A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations

Felix Naef¹, Nicholas D. Socci² and Marcelo Magnasco¹

¹Center for Studies in Physics and Biology and ²Laboratory for Molecular Genetics, Rockefeller University, 1230 York Avenue, NY 10021, USA

Received on March 8, 2002; revised on July 11, 2002; accepted on August 20, 2002

ABSTRACT

Motivation: Despite the success and popularity of oligonucleotide arrays as a high-throughput technique for measuring mRNA expression levels, quantitative calibration studies have until now been limited. The main reason is that suitable data was not available. However, calibration data recently produced by Affymetrix now permits detailed studies of the intensity dependent sensitivity. Given a certain transcript concentration, it is of particular interest to know whether current analysis methods are capable of detecting differential expression ratios of 2 or higher.

Results: Using the calibration data, we demonstrate that while current techniques are capable of detecting changes in the low to mid concentration range, the situation is noticeably worse for high concentrations. In this regime, expression changes as large as 4 fold are severely biased, and changes of 2 are often undetectable. Such effects are mainly the consequence of the sequence specific binding properties of probes, and not the result of optical saturation in the fluorescence measurements. GeneChips are manufactured such that each transcript is probed by a set of sequences with a wide affinity range. We show that this property can be used to design a method capable of reducing the high intensity bias. The idea behind our methods is to transfer the weight of a measurement to a subset of probes with optimal linear response at a given concentration, which can be achieved using local embedding techniques.

Availability: Program source code will be sent electronically upon request.

Contact: felix@funes.rockefeller.edu; soccin@rockefeller.edu; marcelo@zahir.rockefeller.edu.

INTRODUCTION

High-density oligonucleotide arrays manufactured by Affymetrix are among the most sensitive and reliable microarray technology available (Chee et al., 1996; Lipshutz et al., 1999). Based on a photolithographic oligonucleotide deposition process, labeled and amplified mRNA transcripts are probed by 22–40 (depending on chip models) short DNA sequences each 25 bases long. The probes are preferentially picked near the 3’ end of the mRNA sequence, because of the limited efficiencies of reverse transcription enzymes. In addition, the probes come in two varieties: half are perfect matches (PM) identical to templates found in databases, and the other half are single mismatches (MM), carrying a single base substitution in the middle (13th) base of the sequence. MM probes were introduced to serve as controls for non-specific hybridization, and most analysis methods postulate that the actual signal (the target’s mRNA concentration) is proportional to the difference of match versus mismatch (PM−MM).

The purpose of this work is 2-fold. First, we present a detailed calibration study of GeneChips. Specifically, we apply latest analysis methods (MAS 5.0 algorithm and others) to a large yeast calibration data set, in which a number of transcripts are hybridized at known concentrations. We investigate the concentration dependence of both the accuracy and precision of differential expression scores. By accurate, we mean that the reported numerical ratio values are close to the known expression ratios, and therefore have little bias. On the other hand, a measurement is precise if it has a low noise level, also referred to as small variance.

Our results show that the ability of conventional analysis techniques to detect small changes strongly deteriorates toward high transcript concentrations. While the variance is smallest for high concentrations, it appears that the question of the bias in this regime has been neglected. In fact, the bias is strong enough that real changes of 2 (even 4) often can not be detected. This sounds at first counter-intuitive, which we believe is rooted in the following widespread interpretation of hybridization data. Namely, when examining the data from two replicated conditions (Figure 1a) most would focus on the low intensity region, and observe how noisy this regime appears to be in
compare to the high-intensity tail. However, this view
is misleading, as it does not consider the question of the
bias. Turning to a comparison of two different conditions
(Figure 1b), we notice that the noise envelope is essentially
unchanged, and that real changes appear as points lying
distinctly outside the noise cloud. Looking at multiple
such comparisons, we would then conclude that the high
intensity data is almost always very tightly scattered about
the diagonal, and that there are rarely genes in that region
that show fold changes greater than, say, 1.5 or 2. The
interpretation that no differential regulation occurs in
large concentrations. This saturation effect can actually
be observed in Affymetrix’s own data (Figure 7 at http://

Although the issue is not commented there (a qualitative
report has been given in Chudin et al. 2001). The physical
origin for the compression effect invokes non-linear probe
affinities and chemical saturation. This is a separate
issue from optical saturation (cf. Results). Chemical saturation
occurs below the detector threshold and is attributed
to the fact that some probes will exhaust their binding
capacities at relatively low concentrations, simply because
their binding affinities are high. Binding affinities are in
fact very sensitive to the sequence composition, resulting
in measured brightnesses that usually vary by several
decades within a given probe set (Naef et al., 2002b).

