JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 1, 2003

Pp. 95–102

Contrast Normalization of Oligonucleotide Arrays

MAGNUS ÅSTRAND

ABSTRACT

Affymetrix high-density oligonucleotide array is a tool that has the capacity to simultaneously measure the abundance of thousands of mRNA sequences in biological samples. In order to allow direct array-to-array comparisons, normalization is a necessity. When deciding on an appropriate normalization procedure there are a couple questions that need to be addressed, e.g., on which level should the normalization be performed: On the level of feature intensities or on the level of expression indexes? Should all features/expression indexes be used or can we choose a subset of features likely to be unregulated? Another question is how to actually perform the normalization: normalize using the overall mean intensity or use a smooth normalization curve? Most of the currently used normalization methods are linear; e.g., the normalization method implemented in the Affymetrix software GeneChip is based on the overall mean intensity. However, along with alternative methods of summarizing feature intensities into an expression index, nonlinear methods have recently started to appear. For many of these alternative methods, the natural choice is to normalize on the level of feature intensities, either using all feature intensities or only perfect match intensities. In this report, a nonlinear normalization procedure aimed for normalizing feature intensities is proposed.

Key words: oligonucleotide array, normalize, curve-fitting, orthogonal, loess.

1. INTRODUCTION

THE USE OF MICROARRAYS TO MEASURE ABUNDANCE of mRNA sequences in biological samples has emerged the last couple of years. One technology commonly used in this context is Affymetrix oligonucleotide arrays. The starting point of this technology is a sample of cells or tissue from which the researcher isolates RNA from which complementary DNA (cDNA) is generated. Then follows transcription from the cDNA to complementary RNA (cRNA), which after fragmentation is put to hybridize on the array. After the hybridization, excess cRNA is washed off, and the final step before scanning the array is staining. The result, after the researchers efforts, is the scanned intensity image, which is the starting point of the low-level analysis of microarrays such as image analysis, feature extraction, and normalization.

Image analysis and feature extraction are in themselves a great challenge. The aim is to select pixels representing each feature and summarize them into a feature intensity. Since each feature is represented by approximately 8×8 pixels, this a great reduction of the data. However, this issue will not be addressed further here; instead, procedures on the level of feature intensity or higher will be discussed.

Statistical and Mathematical Science, AstraZeneca R & D Mölndal, S-431 83 Mölndal, Sweden.

When analyzing data from oligonucleotide arrays, normalizing is a necessity to allow direct array-to-array comparisons. This is because the overall brightness of the scanned image can differ substantially from one array to the next. An even better understanding of this problem is attained by scatter plots with the feature intensities of one array on the y-axis and the feature intensities of another array on the x-axis (Fig. 1). The main sources of this variation in feature intensity level between arrays are the different steps prior to obtaining the intensity image together with the quality of the arrays.

The most commonly used normalization procedure is probably the one implemented in the Affymetrix software GeneChip. This procedure is based on the Affymetrix expression index *average difference* (AD), which is the average difference between the perfect match intensity (PM) and the mismatch intensity (MM). For each array, a trimmed mean of all probe's AD is calculated, and normalization factors for AD are determined by ratios of such means or by a ratio to a target mean AD. Hence, this is a linear procedure on the level of expression index where all probes are used.

Recently, alternatives to the Affymetrix expression index AD have started to appear, e.g., the model-based expression index (MBEI) introduced by Li and Wong (2001a) and an index based on log(PM-BG), suggested by Irizarry *et al.* (2002), where BG is a global estimate of the background intensity of the array. For such alternative expression indexes, it is more natural to normalize on a lower level of the data, i.e., on the level of feature intensities using only PM or PM and MM together.

