JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 5, 2003

Pp. 751–762

Normalization of DNA-Microarray Data by Nonlinear Correlation Maximization

D. FALLER,[1] H.U. VOSS,[1] J. TIMMER,[1] and U. HOBOHM[2,3]

ABSTRACT

Signal data from DNA-microarray ("chip") technology can be noisy; i.e., the signal variation of one gene on a series of repetitive chips can be substantial. It is becoming more and more recognized that a sufficient number of chip replicates has to be made in order to separate correct from incorrect signals. To reduce the systematic fraction of the noise deriving from pipetting errors, from different treatment of chips during hybridization, and from chip-to-chip manufacturing variability, normalization schemes are employed. We present here an iterative nonparametric nonlinear normalization scheme called simultaneous alternating conditional expectation (sACE), which is designed to maximize correlation between chip repeats in all-chip-against-all space. We tested sACE on 28 experiments with 158 Affymetrix one-color chips. The procedure should be equally applicable to other DNA–microarray technologies, e.g., two-color chips. We show that the reduction of noise compared to a simple normalization scheme like the widely used linear global normalization leads to fewer false-positive calls, i.e., to fewer genes which have to be laboriously confirmed by independent methods such as TaqMan or quantitative PCR.

Key words: DNA-microarray, nonlinear normalization.

INTRODUCTION

IT IS FAIR TO STATE THAT THE INITIAL DNA-chip hype-and-hope phase from a few years ago, which was fueled rather by the excitement about the enormous potential the technology was offering, is now going to be accompanied by a more sound recognition of its traps and pitfalls. Even if the very same RNA sample is put on a series of repetitive chips, we often have to deal with considerable signal variability for the same gene on repetitive chips. Part of this variability is of biological nature, i.e., results from, e.g., different gene expression over cell culture flasks of the same cell type due to subtle different nutrition states, or from different gene expression in organs of separate individuals due to a different genetic background. Biological variability cannot be eliminated by normalization, while systematic variability can. Systematic variability again has different sources, e.g., chip-to-chip manufacturing differences; unsturdy laboratory sample preparation, hybridization, and washing protocols; imprecise signal measurements coming from the scanner; and subtle gene-to-gene differences in hybridization efficiency leading to intergene variability

[1]Freiburg Center for Data Analysis and Modeling, Eckerstrasse 1, 79104 Freiburg, Germany.
[2]F. Hoffmann-La Roche, Ltd., Pharma Research, Building 69, Room 209, CH-4070 Basel, Switzerland.
[3]University of Applied Sciences, KMUB-Bioinformatics, Wiesenstrasse 14, D-35390 Giessen, Germany.

(see Fig. 1, for example). The resulting noise can raise both the number of genes with false-positive calls (genes wrongly called to be differentially expressed) as well as the number of genes with false-negative calls (genes wrongly assigned to be not differentially expressed) into the dozens or even hundreds per experiment. Experimentators in recent publications in high-impact journals therefore use three to four chip repeats per experimental condition (Kerr and Churchill, 2001; Tusher *et al.*, 2001; Ideker *et al.*, 2001) rather than one or two as in the early days, in order to be able to reduce the number of false calls by statistical means.

Considerable effort has been put into method development for the identification of differentially expressed genes, of pairwise gene expression correlations, and the delineation of gene clusters (see Claverie [1999] for overview). With respect to normalization, however, a majority of published experiments normalize by employing linear global normalization procedures which assume that intensities are related by a constant factor, despite the evidence of spatial or intensity dependent signal biases (Tusher *et al.*, 2001).

A normalization algorithm should be able to at least partially correct systematic errors, i.e., to minimize standard deviation and to maximize pairwise correlation over replicate experiments, while maintaining the dynamic signal range. Noise reduction should span the entire dynamic range; i.e., local improvements in noise reduction in a particular signal range should not be confounded by increased noise in another signal range. Also, the transformation should not decrease the information content; i.e., it should be possible to recalculate the original signal using the inverse transformation.

Here we propose a nonparametric nonlinear normalization scheme called simultaneous alternating conditional expectation (sACE), which is a modification of the ACE algorithm by Breiman and Friedman (1985). The sACE algorithm fulfills the above mentioned criteria. It has been tested on 158 chips in 28 sets of repetitive chip experiments, where each set contains between four and nine repetitive chips. Compared to linear global (LG) normalization, sACE decreased noise, as expressed by relative per-gene standard deviation (rSD, i.e., SD divided by mean), averaged over all genes, in all cases. With respect to