Our second goal is to present an analysis method that
reduces the bias at high concentrations. Our approach uses
all PM and MM probes equally, in contrast to the standard
view in which the PMs are thought to carry the signal,
while the MMs serve as non-specific controls. In fact, it
has become clear that the MM probes also track the signal,
usually with lower (although often with higher) affinities
than the PMs (Naef et al., 2002b). In that sense, the MMs
should be viewed as a set of on average lower affinity
probes. It is then reasonable to expect that some MM
probes will more accurate at high intensities, since they
will be less affected by saturation than the the PMs (cf.
Figure 3).

METHODS

The existing methods for the analysis of the raw data
fall into two main classes. The first methods are similar to
Affymetrix’s Microarray Suite software, providing
absolute intensities on a chip by chip basis, or differential
expression ratios from two experiments (Affymetrix,
2001; Naef et al., 2001, 2002a). The second class are
called ‘model-based’ approaches (Li and Wong, 2001), and
attempt to fit the probe affinities from a large number of
experiments.

The method described below belongs to the second class
and is specifically designed for improved accuracy in the
compressive high-intensity regime. It is based on ideas
borrowed from the theory of locally linear embeddings
(Roweis and Saul, 2000).

Notation

We construct the following matrix

$$A_j^i = \begin{cases} PM_i^j & \text{if } 1 \leq j \leq N_p \\
MM_i^{j-N_p} & \text{if } N_p < j \leq 2N_p \end{cases}$$

or in expanded notation

$$A_j^i = \begin{bmatrix}
PM_1^1 & \cdots & PM_1^{N_p} & MM_1^1 & \cdots & MM_1^{N_p} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
PM_{N_p}^1 & \cdots & PM_{N_p}^{N_p} & MM_{N_p}^1 & \cdots & MM_{N_p}^{N_p}
\end{bmatrix}$$

which contains the raw, background subtracted and
normalized data. By background, we mean fluores-
cence background, which we identify by fitting a
Gaussian distribution to the subset of all (PM, MM)
pairs satisfying the criterion that $|PM - MM| < \epsilon$
with $\epsilon = 50$. This provides us with the mean
and SD in the background fluorescence. (details at
http://xxx.lanl.gov/abs/physics/0102010). For a fair com-
parison of compression effects in various methods, we
used the global normalization factors from the MAS 5.0
software in all cases, however, the technique remains
applicable with other normalization schemes. $N_p$ is
the number of probe pairs and $N_e$ is the number of
experiments. We introduce a set of weights $w_i$ such that

$$\sum_{i=1}^{N_e} w_i = 1$$

and define the column means (or center of mass)

$$m^j = \sum_{i=1}^{N_e} w_i \log(A_j^i)$$
Note, we are computing the mean of the logs of the components of $A_j$.

**Local principal component analysis**

Local embeddings are adequate in situations where compression is important because non-linearities (resulting from chemical saturation) affect the one-dimensional manifold $\{PM_j(c), MM_j(c)\}$ (the concentration $c$ is the one-dimensional ‘curve’ parameter) by giving it a non-zero local curvature. The results section, in particular Figure 3 contains ample evidence that these non-linearities are significant (cf. also Figure 2 in Chudin et al. 2001). Our method is a multidimensional generalization of the schematic depicted in Figure 2, which shows the typical situation of two probes in which one of the probes ($PM_2$) saturates at concentrations lower than the other. If both probes were perfectly linear, the curve would be a straight line with slope 1. In the multidimensional case, the directions of largest variation (analogous to D1 or D2 in Figure 2) are computed from the principal components of the matrix

$$\sqrt{w_j} \left( \log(A_j) - m_j \right) = \sum_{k=1}^{N_p} U_{ik} D_k V_j^k$$

which can easily be done via singular value decomposition (SVD). In order to reconstruct the concentrations, one needs first to consider the unspecified sign of the vector $V_j^1$ (when returned by the SVD routine), which has to be chosen such that the total amount of signal comes out positive. This is easily achieved by adjusting the sign of $V_j^1$ such that $\sum_{i,j} \log(A_j^i) V_j^i > 0$. The logarithm of the concentration $s_i$, is then computed by projecting the original matrix onto the first principal component $V_j^1$, corrected by a factor $v_{\text{max}}$. This factor accounts for the fact that the vector $V_j^1$ is $L_2$-normalized ($\sum_j (V_j^1)^2 = 1$) by definition in the SVD. The signal then reads:

$$s_i = v_{\text{max}} \sum_{j=1}^{N_p} \log(A_j^1) V_j^1$$

where $v_{\text{max}} = \max_j |V_j^1|$. In addition, the above procedure automatically yields a signal-to-noise ($S/N$) measure for the entire probe set

$$S/N = \frac{D_1}{\sqrt{\sum_{j=2}^{N_p} D_j^2}},$$

where $\{D_k\}$ are the singular values. Large $S/N$ values imply that the probe set measurements in the $N_e$ experiments had a well-defined direction of variation, and can for instance be used as a filter for identifying genes that exhibit significant changes across the experiments tested. In the results shown in Figure 4, we used the following weights

$$w_i = \frac{1}{W} \left\{ \frac{1}{1/(1+b(i-i_s)^2)}, \quad i_s > \Gamma, \quad i_s \leq \Gamma \right\}$$

where $W = \sum_i w_i$, $\hat{s}_i$ is the signal obtained with uniform weights, $i_s = 20$ (out of 28 experiments), $b = 2$, and $\Gamma = \hat{s}_{i_s+1}$ with $\hat{s}_i$ being the ascendantly sorted $\{\hat{s}_i\}$. In other words, lower concentration points are suppressed according to their rank (computed with uniform weights) using a slowly decaying Cauchy weight function. There are of course other weight functions that could serve the same purpose.

Note that the fitting procedure used in the Li–Wong (Li and Wong, 2001) method is identical to an SVD decomposition, however, with different input data than was used here. The three main differences between our method and the Li–Wong technique are: (i) in the analysis here, we used log transformed $PM$ and $MM$ intensities, rather than the bare $PM$-$MM$ values; (ii) we introduced optional weights, which can account for non-linearities of the probe response in the high concentration regime; and (iii) we subtract the column mean before we compute the principal components, which is crucial for capturing the local directions of variation. Indeed, as can be seen in Figure 2, the principal component would be dominated by the mean itself without subtraction.
RESULTS

Data sets

The yeast Latin square (LS) experiment is a calibration data set produced by Affymetrix, that uses a non-commercial yeast chip. 14 groups of 8 different genes, all with different probe sets, are spiked onto 14 chips at concentrations, in pM, corresponding to all cyclic permutations of the sequence (0, 0.25, 0.5, 1, 2, ..., 1024). Hence, each group is probed at 13 different finite concentrations, logarithmically spaced over a range of magnitudes from 1 to 4096 (in Figures 3 and 4, we refer to these concentrations as (1 = 0.25 pM, 2 = 0.5 pM ... 13 = 1024 pM), and each group is completely absent in one array. Besides the spiked-in target cRNAs, a human RNA background was added to mimic cross-hybridization effects that would occur in a real experiment. In addition, each experiment was hybridized twice leading to 2 groups of 14 arrays called R1 and R2.

There are several reasons why this data set is attractive as compared to the similar human and E. coli data available at www.netaffx.com. First, the E. coli data exhibits severe optical saturation, which interferes with the chemical saturation issue we are trying to address here. The yeast data set, on the other hand, has virtually no optically saturated cells, as can be inferred by the SD in the pixel intensities reported in the raw data (.CEL) files. In total, fewer than 0.1% of the probes have SD = 0 (SD = 0 characterizes optical saturated cells). Further, optical saturation is no longer an issue in GeneChips with the current scanner settings. More important, the present data set permits far better statistics, as the number of spiked-in genes is 112 as compared to 14 for the human chip. In the latter data set, there is only one transcript per concentration group as compared to 8 in the yeast case. In fact, we verified that the compression effects discussed below are virtually identical in the human case (not shown).

Summary of two-array methods

The figures in the results section show the log-ratios as a function of concentration in the form of boxplots. In these plots, the central rectangle indicates the median, 1st and 3rd quartiles (Q1 and Q3). The ‘whiskers’ extend on either side up to the last point that is not considered an outlier. Outliers are explicitly drawn and are defined as laying outside the interval [Q1−1.5×I Q, Q3+1.5×I Q], where I Q = Q3 − Q1 is the interquartile distance. For each method, we show three plots, the top two measure the false negative rate for ratios of 2- and 4-fold respectively, and the last one shows the false positive rate. For the top two plots, all combinations (within R1 and R2 separately) of arrays leading to ratios of 2 and 4 were considered, and plotted as a function of their baseline intensity (the lesser of the two concentrations). For the third, each gene was compared between the groups R1 and R2, at identical concentrations. Of the $8 \times 14 = 112$ transcripts, 8 were left out of the analysis because they did not lead to a signal that was tracking the concentrations at all (presumably due to bad probes or transcript quality).