Such a normalization procedure is described by Li and Wong (2001b). In this procedure, a baseline array is selected to which the other arrays are normalized by fitting a smooth curve. Prior to fitting the curve, a subset of features with small absolute rank differences is selected. The argument for using this kind of subset selection is that we can expect features belonging to an unregulated gene to have similar intensity ranks on two arrays. The curve is then fitted using these features only. In contrast to the normalization procedure in GeneChip, this is a nonlinear procedure on the level of feature intensities that uses a subset of features. Another procedure suggested by Bolstad *et al.* (2002) uses the distribution of all feature intensities. An *average distribution* is derived by first computing the quintiles of each array separately, and then the quintiles are averaged across the arrays. In relation to the method in Li and Wong (2001b), this procedure uses all features. It is also substantially simpler.

In this report, a method for normalizing using smooth curves is proposed. It is a method meant for normalizing the feature intensities, i.e., the PM and MM intensities. But the method can just as well be applied to PM-MM or an expression index derived from the feature intensities. The method proposed by Li and Wong (2001b) uses smooth curves fitted in scatter plots with the baseline array on the y-axis and the array to be normalized on the x-axis. Another solution is to fit a smooth curve in scatter plots with the feature intensity differences on the y-axis and the intensity means on the x-axis (often the intensities are logged before computing the differences and means). This is the basis of the method proposed in this report. We will start by describing the proposed method, termed *contrast normalization* (CN), and then discuss it together with the other methods mentioned. We will also have a look at how these methods perform.

2. RESULTS

2.1. Contrast normalization

Suppose we have a set of k arrays that are to be normalized; each array is represented by n feature intensities. Let the nxk matrix Y denote the intensities of these arrays. Hence, the element in row i and column j of Y is the unnormalized feature intensity of feature i on array j.

2.1.1. Change of basis. In the first step, these intensities are logged and transformed using the matrix M:

$$Z = [x, y_1, \ldots, y_{k-1}] = \log(Y) \cdot M'. \tag{1}$$

Here, M is an orthonormal kxk matrix; i.e., the rows of M are mutually orthogonal unit vectors. Moreover, the first row of M is always the 1-vector times $\sqrt{1/k}$, and then it follows that the other rows are a set of orthonormal contrast. Matrixes such as M will be called *transformation matrixes* hereafter. Note that with

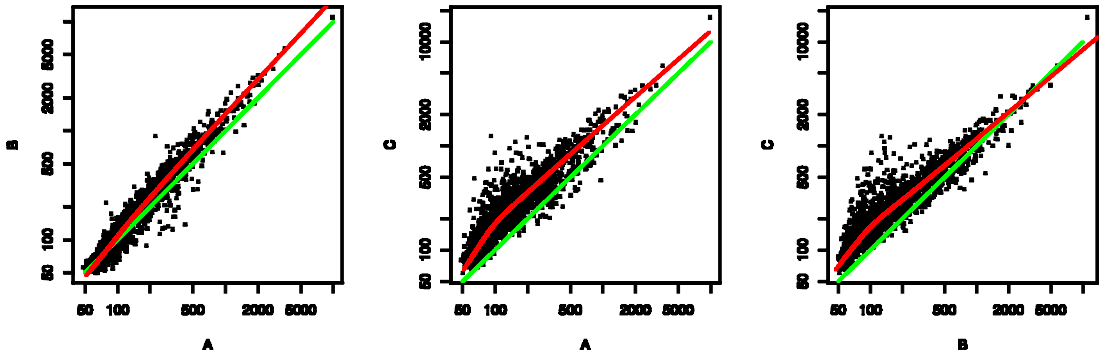

FIG. 1. Scatter plots of 3 arrays, A, B, and C, prior to normalizing. The red curve is the fitted normalizing curve fitted using the alternative basis shown in Fig. 2.