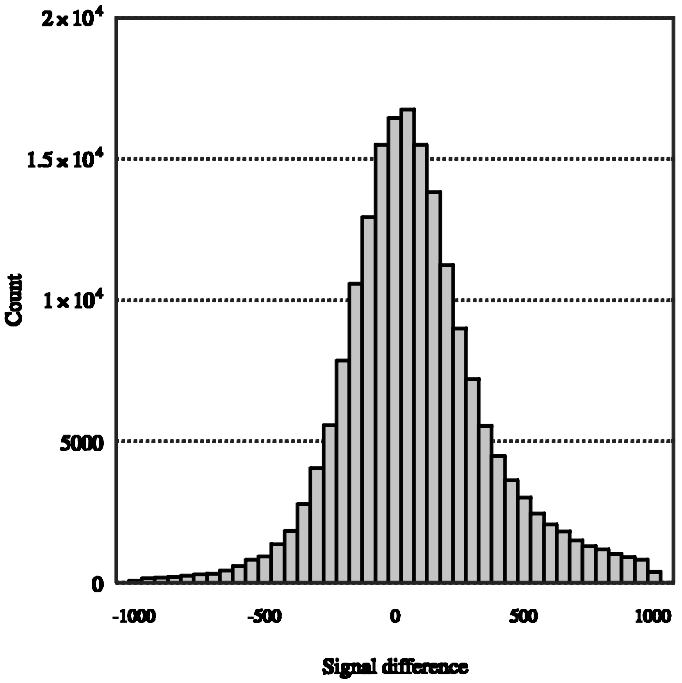


FIG. 1. Signal differences of two chip repeats (chip type HG-U95A) on probe pair level. On Affymetrix chips, the signal for each gene is composed of 16–20 so-called positive match (PM) oligos and the same number of so-called mismatch (MM) oligos. A probe pair is comprised of one PM oligo and one MM oligo, which have the same nucleotide sequence except one central nucleotide. For each XY-location on the chip, the probe pair difference of chip one was subtracted from probe pair difference of chip two ((PM1 − MM1) − (PM2 − MM2)). The histogram illustrates that signals from the same chip location can be very different on two repetitive chips.

false-positive calls, sACE reduced the number of false-positive calls by 57%, as tested on 12 experiments with 6 or more repeats. If one assumes that each differentially expressed gene is verified by an independent method (TaqMan, qPCR, Northern) with a time requirement of about one day per gene, one can estimate how time saving improved normalization of DNA-microarray data can be.

METHODS

Chip usage

Microarray data from 158 Affymetrix chips kindly provided by four different working groups have been used to assess the noise reduction capability of a nonparametric nonlinear normalization procedure (sACE). Chips were from 12 different experimental conditions. For four conditions (C1, C2, C3, C4), we used the entire Hu42K chipset with about 42,000 "genes" (most of them from EST sequences). The Hu42K chipset consists of five chip subtypes (Hu6800, Hu35KsubA, Hu35KsubB, Hu35KsubC, Hu35KsubD). Five chip types with four conditions result in 20 repeat groups. Biological samples were from cell cultures of human macrophages, where each chip represented one cell culture flask. In two conditions (EM, EMI), only the Hu6800 chip was used. In six other experimental conditions (C, NF, HF, FED, VV7, VV8), the rat RgU34A chip was used. Biological samples were tissues from individual rats. Altogether, 28 repeat groups (22 human, 6 rat) were used to assess normalization, with 4–9 repetitive chips per group (see Table 1).

Chip preparation

Chip hybridization, washing, and staining with a strepta vidin-phycoerythrin conjugate were performed using Affymetrix instrumentation according to the companies' recommended protocols.

Chip signal calculation

Per-gene signals were calculated from 40 subsignals of individual oligo probes using the standard algorithm of the Affymetrix GeneChip software called ADI (average difference intensity). ADI may generate negative signals, because the so-called mismatch oligo probes, which are aimed at representing the cross-hybridization portion of the signal, sometimes show a higher signal than the so-called positive match oligo probes. Since negative signals on mRNA expression do not make biological sense, all signals below 10 were adjusted to 10 prior to normalization.

Normalization method

The aim of every normalization algorithm is to minimize the standard deviation and to maximize the pairwise correlation of the repeats. The simplest normalization procedures are linear and global (LG), such as mean or average normalization. Here one tries to adapt the mean or median of different repeats by multiplying each repeat with a constant factor (Alon *et al.*, 1999). (See Zien *et al.* [2001] for a more sophisticated version.) More advanced normalization schemes use nonlinear methods (Yang *et al.*, 2000; Amaratunga and Cabrera, 2000, 2001; Schadt *et al.*, 2001, 1999; Schuchhardt *et al.*, 2000; Dudoit *et al.*, 2000).

In order to present our extended version of the ACE algorithm, we will first give a short review of the standard ACE algorithm in the bivariate case. Suppose we have a microarray experiment with two repeats, the values of the expression level of gene g in repetition i denoted by X_{ig}. If there were no experimental errors, the differences in the measured gene expression could clearly be traced back to the biological variability of the two samples, and there would be no need to normalize the data. However, real world experiments introduce both systematic errors and noise. In order to reduce the noise, several replications are made, but this approach of course does not reduce the systematic errors. Since systematic errors reduce the correlation, one is looking for transformations which increase the correlation between different repeats. This is exactly what the ACE algorithm is designed for.

TABLE 1. NOISE REDUCTION OF sACE NORMALIZATION COMPARED TO LINEAR GLOBAL (LG) NORMALIZATION, EXPRESSED AS AVERAGE RELATIVE STANDARD DEVIATION (rSD)[a]

