curve, both scaled to the range of data values in Figure 2. Clearly, the two curves are virtually indistinguishable. In light of this, we fit logistic curves to the PM and MM data in Figure 2. The logistic curves were fit simultaneously where the PM curve only differed from the MM curves in the location of the inflection point. In other words, the PM and MM curves have identical shapes but the MM curve is shifted, usually to the right of the PM curve. As can be seen in Figure 2, the logistic function fits the data well.

3.3 Graphical comparison of error models

The ideal estimate of gene expression for an experiment would use the probe curves from Figure 2 directly; which is unfortunately not possible since the curves are unknown. But let us assume for a moment that the calibration curve is S-shaped, the corresponding ideal error function for the data will also be S-shaped. All other functions can be compared to this

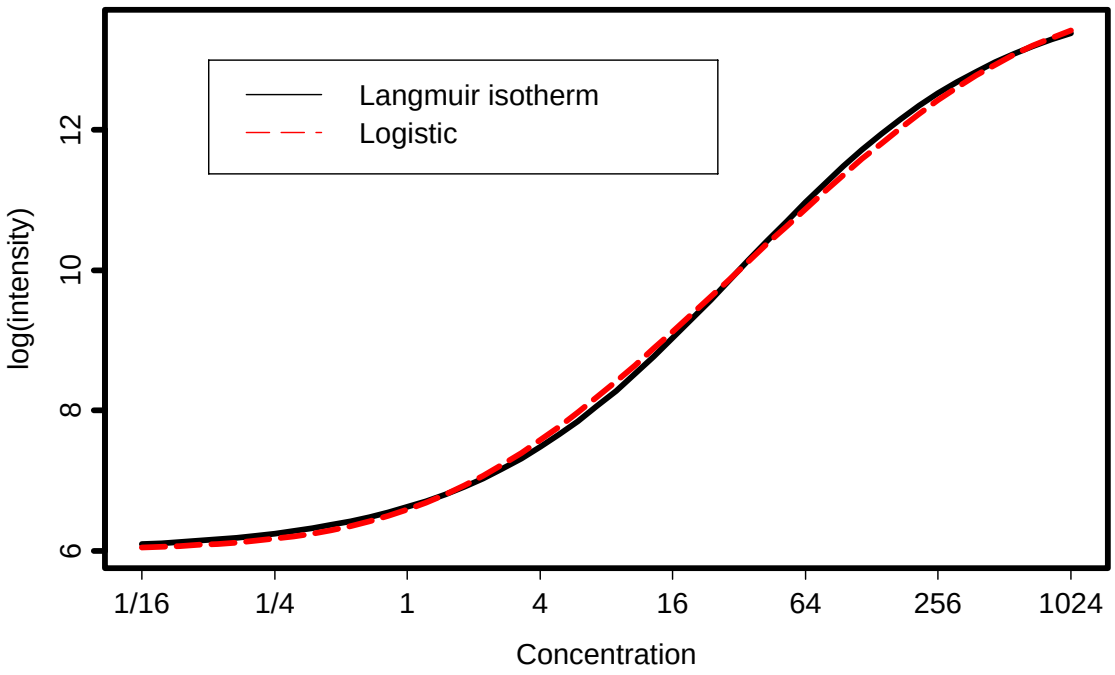

Figure 3: A logistic curve and Langmuir isotherm curve.

error function. We compare the error functions under for two different true concentrations for a probe, $\mu = 512$ and $\mu = 256$ (9 and 8 on the $\log_2$ scale, respectively), with a known binding background level of 64 (6 on the $\log_2$ scale). Specifically, we compare the error as a function of the estimated μ, i.e. $\hat{\mu}$, values under the MAS5 model and PLIER model to the ideal error curve (from a S-shaped calibration curve). The error functions for MAS5 and PLIER are a function of the observed pm and mm values. The PLIER (and MAS5) functions presented were applied to non-background adjusted data. However, PLIER (and MAS5) is applied to global background adjusted data in practice and so we show the error curve for PLIER applied to background-adjusted data. The global background adjustment we used was 64, which roughly corresponds to the 0.02 quantile of all the probe values (this is the default global background correction of MAS5). Figure 4 shows the form of the error functions on the same plot for different observed values of the pm and mm values. Note that these error functions are idealized in that they have been shifted so that they all have the smallest error (zero error if possible) at the true binding intensity value. The amount of shift necessary differs for the different functions and would be unknown in practice. Hence, this is a comparison of errors under perfect conditions for each function.