In Figure 3, we summarize the results obtained by the Microarray Analysis Suite 5.0 (MAS 5.0) software and the ‘2 chips’ method discussed in Naef et al. (2002a). The later method computes for each gene probed in two arrays a ratio score $R$ such that

$$\log(R) = \sum_j \log(R_j)$$

is a robust geometric mean (a least trimmed squares estimator) of the probe ratios $R_j$. Figure 3 shows the cases where

$$R_j = \frac{PM_j - MM_j}{PM_j^2 - MM_j^2}$$

and

$$R_j = \frac{MM_j}{MM_j^2}$$

In both cases, only probes with numerator and denominator above background are retained. The first case $(PM - MM)$ is in essence similar to the MAS 5.0 program, differences are in the choice of the Tuckey bi-weight mean as the robust estimator, and in the treatment of negative differences. For our purpose here, we like to think of the Affymetrix method as two-array, $PM - MM$ based method. In all the results presented below, the arrays were scaled according to the MAS 5.0 default settings.

The main features of Figure 3 are: there is an optimal range of baseline concentrations (≈ 1–16 pM) in which the ratio values from both $PM - MM$ methods (the two first columns) are fairly accurate, for both ratios of 2 and 4. For both lower and higher concentrations, there is a noticeable compression effect, which is most dramatic at the high end. At the highest baseline concentration (512 pM for the ratios of 2 and 256 pM for ratios of 4), changes of 2 are basically not detected and real changes of 4 are compressed on average to values around 1.25. The analysis of the false positive rate (last row) shows that both methods yield very tight reproducibility: the log2 ratio distributions are well centered around 0 and the interquartile distances are roughly intensity independent and smaller than 0.2, meaning that 50% of the measurements fall in the ratio interval [0.93, 1.07]. To be fair, we should point out that as a $(PM - MM)$ method, the MAS 5.0 algorithm is on average a bit cleaner, having slightly fewer outliers. However, we like to emphasize that the qualitative behavior in the two
Accuracy and precision in oligonucleotide arrays

Fig. 3. Comparison of ‘two array’ methods: MAS 5.0 (first column), PM-MM (second) and MM only (third) of (Naef et al., 2002a). Boxplots show the log base 2 ratio distributions for each baseline concentration group (cf. text). Row (A): Fold-change of 2, (B): Fold-change 4, (C): Negative controls (false positives). The central rectangle indicates the median, 1st and 3rd quartiles ($Q_1$ and $Q_3$). The ‘whiskers’ extend on either side up to the last point which is not considered to be an outlier. Outliers are explicitly drawn and are defined as laying outside the interval $[Q_1−1.5×IQ, Q_3+1.5×IQ]$, where $IQ = Q_3−Q_1$ is the interquartile distance. Notice the two first rows are qualitatively similar, with the MAS 5.0 being marginally cleaner. Both methods show a strong high concentration compression, but have excellent reproducibility (cf. text). The third column illustrates that MM probes contain valuable signals, often leading to more accurate ratios at high concentrations.

(MM) methods is unchanged, especially as far as the high-intensity compression is concerned. Further, similar behavior is also found using the (PM-MM) Li–Wong method (data not shown). The above observations are consistent with what was reported in Chudin et al. (2001), confirming that these effects are independent of the chip series.

The third column in Figure 3 illustrates our contention that the MM are in essence a set of lower affinity probes. We notice that using only the MM measurements in the two-array method changes the picture qualitatively. Whereas the low concentration regime is far worse than in the (PM-MM) methods, the behavior toward the high end has changed and the drop off occurs now at higher concentrations: approximately 256 pM for the ratios of 2 and 128 pM for ratios of 4. On the other hand, even in the optimal range, the magnitude of the medians are always a bit lower than the real ratios, and the false positive rate also suffers. To summarize, this result suggests that if one is interested in accuracy at high concentrations, then the MM-only methods offer the best two-array alternative. We have tried other variations: PM only, or the double size set consisting of the merged PM and MM, both being worse at high concentrations than the MM only method.