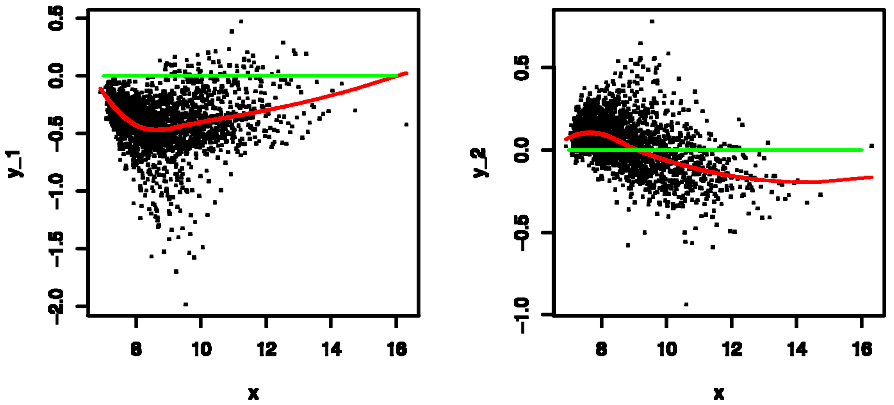

FIG. 2. Contrast plots. Scatter plots of the 2 contrasts against the mean for 3 arrays, A, B, and C, prior to normalizing. The red curve is the fitted normalizing curve, and the green line is the reference line.

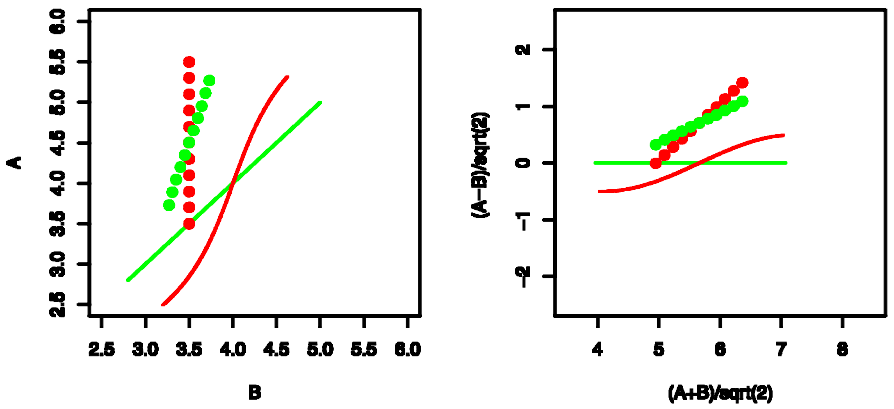

FIG. 3. Normalizing prior change to original basis. This figure illustrates how feature intensities from two arrays are normalized using the alternative basis. The graphs show logged intensities in original basis (left graph) and alternative basis (right graph). Red line is the fitted normalizing curve (fitted using the alternative basis in the right graph), and green line the reference line. The graphs show a couple of features which all have intensities equal to 3.5 on array B, and intensities ranging from 3.5 to 5.5 on array A prior to normalizing (red dots). The intensities are normalized using the alternative basis in the right graph yielding the normalized intensities shown as green dots. Now the features have intensities ranging from 3.25 to 3.75 on array B.

this specification M is unique for k equals 2, but this is not the case for $k > 2$. When k equals 2, we get $M = M_2$, and when k equals 4 we can use $M = M_4$:

$$M_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \sqrt{\frac{1}{2}} \qquad M_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \frac{1}{2} \tag{2}$$

This use of an orthonormal matrix is just a change of basis, where the rows of M form the new basis, denoted the *alternative basis* from now on. When k equals 2, we see that, besides a constant, the alternative basis corresponds to what is called a Bland and Altman plot of the log feature intensities, i.e., a plot of the difference versus the mean. The change to logarithmic scale before performing the change of basis is used to make the error variances more homogenous.

2.1.2. Fitting the normalizing curve. Using the alternative basis, we then fit the normalizing curve: we use the first column of the transformed intensities in Z, i.e., x, as a predictor for column $2, \ldots, k$ of Z, i.e., $y_1, \ldots, y_{k-1}$. When doing this, it's important to have in mind that the set of orthonormal contrasts is not unique. Thus, the method for fitting the curve should be invariant with respect to choice of contrast. Suppose that $[x, y_{a1}, \ldots]$ and $[x, y_{b1}, \ldots]$ are the intensities obtained when transforming according to (1) using the transformation matrixes M_a and M_b, respectively. If $\hat{y}_{a1}, \ldots$ and $\hat{y}_{b1}, \ldots$ are the corresponding fitted curves, $[x, \hat{y}_{a1}, \ldots] \cdot M_a$ should equal $[x, \hat{y}_{b1}, \ldots] \cdot M_b$.