Exp-Name	*# Repeats*	*Chip type*	*# Genes*	*# Genes used*	*LG*	*sACE*
C1	8	Hu35KsubA	8907	7347	0.68	0.50
C2	4	Hu35KsubA	8907	7199	0.70	0.45
C3	8	Hu35KsubA	8907	7116	0.76	0.52
C4	4	Hu35KsubA	8907	7075	0.66	0.47
C1	8	Hu35KsubB	8924	5919	0.67	0.43
C2	4	Hu35KsubB	8924	6285	0.67	0.51
C3	7	Hu35KsubB	8924	5882	0.67	0.47
C4	4	Hu35KsubB	8924	6035	0.66	0.40
C1	8	Hu35KsubC	8928	6602	0.78	0.55
C2	4	Hu35KsubC	8928	6282	0.73	0.52
C3	8	Hu35KsubC	8928	5753	0.67	0.37
C4	4	Hu35KsubC	8928	5752	0.64	0.43
C1	8	Hu35KsubD	8928	6318	0.62	0.37
C2	4	Hu35KsubD	8928	6810	0.80	0.57
C3	8	Hu35KsubD	8928	6339	0.63	0.37
C4	4	Hu35KsubD	8928	6947	0.78	0.55
C1	8	Hu6800(E)	7129	5240	0.54	0.28
C2	4	Hu6800(E)	7129	5095	0.47	0.33
C3	9	Hu6800(E)	7129	5120	0.49	0.29
C4	4	Hu6800(E)	7129	5043	0.44	0.27
EM	6	Hu6800(E)	7129	5115	0.42	0.24
EMI	6	Hu6800(E)	7129	5122	0.36	0.23
VV7	4	RgU34A	8798	6361	0.37	0.32
VV8	4	RgU34A	8798	6428	0.41	0.32
C	5	RgU43A	8798	6336	0.35	0.27
FED	4	RgU43A	8798	6296	0.36	0.33
HF	5	RgU43A	8798	6311	0.36	0.30
NF	4	RgU43A	8798	6380	0.35	0.26

[a]Since genes with arbitrary low ADI signals are adjusted to 10 up-front, and since genes with very high signals may show scanner saturation effects, only genes with a mean raw signal between 20 and 5,000 (mean calculated over repeats) were used to calculate rSD, representing the bulk of genes with reliable signals. The average rSD values obtained with the two normalization methods are 0.57 ± 0.15 for LG normalization and 0.39 ± 0.11 for sACE normalization. The number of genes used to calculate rSD is given in column "#genes used."

A straightforward application of ACE constructs two transformations $\Phi_i(X_{ig})$, $i = 1, 2$, which minimize the fraction of variance e^2 not explained by a regression of $\Phi_1(X_{1g})$ on $\Phi_2(X_{2g})$,

$$e^2 = \frac{\sum_g \left[\Phi_1(X_{1g}) - \Phi_2(X_{2g})\right]^2}{\sum_g \Phi_1^2(X_{1g})}. \tag{1}$$

These functions, called optimal transformations, also optimize the maximal correlation Ψ^* between the two repeats

$$\Psi^*(X_{1g}, X_{2g}) = \max_{\Phi_1, \Phi_2} R(\Phi_1(X_{1g}), \Phi_2(X_{2g})), \tag{2}$$

where R is the correlation coefficient.

This is achieved by a rather simple iterative algorithm based on alternation between conditional expectations (ACE). The basic idea is that if, e.g., $\Phi_1(X_{1g})$ is known, then $\Phi_2(X_{2g})$ can be computed as the

conditional expectation value of $\Phi_1(X_{1g})$ with given X_{2g},

$$\Phi_2(X_{2g}) = E[\Phi_1(X_{1g}) \mid X_{2g}], \tag{3}$$

where $E[.|.]$ denotes the conditional expectation value. By iterating the computation of the conditional expectation values and introducing normalization factors, the ACE algorithm computes the so-called optimal transformations.

A generalization of this algorithm to the multivariate case is available (see Breiman and Friedman, 1985; Härdle, 1990; and Schimek, 2000). In this case, one obtains transformations which minimize

$$e^2 = \frac{\sum_g \left[\Phi_1(X_{1g}) - \sum_{k=2,N} \Phi_k(X_{kg}) \right]^2}{\sum_g \Phi_1^2(X_{1g})}. \tag{4}$$

This common approach needs to be generalized for the present setting. In an experiment with more than two repeats, the goal is to minimize the pairwise residuals given by

$$e^2 = \sum_{i<j} \frac{\sum_g \left[\Phi_i(X_{ig}) - \Phi_j(X_{jg}) \right]^2}{\sum_g \Phi_i^2(X_{ig})}. \tag{5}$$

To be able to apply this algorithm to DNA microarray data, one needs a modification of ACE which will be presented now.

The main idea is to apply the standard ACE algorithm simultaneously to all pairs of repeats. This leads to $n-1$ different transformations for each repeat. Thus, after every iterative step, the transformations for one repeat are initialized to their mean computed from the previous iteration step. This results in the following algorithm, called sACE:

Initialize $\Phi_i(X_{ig}) = X_{ig} / \sqrt{\frac{1}{G} \sum_g X_{ig}^2}$
Repeat
  For all pairwise comparisons i, j
    $\Phi_j(X_{jg}) = E\left[\Phi_i(X_{ig}) | X_{jg}\right]$
    $\Phi_i(X_{ig}) = E\left[\Phi_j(X_{jg}) | X_{ig}\right] / \sqrt{\frac{1}{G} \sum_g E\left[\Phi_j(X_{jg}) | X_{ig}\right]^2}$
  End For
  Compute mean of $\Phi_i|_{\text{same repeat}}$
while $e^2(\Phi_i, \Phi_j)$ decreases