As can be seen from the panel of plots, the implied error function for MAS5 differs dramatically from the ideal error function. Foremost, the shape of the error function is concave rather than S-shaped. This explains the poor behavior of MAS5 for estimating expression values for low RNA concentration levels, which as been cited extensively in the

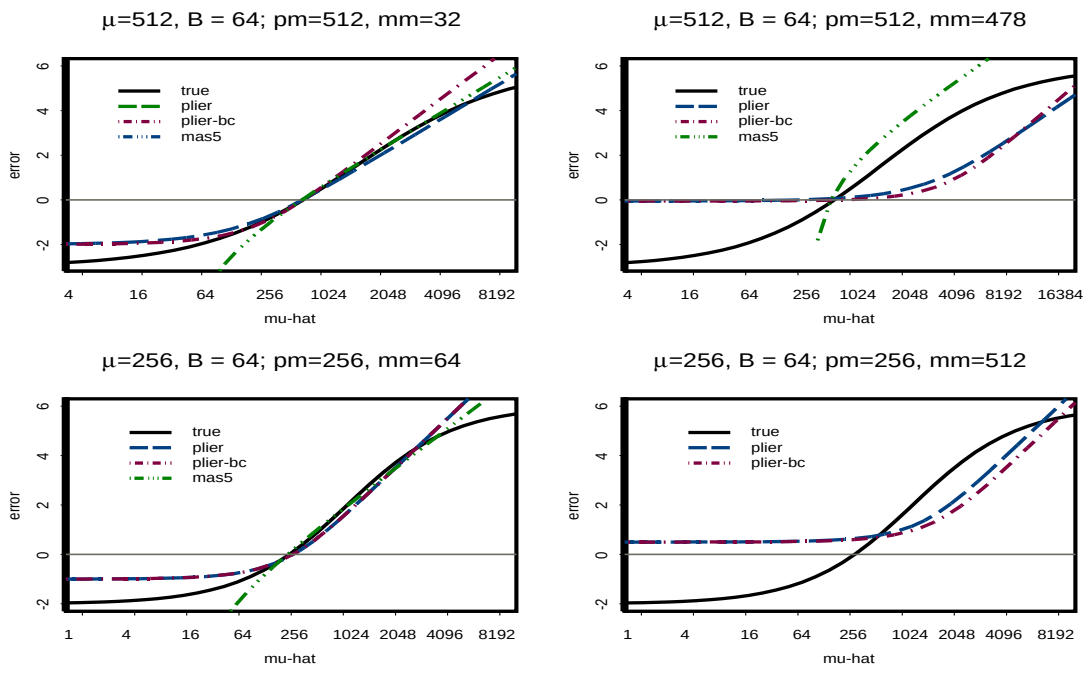

Figure 4: The functions for the ideal error (true), the MAS5 error (mas5) , PLIER error(plier), and PLIER applied to background-adjusted data error(plier-bc). The true gene intensity (μ) and background (B) and the observed pm and mm values are on each plot.

literature. Also note that there is no curve for the MAS5 model when $pm = 256$ and $mm = 512$. The reason is that when the mm value is larger than the pm value, the expression value is undefined. This is not technically true for MAS5 because in instances where $pm > mm$, the algorithm employs an ideal mismatch value; the ideal mismatch value is selected such that it is less than the pm value. However, when $pm > mm$, which occurs for the majority of the probe pairs, the error functions in 4 are correct.