In order to achieve this, we fit a smooth curve using a local regression model (loess) to each vector y_i. For the curve to be less sensitive for outliers, we use a redescending M estimator with the bisquare weight function as is done in the R-function loess (Chambers and Hastie, 1997), but with one important modification. If $\hat{y}_1, \ldots, \hat{y}_{k-1}$ are the vectors of the fitted values, we take $\hat{\epsilon}$ as the Euclidian distance between the rows of the $n\text{x}(k-1)$ matrixes $[y_1, \ldots, y_{k-1}]$ and $[\hat{y}_1, \ldots, \hat{y}_{k-1}]$.

$$\hat{\epsilon} = \sqrt{\sum_{i=1}^{k-1} (\hat{y}_i - y_i)^2} \tag{3}$$

Thus, in each iteration, the same set of robust weights is used for each of the $k-1$ contrast vectors, and these weights are invariant to the choice of orthonormal contrasts. Further, since the local regression model is fitted using weighted least squares, the fitted curve is invariant to the choice of orthonormal contrasts.

2.1.3. Normalizing the arrays. The normalizing curve can be represented with the matrix $[x, \hat{y}_1, \ldots, \hat{y}_{k-1}]$. These sets of points can be viewed either using the original basis or the alternative basis, the red curves in Figs. 1 and 2, respectively. Hence, we still could choose a baseline array and normalize the others by the fitted normalizing curve, e.g., using the two rightmost graphs in Fig. 1 and normalizing A and B to the baseline array C. If doing so, we have used a normalizing curve that is invariant to the choice of baseline array. But the scale to which we normalize still depends on which baseline array we choose.

Another way of normalizing the arrays using the fitted curve is to simply subtract the fitted values using the alternative basis, i.e., Fig. 2, and then go back to the original basis using the matrix M. In this case, the normalized and unlogged intensities would be

$$\exp\left\{[x, y_1 - \hat{y}_1, \ldots, y_{k-1} - \hat{y}_{k-1}] \cdot M\right\}. \tag{4}$$

But this results in a nonsmooth normalizing procedure, in the sense that intensities being equal on one array prior to normalizing may not be equal after. Figure 3 shows why this is the case.

However, we can still use the normalizing curve with the alternative basis. The matrix $[x, \hat{y}_1, \ldots, \hat{y}_{k-1}]$ is a representation of the normalizing curve, and the matrix $[x, 0, \ldots, 0]$ is a representation of what the curve should be after normalization. Hence, the mapping

