Improved 2-Color "Exponential" Normalization For Microarray Analyses Employing Cyanine Dyes.

Thomas M. Houts

Introduction

Normalization continues to be an area of controversy and difficulty in gene expression analysis. In single-channel expression studies, such as those involving radioisotopic detection from membranes, the use of "housekeeping genes" as a means to normalize the data has been widely used, with limited success. Others have proposed normalizing single-channel data internally to the distribution of the data.

The advent of fluorescence-based two-color gene expression microarrays promised better precision and discrimination power in identifying differentially expressed genes. However, most normalization schemes have been adapted from those devised for single channel work, including "housekeeping genes", or multiplying by a constant, such as the ratio of the average signal for all genes. The work by Spellman et al[1] takes the further step of looking at log ratios, and searching for a normalization routine to render the average log ratio to a value of zero. This begins the process of doing normalization to achieve the best two-color performance.

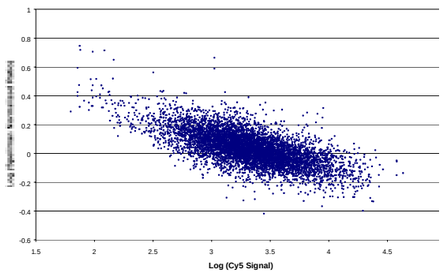
We have extended the use of two-color normalization to account for an unexpected observation, that the required normalization factor is not a constant, but is a function of the Cy™5 signal. The magnitude of the correction required varies from hybridization to hybridization, and is therefore determined empirically from the data for each hybridized image.

Data Sets and Data Analysis

Data from the Alpha Factor and Elutriation experiments in CAMDA 2000 Data Set 1 were used. Spots flagged as "Control", "Empty", "PCR Fail" or "Poor visual quality" were removed from the calculations for normalization. Background-subtracted signals were used as the measure of signal for each spot (CH1D and CH2D). The Spellman et al value for normalized ratio was used as either RAT1N, or 1/RAT2N when RAT1N was less than 1.

Figure 1 presents the normalized log ratios calculated by Spellman et al, for each spot as a function of the Cy5 signal for slide y744n82.

Figure 1: Original Analysis for y744n82



There is a correlation between the Cy5 signal and the ratio, with results distorted toward higher values at low Cy5 signal intensities. Table 1 shows the correlation values (r^2) of the log (Cy3/Cy5) ratios with both the log (Cy5) and log (Cy3) signals for the 32 slides in the alpha factor release and elutriation series experiments. The table, which is sorted by the correlation for Cy5, shows that the correlation is stronger with the Cy5 signal than the Cy3 signal for all except y744n98, which shows hardly any correlation at all.

Table 1. Correlation of log(ratio) with log(signal)

Experiment	r^2 for Cy5	r^2 for Cy3
y744n82	0.453	0.202
y744n81	0.440	0.205
y744n79	0.435	0.214
y744n83	0.434	0.151
y744n43	0.417	0.184
y744n60	0.390	0.195
y744n58	0.354	0.110
y744n63	0.324	0.133
y744n61	0.257	0.000
y744n100	0.242	0.073
y744n84	0.228	0.058
y744n77	0.199	0.044
y744n103	0.198	0.060
y744n70	0.177	0.015
y744n76	0.167	0.013
y744n69	0.158	0.009
y744n89	0.146	0.023
y744n102	0.128	0.010
y744n52	0.119	0.019
y744n40	0.103	0.001
y744n96	0.089	0.000
y744n90	0.089	0.001
y744n97	0.089	0.014
y744n94	0.071	0.002
y744n101	0.071	0.001
y744n41	0.070	0.007
y744n88	0.069	0.001
y744n51	0.049	0.000
y744n73	0.040	0.000
y744n98	0.033	0.048
y744n32	0.017	0.011
y744n72	0.017	0.008

In order to better understand this relationship, we started with the background-subtracted signal intensities for each spot. For each spot, the logarithm (base 10) was calculated for both the Cy3 and Cy5 signals. The difference between these was also calculated, giving the logarithm of the ratio (Cy3 signal / Cy5 signal). Subsequently, the range of log (Cy5 signal) was divided into 50 equally-spaced bins. Each spot was assigned to a bin based on its log(Cy5 signal). For each bin, the average log signal was calculated, as well as the average log ratio. Those results are shown as the red diamonds in Figure 2, with error bars representing the 95% confidence interval for the mean. Another approach to estimating the normalization relationship is shown in the green curved line, obtained by fitting an exponential decay curve to all the points. This fitted curve is an excellent match to the relationship identified by the binned means.

Figure 2. y744n82 Binned Means

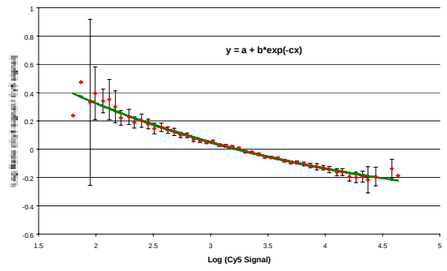


Figure 3: y744n82 Combined

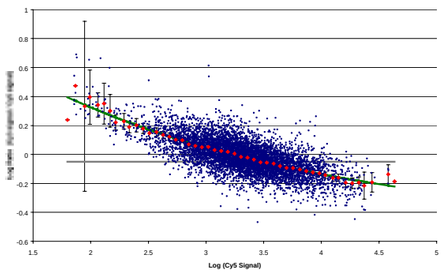
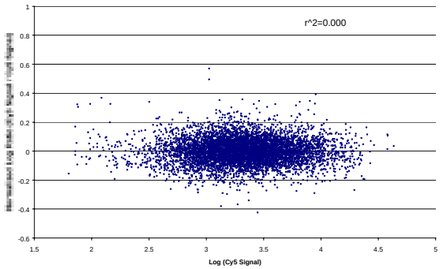


Figure 3 overlays the elements from figure 2 on the individual point data from figure 1.