On the other hand, the implied error curve for the PLIER model has the correct shape for the left portion of the function. This explains the observation that PLIER yields improved estimations of expression values for low RNA concentration levels compared to MAS5. Neither PLIER nor MAS5 error functions have the correct shape for the right-hand portions of the plots. In practice, the effect of differences from the ideal error for the right portion of the function is not as serious as differences in the left portion. This is because for actual experiments employing collected biospecimens of interest (cell lines, animal tissue, or human tissue), saturation of the probes is rarely reached. However, when the MM value is far above background, as it is for the spike in experiment when the observed MM values are greater than 256, the overly high lower threshold of the PLIER error function can cause overestimation. Finally, PLIER applied to global background adjusted data does not perform as well as PLIER applied to unadjusted data. There are several variations of

PLIER, e.g. PLIER+16 and PLIER+32, which add the constants 16 and 32, respectively, back to global background adjusted data. These constants are on the order of magnitude of the values that were subtracted for global background adjustment and largely "undo" the global background adjustment; from above, we see that PLIER performs better on data that has not been globally background adjusted.

4 Properties of the error function

From the graphical display of the error functions for MAS5 and PLIER, it appears as though a possible explanation for why PLIER performs so well is that in the crucial part of the error function, it has characteristics similar to the ideal error curve. What are the general characteristics of the ideal error function?

Assume that the true assay binding calibration function is a logistic curve, or something quite like it, so

$$\log(pm_{ij}) = f(\mu_{ij}) + \epsilon_{ij}$$

with ϵ from a symmetric distribution, which is equivariant across the range of the data. The true concentration μ_{ij} will be estimated with a model of interest such as array + probe effects. A rational approach for estimating the parameters is to minimize the overall error

$$\begin{aligned} E &= \sum_{ij} [\log(pm_{ij}) - f(\hat{\mu}_{ij})]^2 \\ &\equiv \sum_{ij} e(pm, \hat{\mu}, f) \end{aligned}$$

What is the form of this *error* function e?

Let us assume that f has a lower threshold or background, $\log_2(b)$, which corresponds to the scanner effect and non-specific binding when the target gene is not expressed. To the right of this threshold, assume f is linear or nearly linear on the $\log_2$ scale, and is smooth. Under these conditions, the error function would have the following properties.

1. For $\hat{\mu}_{ij}$ small, $e \to \log(pm_{ij}) - \log(b) = \log(pm_{ij}/b)$.
2. For $\hat{\mu}_{ij}$ large enough so that $f(\hat{\mu}_{ij})$ is in the linear part of the curve (i.e. sufficiently larger than $\log(b)$), the derivative of e with respect to $\hat{\mu}_{ij}$ will be a constant.
3. The behavior described in 2 is independent of the value of pm_{ij}.

For PLIER, we can verify 1 and 2 above, algebraically; this confirms the behavior observed in Figure 4 for the general case.

For property 1, the error in equation (2) is placed on the $\log_2$ scale and multiplied by -2 to get ϵ^*,

$$\epsilon^* = -2 \times \log_2(\epsilon)$$

$$= \quad -2 \times \log_2 \left[\frac{\mu_{ij}/pm_{ij} + \sqrt{(\mu_{ij}/pm_{ij})^2 + 4(mm_{ij}/pm_{ij})}}{2} \right]$$

As $\mu_{ij} \to 0$ (so $\hat{\mu}_{ij} \to 0$), we get

$$\begin{aligned} \epsilon^* &\to -2 \times \log_2 \left[\frac{0 + \sqrt{0 + 4(mm_{ij}/pm_{ij})}}{2} \right] \\ &\to \log_2 (pm_{ij}) - \log_2 (mm_{ij}) = \log_2(pm_{ij}/mm_{ij}) \end{aligned}$$

Under the Affymetrix assumption, the PM probe measures the target gene concentration and the MM probe measures the background level. Since pm estimates the signal level and mm estimates MM, or background, respectively, this satisfies the first property. Note, it has been established that MM does not measure background alone but also measures signal. However, as the true concentration level, μ, becomes small, MM becomes a better estimate of background, i.e. it is less likely to also measure signal. Hence, the PLIER error function is reasonably consistent with property 1.