$$[x, \hat{y}_1, \ldots, \hat{y}_{k-1}] \longmapsto [x, 0, \ldots, 0] \tag{5}$$

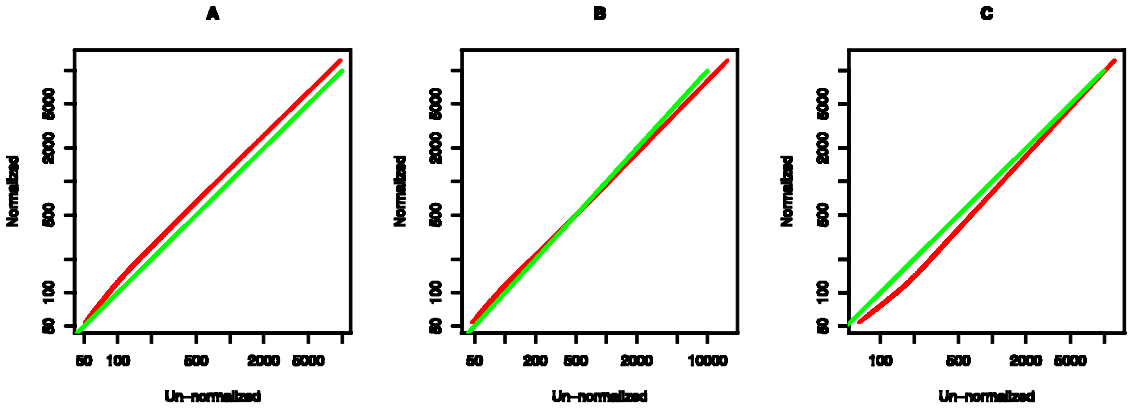

FIG. 4. Normalizing functions. The normalizing functions defined through the mapping of the fitted normalizing curve on to the reference line in Fig. 2 are shown for three arrays, A, B, and C (red line). The green line is the reference line with slope one and zero intercept.

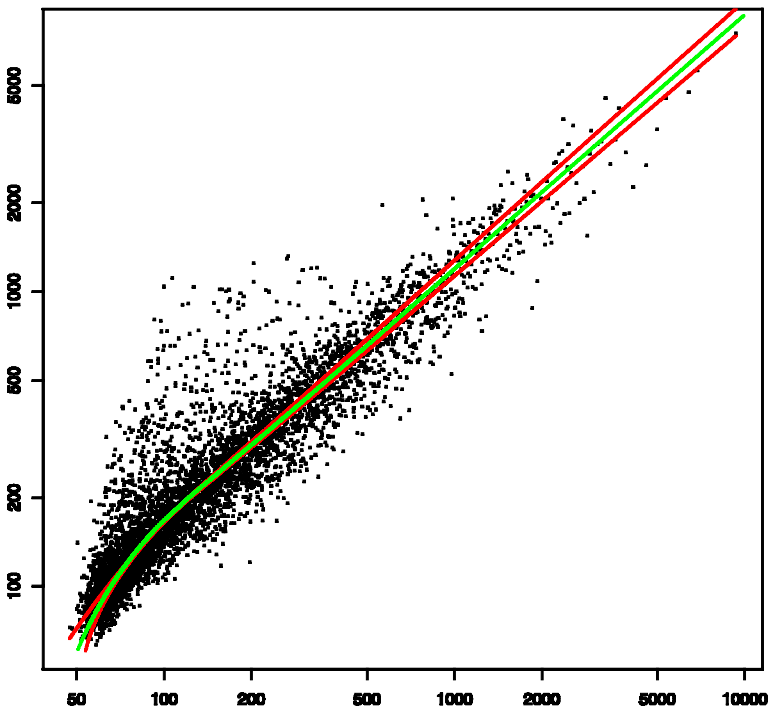

FIG. 5. Scatter plot of 2 arrays, A and B. The two red lines are fitted loess curves from using A (and B) as a predictor for B (and A). The green curve is a loess curve fitted using the alternative basis.

defines a transformation that does the job of evening out the contrast for the alternative basis. Moreover, the mapping

$$\exp\left\{[x, \hat{y}_1, \ldots, \hat{y}_{k-1}] \cdot M\right\} \longmapsto \exp\{[x, 0, \ldots, 0] \cdot M\} \tag{6}$$

defines the same transformation but for the original basis and anti-logged scale. This transformation forms a function $F : R^k \mapsto R^k$ that row-by-row normalizes the matrix of intensities Y. Thus, if $F((x_1, \ldots, x_k)) = (f_1(x_1), \ldots, f_k(x_k))$, f_j is function that normalizes array j. These functions, f_1, f_2, and f_3 for the set of three arrays A, B, and C, are shown in Fig. 4.

One may note that

$$[x, 0, \ldots, 0] \cdot M = \frac{1}{\sqrt{k}} \cdot [x, x, \ldots, x] \tag{7}$$

where $x/\sqrt{k}$ equals $\overline{\log(Y)}$, i.e., the mean across the rows of $\log(Y)$. Hence, this procedure normalizes to a scale determined by $\exp\{\overline{\log(Y)}\}$, i.e., the geometric mean of the arrays.