Multiarray methods

The data analyzed using our new method is shown in Figure 4.

It is clear that both are capable of reducing the high intensity compression, as compared to existing methods. The second column explicitly shows the higher accuracy of the local method. It should be noted, however, that the precision is significantly lower than with MAS 5.0, which is the trade-off to pay for higher signal detection.
As compared to the ‘two chip’ MM method, which was previously the least compressive in this regime, the medians are systematically more accurate. Also, the method does not perform well at low-concentrations which is expected since it was not designed for that range.

**Significance scores**

Although ratio score may suffer severe compression, there remains the possibility that they would be attributed a significant increase or decrease call.

Figure 5 displays the relation between the MAS 5.0 log-ratios and their associated $P$-values. MAS 5.0 change ‘$P$-values’ $p_M$ are symmetric about 0.5 and designed such that the ratio score is called increased when $p_M < \gamma_1$ and decreased (D) when $p_M > 1 - \gamma_1$, with a default $\gamma_1 = 0.0025$. This definition is not well suited for plotting purposes, we therefore work with $p_{MAS} = p_M$ when $p_M < 0.5$, and $p_{MAS} = 1 - p_M$ otherwise. This way, both I and D genes have $p_{MAS} < \gamma_1$, the direction being given by the sign of the log-ratio. The results show that there are remarkably few false positive calls: only 4 out of 624 for concentrations $c \leq 8$ pM, and 6 of 728 when $c \geq 16$ pM. Fold changes of 4 are also well detected despite the compression at high intensities: there are 21 false negatives (and 3 false positives having ratios with the wrong sign) out of 1248 for $c \leq 8$ pM, and 84 of 1040 false negatives for $c \geq 16$ pM. The situation deteriorates for fold changes of 2, with 124 false negatives (and 3 false positives) out of 1248 for $c \leq 8$ pM, and 425 of 1040 false negatives for $c \geq 16$ pM.

**DISCUSSION**

We have shown that high-concentration bias is a serious issue in GeneChips, which is probably related to chemical saturation in the adsorption process of the target to the probes. Exploiting the broad range of affinities of different probes ($PM$ and $MM$ included) offers an approach toward improvements. However, the gain in accuracy comes with an expected decrease in precision, since effectively, the weight of a measurement is transferred to a smaller set of probes. Hence, the reduction in noise levels resulting from averaging over probes is diminished.

Our method should serve as a useful complement to
CONCLUSION

We have summarized the performance of existing methods for analyzing Affymetrix GeneChip data, using the yeast calibration data set from Affymetrix. The results show unambiguously the compressive tendency of GeneChip measurements in the high-intensity range, namely that fold changes as large as 4 in expression levels can be reduced to fold changes barely larger than 1 (Figure 3). Interestingly, we showed that among the standard techniques, the one using only the MM signals offers the highest accuracy at high concentrations. Additionally, we have how it is possible to achieve higher accuracy at high concentrations by exploiting the probe set’s wide affinity range. One should realize, however, that saturation problems of the sort encountered present a hard challenge in signal processing, and it is therefore expected that higher accuracy is obtained at the expense of reduced precision.

Our observations raise the following design issue for oligonucleotide arrays. Since it will likely be difficult to manufacture oligonucleotide probes with linear responses over 4 or more decades in concentration, an option would be to optimize the design of probe sets such that each of its probe would be optimally linear in smaller ranges (say 2 decades at most) centered around graded concentrations. In this way, the weights of a measurement could be transferred to an appropriate subset of probes known to be optimal in a given range. Hence, one would use a different set of probes for high or low concentration values to increase the overall dynamic range of a probe set.

ACKNOWLEDGEMENTS

The authors are thankful to Affymetrix for having provided useful calibration data. F.N. is a Bristol-Meyers Squibb Fellow and acknowledges support from the Swiss National Science Foundation. This work was also supported by NIH/NINDS Grant NS39662 (N.D.S) and the Meyer Foundation (M.M.)

REFERENCES


Fig. 5. $P$-values versus log-ratios in for MAS 5.0. The plotted $P_{\text{MAS}}$ is the transformed MAS 5.0 $P$-value (cf. text). The dotted line indicates the default $\gamma_1 = 0.0025$, below which MAS 5.0 scores are considered increased (I) or decreased (D) (for the transformed $P$-value). Baseline intensities are grouped: 0.25–8 pM (left panel), 16–1024 pM (right panel).