Here, $E[.|.]$ denotes the estimate of the conditional expectation value, and G is the total number of genes. Note that the conditional expectation value $E\left[\Phi_i(X_{ig}) | X_{jg}\right]$ is a function depending on the random variable X_{jg} and thus is a random variable itself. It is estimated by smoothing the scatterplot $\Phi_i(X_{ig})$ versus X_{jg} using a triangular window over 400 neighboring genes,

$$\Phi_j(X_{jg}) = E\left[\Phi_i(X_{ig}) | X_{jg}\right] = \sum_{l \in I} w_{lg} \Phi_i(X_{il}) / \sum_{l \in I} w_{lg} \tag{6}$$

with

$$I = \{g\} \cup \{\text{indices of } 2n \text{ nearest neighbours of } X_{jg}\} \quad \text{and} \quad w_{lg} = 1 - \left| \frac{l-g}{n} \right|. \tag{7}$$

The parameter n plays the role of a regularization parameter and controls the smoothness of the resulting transformation. In each iterative step, one obtains several different transformations for one replication, $\Phi_i|_{\text{same repeat}}$. After each iterative step, their average value is used to initialize Φ_i for the next iterative step.

To apply the algorithm, the experimental data is first rank ordered (Voss, 2001), before sACE is applied. Thus, the calculation of the conditional expectation in the above scheme is independent of the original distribution or a monotonic transformation of the raw data, e.g., logarithmic transformations.

The "optimal" transformed data is again transformed with a joint transformation to have a similar distribution as the original data. This joint transformation is constructed by mapping the averaged optimal data to the averaged mean raw data. This transformation is then applied to every "optimal" normalized dataset. Raw and transformed data are directly comparable, which simplifies the biological interpretation but may not be necessary if one is interested in statistical tests of significance only.

It is important to note that any systematic part of experimental noise generates statistical dependency, and hence DNA microarray experiments may produce data which are statistically not independent. It can be shown that in two dimensions (as it is the case in sACE) any statistical dependence of different repeats is detected and corrected for by the ACE algorithm (Rényi, 1959). The proposed normalization algorithm is designed to find smooth functions over the intensity of the measurements which do correct for this type of error.

DISCUSSION

The human chips used here represent the first generation of Affymetrix chips where the 40 oligo probes belonging to one gene are in close spatial proximity, while the rat chips used here represent the second generation of Affymetrix chips where the 40 oligo probes belonging to one gene are distributed over the chip. Distributing oligos makes the per-gene signals less vulnerable to local defects and gradients. Gradients are not accounted for using standard linear normalization schemes. Hybridization on the human chips was done about one year before the rat chips, when laboratory protocols were still under improvement. Both improvements on chip design and wet lab procedures led to a significantly lower noise for the rat chips, as expressed by the lower average rSD of the nonnormalized signals. Since biological variability is higher within rat individuals compared to cell cultures (data not shown), the technological improvement is even larger with the new chips than reflected by rSD differences.

Both normalization methods (linear global, sACE) generated signals in the same range (between 10 and 40,000) and with a similar distribution of raw signals (see Fig. 2 as an example). While average rSD calculated over all genes represents a rough estimate of normalization efficiency, a more detailed look at rSD as a function of signal intensity (Fig. 3) shows that sACE is particularly efficient with low signals, without falling behind LG for higher signals. For instance, for the experiment C3-E (Hu6800 chip), sACE generates a considerably lower rSD for genes in the mean raw signal range 100–800, which represent about 30% of all genes on the chip, compared to LG (Fig. 3b). A similar behavior on low-signal genes can be observed for the other experiments (see Figs. 3c–3e for example). Over 22 experiments with first generation chips (chip type Hu*) and comparatively low biological variability (samples from cell culture), sACE reduced average rSD by 24%–48%; over 6 experiments with second-generation chips (chip type RgU34A) and presumably higher biological variability (samples from different individuals), sACE reduced average rSD by 8–26% (Table 1). An edge effect can be observed with sACE in a few cases (2/28) for high-signal genes with chips of particularly low quality (see experiment C1, Hu6800 chip (E-chip) in Fig. 3a as an example), where the rSD increases compared to LG normalization for genes with a mean raw signal above 3,000. However, this adverse behavior afflicts in this case only 83 out of 8,798 genes and can never afflict more than 200 genes at each end of the intensity scale, since smoothing within sACE occurs over 200 neighboring genes in each direction. On the positive side, more than 3,200 genes gain reduced rSD over repeats (see Fig. 3a: improved rSD is obtained for genes with signals between 100 and about 200, i.e., percentage of genes between 0.4 and 0.8, which amounts to 40% of all genes or about 3,200). That is, the majority of genes gains reduced rSD with sACE. An example for a typical transformation obtained by the sACE algorithm is shown in Fig. 4.