2.1.4. Adding arrays. Suppose a set of arrays have been normalized and further analyzed, e.g., expression indexes have been computed. Now we have an additional set of arrays that we would like to add to the original ones to use in the same analysis. We would like to do this without affecting the intensities of the original set. This can be done by first normalizing the new set of arrays separately. These arrays are then normalized to a scale determined by their geometrical mean. Thus, we have to transform these to the same scale as the original set, i.e., the scale determined by the geometrical mean of the arrays in the original set.

Let Y_1 and Y_2 be the normalized intensities of the original and new set of arrays, respectively. Also, let x_1 and x_2 be the mean across the rows of $\log(Y_1)$ and $\log(Y_2)$, respectively. To find a transformation that transforms the new arrays with the same scale as the original arrays, we apply the normalizing method treating x_1 and x_2 as log-intensities of two "arrays." If $\hat{y}_1$ is the fitted values for the contrast of these "arrays," we have the mappings

$$\exp\left\{\frac{x_1 + x_2}{2} + \frac{\hat{y}_1}{\sqrt{2}}\right\} \longmapsto \exp\left\{\frac{x_1 + x_2}{2}\right\}, \tag{8}$$

$$\exp\left\{\frac{x_1 + x_2}{2} - \frac{\hat{y}_1}{\sqrt{2}}\right\} \longmapsto \exp\left\{\frac{x_1 + x_2}{2}\right\} \tag{9}$$

that form the functions f_1 and f_2 that would normalize the two "arrays" to a common scale. However, we only want to change the scale of the second one (x_2). To do this, we apply $f_1^{-1} \circ f_2$ formed by the mapping

$$\exp\left\{\frac{x_1 + x_2}{2} - \frac{\hat{y}_1}{\sqrt{2}}\right\} \longmapsto \exp\left\{\frac{x_1 + x_2}{2} + \frac{\hat{y}_1}{\sqrt{2}}\right\} \tag{10}$$

on the intensities of the second "arrays." Hence, the $f_1^{-1} \circ f_2$ is the function that transforms the intensities of the new set of arrays to the scale of the original set.

2.1.5. Software. The contrast normalization as described above is included in the R package affy as the "contrast" option in the "normalize" method (Irizarry *et al.*, 2003). The affy package is available through the open source software project Bioconductor (*www.bioconductor.org/*).

3. DISCUSSION

The usage of curve fitting by normalizing to a baseline array is perhaps the most intuitive way. However, it has one obvious drawback: we have to choose the baseline array, i.e., the array to place on the y-axis. How important this drawback is for the result of the downstream analysis of the raw intensities, i.e., computing expression indexes, is hard to tell. However, by normalizing and computing MBEI (Li and Wong, 2001a) of two arrays, A and B, first using array A as the baseline array, and then a second time using array B as the baseline array, we get an indication that it's not negligible: For each choice of baseline array, the ratios of MBEI, array A to array B, was computed. Of the 8,799 probes, the ratio differed more then 10% (10% greater or smaller) for 1,603 probes (18%), when using array A as baseline instead of array B. The difference was most notable among probes with small indexes. But even among the 2,500 probes with highest expression indexes, the ratio differed more than 10% for 10% of the probes. The two sets of probes filtered out, based on the confidence interval and absolute difference, for each choice of baseline array, differed. There were 469 and 298 probes in the two sets of which 265 were contained in both sets.