For property 2, we again place the error on the $\log_2$ scale, drop the subscripts, and we get

$$\epsilon^* = \log_2 \left[\frac{\hat{\mu}/pm + \sqrt{(\hat{\mu}/pm)^2 + 4(mm/pm)}}{2} \right]$$

Now we take the derivative with respect to $\log_2(\hat{\mu})$

$$\frac{d\epsilon^*}{d\log_2(\hat{\mu})} = \left(\frac{1}{\ln 2}\right) \left(\frac{1}{\hat{\mu}/pm + \sqrt{(\hat{\mu}/pm)^2 + 4(mm/pm)}} \right) \left(\frac{1}{pm} + \frac{\hat{\mu}/pm^2}{\sqrt{(\hat{\mu}/pm)^2 + 4(mm/pm)}} \right) \left(\frac{\hat{\mu}}{\ln(2)} \right)$$

If we assume that background is small compared to signal (i.e. as we move away from background levels) and that mm is a good estimate of background, then $(mm/pm) \to 0$ as pm increases. Under these assumptions, as pm increases, we get

$$\begin{aligned} \frac{d\epsilon^*}{d\hat{\mu}} &= \left(\frac{1}{\ln 2}\right) \left(\frac{1}{\hat{\mu}/pm + \sqrt{(\hat{\mu}/pm)^2 + 4(0)}} \right) \left(\frac{1}{pm} + \frac{\hat{\mu}/pm^2}{\sqrt{(\hat{\mu}/pm)^2 + 4(0)}} \right) \left(\frac{\hat{\mu}}{\ln(2)} \right) \\ &= 1 \end{aligned}$$

So again, under somewhat reasonable assumptions, the PLIER error is consistent with the second property of the ideal error function. In addition, it is also consistent with property 3.

PLIER is of course making the assumption that log(MM) = background + error; in particular, it assumes that mm does not measure any gene signal. If this assumption is true, we see from above that the PLIER error model has the characteristics of the ideal error model, especially in the region of the plot that is the hardest, the low end. This explains why it does perform better than MAS5. However, the more these assumptions are violated—i.e. the more signal the observed mm measures in addition to non-specific

binding, the more the PLIER error function will deviate from the ideal error function. As mentioned previously, it is fairly well established in the literature that *mm* does measure signal in addition to non-specific binding, which may explain why PLIER is not the best performing algorithm of those entered in Affycomp.

5 Conclusions

In light of the fact that the MM probes are not good estimates for probe background level, the PLIER algorithm could likely be improved with a better estimate of background binding, perhaps along the lines of that proposed by Naef et al. [8]. Another question, one which we did not address here, is whether a robust average, such as that employed by PLIER, is really necessary. This is based on the fact that on a log scale, the spike-in data appear relatively equivariant, with few outliers. However, these considerations are of secondary importance. Of major concern is the fact that the error model is based upon an implausible assumption regarding the relationship between the error of the PM values and MM values.

Overall, we found that in spite of the non-intuitive assumption regarding the PM and MM errors made as part of the derivation for PLIER, the resulting model does capture the key characteristics of the ideal error curve, assuming MM probes only measure non-specific binding and no signal. Our only explanation for why this should be is good fortune.

This note has only considered the shape of the PLIER influence function for a single probe. When averaging over multiple probes not only the shape but the relative shifts of the per probe influence curves from one another will affect the effectiveness of the final estimate; our paper does not predict how PLIER will fair in comparison to other methods. In particular, we believe the deviations of the individual influence functions from the ideal error functions likely will be compounded when performing the averaging across the probes in a probeset. Our belief is based on the observation that although PLIER performs better than MAS5, it does not perform as well as other algorithms entered in Affycomp, most of which are based on more biologically plausible assumptions.

References

[1] K. V. Ballman and T. M. Therneau. A exploration of Affymetrix probe-set intensities in spike-in experiments. Technical Report 74, Mayo Clinic College of Medicine, March 2005.

[2] L. M. Cope, R. A. Irizarry, H. Jaffee, Z. Wu, and T. P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 1(1):1–13, 2003.

[3] D. J. Finney. Radioligand assay. *Biometrics*, 32:721–740, 1976.

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

[5] R. Irizarry, B. Hobbs, F. Collins, Y. Beazer-Barclay, K. Anntonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[6] 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:e15, 2003.

[7] R. A. Irizarry, Z. Wu, and H. A. Jaffee. Comparison of Affymetrix GeneChip expression measures. Technical Report Working Paper 86, Johns Hopkins University, Dept. of Biostatistics Working Papers, September 2005.

[8] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonecleotide arrays. *Physical Review E*, 68:011906–1–011906–4, 2003.