In this work, we concentrate on the normalization of DNA microarray data from samples of the same biological entity. Algorithm sACE is equally applicable to the analysis of arrays from different biological

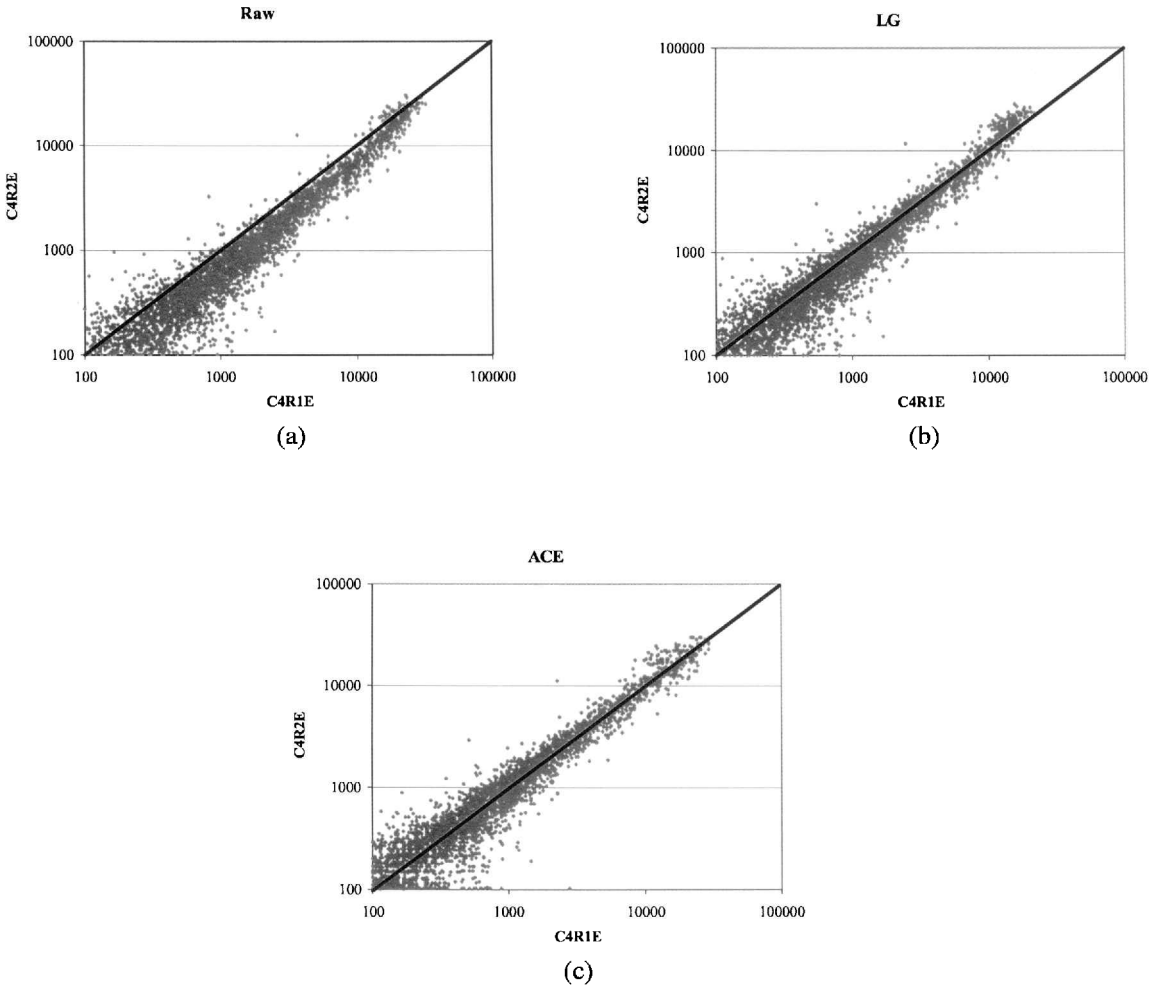
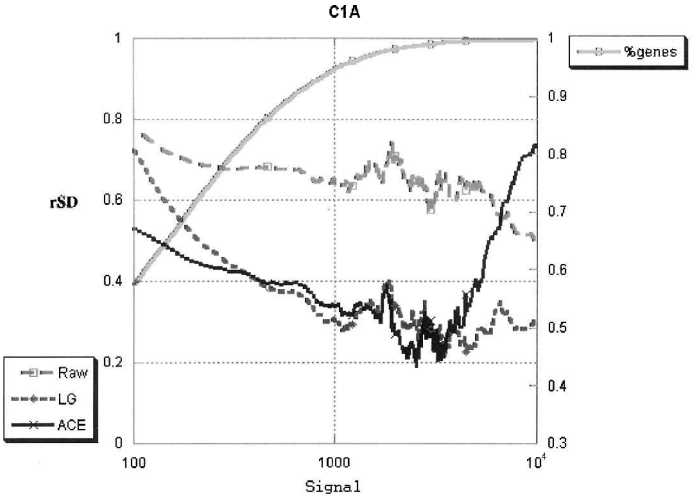


FIG. 2. Signal distribution of raw signals (**a**) after normalization with LG, (**b**) and normalization with sACE, (**c**) exemplified by the two chip repeats C4R1E and C4R2E.

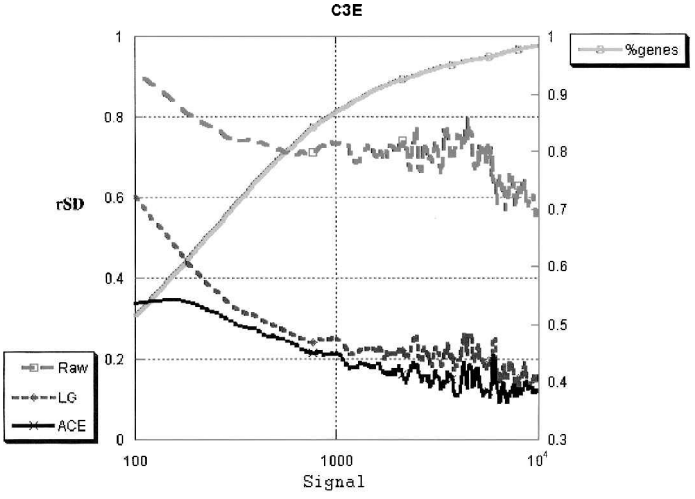
samples, e.g., treatment versus control settings. For multiple conditions, one can apply a first normalization run on chip repeats and a second normalization run using, e.g., a linear normalization method over all chips (see below "False-negative rate" and Table 3). Chip data analysis using a two-step normalization procedure may be the subject of forthcoming work.