TABLE 1. STANDARD DEVIATION COMPARISON[a]

Baseline dataset	*Comparative dataset*					
	UN	*LR*	*CN2*	*CN1*	*QN*	*CN3*
UN	-(-)	89(2.44)	89(2.55)	90(2.51)	90(2.51)	90(2.53)
LR	11(1.42)	-(-)	57(1.18)	57(1.21)	58(1.19)	57(1.22)
CN2	11(1.43)	43(1.14)	-(-)	50(1.04)	53(1.05)	52(1.03)
CN1	10(1.41)	43(1.16)	50(1.04)	-(-)	52(1.02)	53(1.03)
QN	10(1.40)	42(1.14)	47(1.05)	48(1.02)	-(-)	51(1.03)
CN3	10(1.42)	43(1.17)	48(1.02)	47(1.03)	49(1.03)	-(-)

[a]Using two sets of replicated Mu11KsubA arrays (4 replicates in each set), the standard deviation (STD) for each feature across the replicates was computed. The Mu11KsubA array has a total of 262,560 features, which are used for 6,584 probes. The STD for each feature was computed using 6 different sets of intensities: Un-normalized intensities (UN), intensities normalized using linear regression (LR), normalized using QN (QN), and using CN, using all features and a loess span equals 2/3 (CN1), using a subset of 10,000 features and loess span equals 2/3 (CN2) and 0.2 (CN3). The values in upper triangle show the percentage of features of which the comparative dataset had smaller STD then the baseline dataset. For those features, the median STD ratio (baseline dataset to the comparative dataset) is shown within brackets. The values in lower triangle show the percentage of features of which the baseline dataset had smaller STD than the comparative. For those features, the median STD ratio (comparative dataset to the baseline dataset) is shown within brackets.

Moreover, Fig. 5 shows a scatter plot of the raw intensities of the same two arrays together with three curves. The two red curves are the loess curves fitted using the original basis with A as the predictor of B and vice versa. The green curve is the loess curve but fitted using the alternative basis. There is a notable difference between the two red curves with the green curve lying in between. Again, there is a clear indication that the choice of baseline array is not just a theoretical matter.

On the other hand, this approach is simple to apply to a set of k arrays: simply choose a baseline array and normalize the other to that array. Also, if an analysis has been done on a set of arrays, it's easy to add arrays without affecting the analysis of the original set. Just use the baseline array of the original arrays to normalize the new arrays.

When using CN, or the quintile normalization (QN) suggested by Bolstad *et al.* (2002), there is no choice of baseline array; instead all arrays are treated uniformly. But it's not as straightforward to add extra arrays without affecting the result of the analysis of the original set. But the solution in Section 2.1.4 does the job when using CN, and a similar solution for the QN method is to simply use the average distribution of the original set of arrays for the new arrays. Both methods normalize to a scale determined by the geometrical mean across the arrays, in contrast to normalizing using a baseline array where the baseline array determines the scale.

The normalizing procedure described in Li and Wong (2001b) uses a baseline array to which the other arrays are normalized. As mentioned, a curve is fitted using a subset of features. These subsets are derived separately for each of the arrays that are normalized and the baseline array. Hence, there is a risk of using different features when normalizing array 1 as when normalizing array 2. Another way of finding a subset of features is to compare the ranks across all arrays. This could be done using the mean square error (MSE) of the ranks or the range of the ranks. The later was used in the comparisons of Table 1.

In Table 1, the normalized feature intensity standard deviations (STD) over two sets of replicated Mu11KsubA arrays are compared. All methods reduced the standard deviation compared to unnormalized intensities, and CN and QN show somewhat smaller STD's than linear normalization. The different versions of CN (using a subset or all features and different span parameters for the loess model) and QN perform similarly with a slight tendency toward QN and CN using a subset and a small span parameter being the better ones.

REFERENCES

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. 2002. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* (To appear).

Chambers, J.M., and Hastie, T.J. 1997. *Statistical models in S.*, Chapman and Hall.

Irizarry, R., Gautier, L., and Cope, L. 2003. An R package for analyses of affymetrix oligonucleotide arrays. In Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L., eds., *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York. (To appear).

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. 2002. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* (To appear).

Li, C., and Wong, W.H. 2001a. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* 98(1), 31–36.

Li, C., and Wong, W.H. 2001b. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol.*, 2(8), research 0032.1–0032.11.

Address correspondence to:
Magnus Åstrand
AstraZeneca R & D Mölndal
S-431 83 Mölndal
Sweden

E-mail: magnus.astrand@astrazeneca.com