False-positive rate

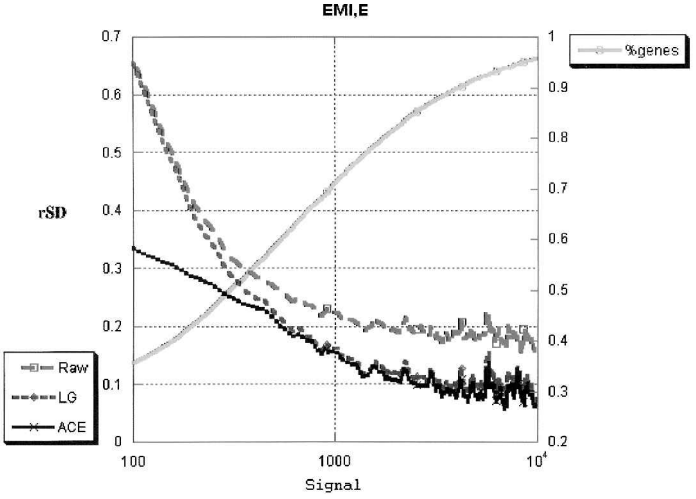
To determine the false-positive rate, we split experiments with six or more repeats into two groups (a minimum of three repeats per condition is needed to apply *t*-test), each group artificially representing a different biological condition. Genes were called false positive if they showed an expression difference of more than two fold (up or down) between conditions. Chip repeats should not show any differential expression. Because of experimental artifacts, it is realistic that there may be genes which are falsely reported to be differentially expressed, though. The number of such genes with such chance differential expression (Table 2) was similar in magnitude to the numbers found by Golub *et al.* (1999) (173/6,817 and 136/6,817, resp.) using a different method. Only genes with a mean raw signal between 100 and 5,000 were considered here. In 10 out of 12 such comparisons, sACE reduced the number of false positives, compared to LG normalization. If false positives were additionally filtered by *t*-test ($p < 0.05$), sACE produced the same number of false positives in 1/12 comparisons and reduced the number of false positives in 11/12 comparisons (Table 2), with an overall reduction in number of false positives by 57%.


C1A
rSD
%genes
Raw
LG
ACE
Signal
100
1000
10^4

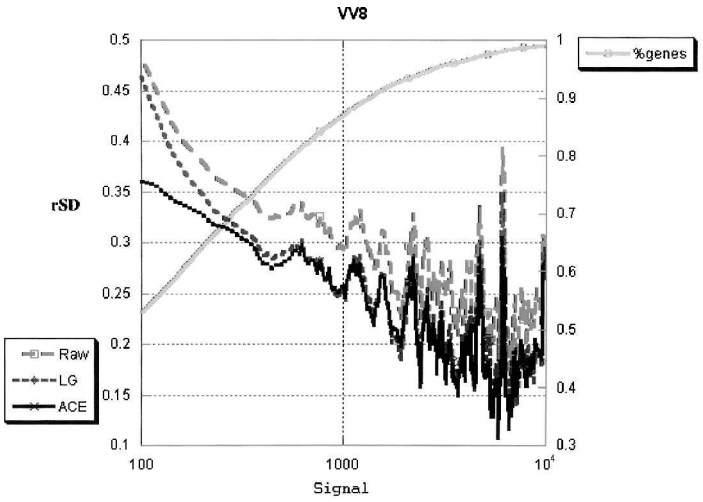
(a)


C3E
rSD
%genes
Raw
LG
ACE
Signal
100
1000
10^4

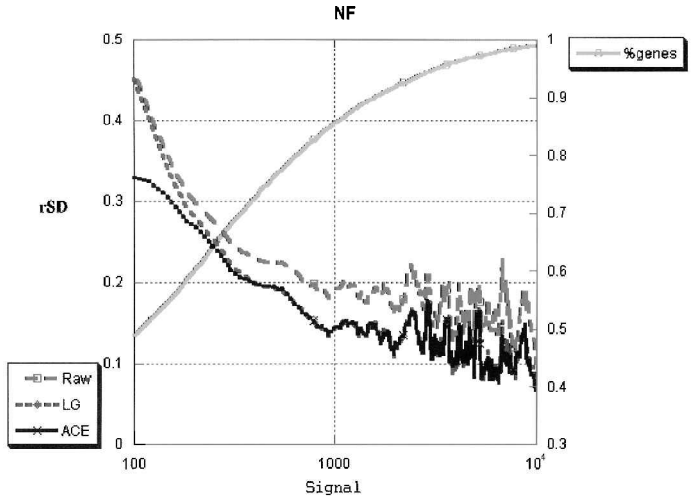
(b)


EMI,E
rSD
%genes
Raw
LG
ACE
Signal
100
1000
10^4

(c)



(d)



(e)

FIG. 3. Dependency of signal variability on signal intensity. Gene signals are binned into windows of size 20 (with mean raw signal used as basis for window binning) and rSD over gene repeats is averaged within each window. Plotted values are smoothed over 9 consecutive windows. The rSD of signals calculated without normalization, with linear global normalization, and sACE normalization for 5 experimental groups C1 (Human35KsubA-chip), C3 (Hu6800 chip), EMI (Hu6800 chip), VV8 (rat chip A), and NF (rat chip A). Experiments were prepared by four different working groups. The cumulated number of genes is shown on the right y-axis (fraction of genes with respect to entire chip); e.g., in Fig. 3e about 85% of all genes have a signal below 1,000.

False-negative rate

The aim of normalization is reduction in variability over repeats. However, this reduction must not go too far and generate compression. Compression would result in false-negatives, i.e., genes which are up- or down-regulated in vivo but not recognized as such *in silico*. The false-negative rate can only be estimated by comparison with an independent method of mRNA expression level measurement, such as Northern-blot, TaqMan, or qPCR. To do this on thousands or even just hundreds of genes is extremely laborious and outside the scope of this work. From principal considerations, we expect that the noise reduction capability of sACE is effective on both sides, i.e., that sACE reduces the number of false negatives in a similar order of magnitude as the number of false-positives, since noise reduction implies significance

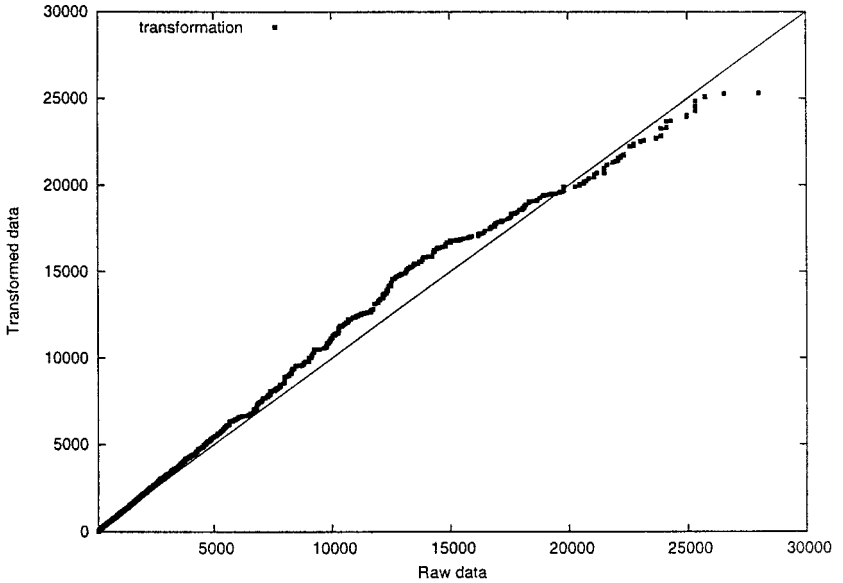


FIG. 4. sACE transformation of experiment C4R2E as example, with signals before (abscissa) and after (ordinate) transformation.

TABLE 2. REDUCTION OF FALSE POSITIVE RATE BY NORMALIZATION[a]

Exp-Name	*# Repeats*	*Chip type*	*# Genes used*	*# False pos. LG*	*Passing t-test*	*# False pos. sACE*	*Passing t-test*
C1	8	Hu35KsubA	4166	70	13	45	5
C3	8	Hu35KsubA	2646	104	6	115	4
C1	8	Hu35KsubB	2043	59	18	24	9
C3	7	Hu35KsubB	1856	31	1	24	0
C1	8	Hu35KsubC	2016	145	6	169	4
C3	8	Hu35KsubC	1372	8	1	5	0
C1	8	Hu35KsubD	2855	53	11	21	7
C3	8	Hu35KsubD	2840	38	8	23	8
C1	8	Hu6800(E)	3531	157	28	26	1
C3	9	Hu6800(E)	3477	103	23	31	10
EM	6	Hu6800(E)	4047	65	7	25	3
EMI	6	Hu6800(E)	4162	55	4	19	3

[a]False positives here are genes with a 2-fold expression difference (up or down) after splitting a set of chip repeats into two different artificial conditions. Only genes with a mean raw signal between 100 and 5,000 were considered for calculation of false-positive rate. The *t*-test was applied with an error probability of less than 0.05 to the genes which showed at least a 2-fold expression difference.

improvement, and significance improvement propagates directly into treatment versus control settings. We have tested this assumption on 10 genes in two different control versus treated settings and find no increase in false negatives with sACE, while significance is improved (*t*-test is smaller overall for sACE plus LG normalized data versus LG normalized data; see Table 3).

CONCLUSION

Here we propose sACE a modification of the ACE algorithm of Breiman and Friedman (1985), for normalization of DNA microarray data. Compared to the widely used linear global normalization, sACE decreases systematic error, as expressed by per-gene standard deviation on more than two repetitive chips,

TABLE 3. COMPARISON BETWEEN TAQMAN, LINEAR GLOBAL NORMALIZATION ALONE (LG) AND SACE PLUS LG NORMALIZATION WITH RESPECT TO FALSE NEGATIVES[a]

	Time 6					*Time 24*				
	LG		*sACE plus LG*		*TaqMan*	*LG*		*sACE plus LG*		*TaqMan*
Gene	*FC*	*t-test*	*FC*	*t-test*	*FC*	*FC*	*t-test*	*FC*	*t-test*	*FC*
31998_at	3.5	0.0001	3.3	0.00001	7.0	3.0	0.0006	2.8	0.00724	4.9
32855_at	1.1	0.81647	1.1	0.72167	1.6	2.0	0.0048	2.0	0.00277	2.1
34776_at	1.5	0.17305	1.4	0.21118	1.8	1.5	0.18498	1.4	0.12129	1.9
36873_at	1.5	0.06182	1.3	0.11863	1.6	4.0	0.00177	3.5	0.00001	6.7
39498_at	1.9	0.35989	1.6	0.19476	1.3	3.1	0.05013	2.0	0.01185	0.9
39950_at	3.0	0.00099	2.8	0.00003	5.2	3.8	0.00045	3.9	0.0002	5.2
41362_at	7.4	0.00178	6.1	0.00014	67.1	9.0	0	8.0	0.00039	31.7
41764_at	1.5	0.00238	1.4	0.00092	1.7	2.9	0.00002	2.7	0.00018	4.4
608_at	1.2	0.06042	1.2	0.05873	1.1	1.7	0.03654	1.5	0.02067	2.2
649_s_at	1.2	0.04974	1.2	0.02752	3.0	2.1	0.00672	2.2	0.01645	3.2

[a]The normalization method sACE plus LG corresponds to applying a linear global normalization after the sACE normalization. For 10 genes in two experimental control vs. treatment settings (Time-6 and Time-24) up- or down-regulation fold-change (FC) was measured both by a DNA-microarray (Affymetrix Hu95A chip) and by two TaqMan experiments. Using a 5% test niveau, no false-negative (i.e., a gene which is reported as differentially expressed by LG alone and TaqMan, but not by sACE plus LG) was found in these 20 cases (data kindly provided by Thomas Grübl and Matthew Wright, Roche Basel).

with a particular positive effect on the majority of small signal genes, which are often the most interesting ones. This noise reduction leads to a substantial reduction in number of false-positively called genes (57% less) as tested in 28 experiments including 158 chips. A similar positive effect may be presumed with respect to false-negatively called genes. Reduction of error in absolute gene expression translates directly into reduction of error of differentially expressed genes, when two or more experiments are compared, such that we expect a reduced number of genes wrongly assigned up- or down-regulated (false positives) and a reduced number of genes differentially expressed but not recognized as such (false negatives). Since in laboratory practice each interesting differentially expressed gene should be verified by an independent method such as TaqMan, qPCR, or Northern-blot, any reduction in the number of genes to be verified should be cordially welcomed by wet-lab biologists.

ACKNOWLEDGMENTS

We thank Angela Schönfelder, Edward Murray, Veronique Voisin, and Dagmar Kube for preparation of chip experiments, Thomas Grübl for TaqMan data, and Martin Beibel for critically reading the manuscript.

REFERENCES

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci USA.* 96(12), 6745–6750.

Amaratunga, D., and Cabrera, J. 2000. Outlier resistance, standardization, and modeling issues for DNA microarray data. Preprint.

Amaratunga, D., and Cabrera, J. 2001. Analysis of data from viral DNA microchips. *J. Am. Statist. Assoc.* 96(456), 1161–1170.

Breiman, L., and Friedman, J.H. 1985. Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Assoc.* 80, 580.

Claverie, J.M. 1999. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* 8, 1821–1832.

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. 2000. *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Technical Report No. 578, Department of Biochemistry, Stanford University.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, D., and Lander, E. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 929–934.

Härdle, W. 1990. *Applied Nonparametric Regression*, Cambridge University Press, New York.

Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Bumgarner, R., Aebersold, R., and Hood., L. 2001. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science* 292, 929–934.

Kerr, M.K., and Churchill, G.A. 2001. Experimental design for gene expression microarrays. *Biostatistics* 2, 183–201.

Rényi, A. 1959. On measures of dependence. *Acta. Math. Acad. Sci. Hungary* 10, 441–451.

Schadt, E.E., Li, C., Ellis, B., and Wong, W.H. 1999. *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* Technical Report No. 303, Department of Statistics, UCLA.

Schadt, E.E., Li, C., Su, C., and Wong, W.H. 2001. Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* 80, 192–202.

Schimek, M., ed. 2000. *Smoothing and Regression. Approaches, Computation and Application*, Wiley, New York.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. 2000. Normalization strategies for cDNA microarrays. *Nucl. Acids Res.* 28(10), e47.

Tusher, V.G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation responses. *Proc. Natl. Acad. Sci.* 98, 5116–5121.

Voss, H.U. 2001. Analyzing nonlinear dynamical systems with nonparametric regression, *in* A. Mees, ed., *Nonlinear Dynamics and Statistics*, 413–434, Birkhäuser, Boston.

Yang, Y.H., Dudoit, S., Luu, P., and Speed, T.P. 2000. *Normalization of cDNA microarray data.* Technical Report, Department of Statistics, UC Berkeley.

Zien, A., Aigner, T., Zimmer, R., and Lengauer, T. 2001. Centralization: A new method for the normalization of gene expression data. *Bioinformatics* 17, 323–331.

Address correspondence to:
U. Hobohm
University of Applied Sciences
KMUB-Bioinformatics
Wiesenstrasse 14
D-35390 Giessen, Germany

E-mail: uwe.hobohm@tg.fh-giessen.de