In this new "exponential normalization" routine, the normalized log ratios are calculated by subtracting the value for the fitted line (based on the Cy5 signal) from the observed log ratio for each spot. The grey line represents the constant value subtracted from each log ratio in the constant normalization method employed by Spellman, et al. The difference between the green line and the grey line represents how much the normalized log ratios would differ between the exponential method and the original method used by Spellman, et al. For the weakest spot on this slide, the difference in log ratio is 0.45, or a difference of about 2.8 fold. The normalized log ratios calculated for slide y744n82 is shown in figure 4.

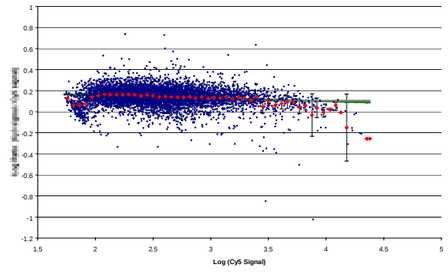
Figure 4: y744n82 Normalized Log Ratios



The results are satisfying in that the normalized log ratios cluster around zero, and there is no correlation with Cy5 signal. There is also no correlation of these normalized log ratios with the Cy3 signal (r^2 = 0.071). In fact, even though achieving a mean value of zero was not a criterion of the regression, the mean normalized log ratio for all spots in the slide is less than 0.001. This pattern continued among all of the slides analyzed for this study, with the mean normalized log ratio being less than 0.01 for each of the 32 slides.

This exponential normalization is robust, in that it approaches the constant normalization of Spellman et al when there is no Cy5 signal correlation to correct for. Figure 5 shows the same elements as Figure 3, but for experiment y744n72, which showed the weakest correlation between signal and average log ratio.

Figure 5: y744n72 Combined



Another advantage of the exponential normalization is that this systematic bias is removed from estimates of the experimental and biological noise of the system. Table 2 compares the distributions of normalized log ratios calculated using the constant normalization of Spellman, et al, and the exponential normalization.

Table 2: Distribution comparison for normalized ratios

	Constant Normalization		Exponential Normalization	
Experiment	Mean	Std. Dev.	Mean	Std. Dev.
y744n82	0.037	0.116	0.000	0.085
y744n81	0.040	0.109	0.000	0.081
y744n79	0.034	0.098	0.000	0.074
y744n83	0.066	0.132	0.000	0.098
y744n43	0.026	0.106	0.000	0.081
y744n60	0.058	0.113	0.000	0.088
y744n58	0.073	0.129	0.000	0.103
y744n63	0.041	0.098	0.000	0.080
y744n61	0.036	0.201	0.000	0.173
y744n100	0.037	0.108	0.000	0.094
y744n84	0.032	0.101	0.000	0.089
y744n77	0.043	0.105	0.000	0.094
y744n103	0.033	0.095	0.000	0.085
y744n70	0.031	0.112	0.000	0.102
y744n76	0.044	0.119	-0.001	0.109
y744n69	0.040	0.117	0.000	0.107
y744n89	0.011	0.106	0.000	0.098
y744n102	0.029	0.112	0.000	0.104
y744n52	0.079	0.087	0.000	0.082
y744n40	0.029	0.149	0.000	0.141
y744n96	0.018	0.125	0.000	0.119
y744n90	0.013	0.104	0.000	0.100
y744n97	0.016	0.076	0.000	0.072
y744n94	0.030	0.097	0.000	0.094
y744n101	0.031	0.137	0.000	0.132
y744n41	0.021	0.143	0.000	0.139
y744n88	0.022	0.114	0.000	0.110
y744n51	0.031	0.095	0.000	0.093
y744n73	0.035	0.093	0.000	0.092
y744n98	0.035	0.165	0.000	0.162
y744n32	0.023	0.108	0.004	0.108
y744n72	0.035	0.092	0.000	0.091

Discussion

Normalization of ratio data in microarrays is an essential first step in correct interpretation of two-color gene expression experiments. The normalized ratios originally published by Spellman, et al for this data set were calculated by multiplying the Cy5 signals by a constant in order to achieve a mean of zero for the log ratios of all well-measured spots.

The exponential normalization method presented here corrects for a previously unrecognized artifact in two-color gene expression studies using Cy3 and Cy5 fluorescent dyes. The magnitude of the artifact varies among all the hybridized slides, and therefore the normalization needs to be determined empirically for each slide. For other data sets, the magnitude of the artifact has been enough to distort the ratios by more than a factor of 10. This method improves normalization, particularly for weakly expressed genes, and results in better centering and a tighter distribution for the non-differentially expressed genes. Although there was insufficient replication in the CAMDA data set to demonstrate the effect, other data sets have shown an improved precision in normalized log ratios using this normalization technique.

The results obtained using exponential normalization with these data sets confirm our experience with this new normalization method in other 2-color gene expression systems:

- Brings average log ratio for all spots to within 0 +/- 0.01 (ratio within 1.00+/- 0.03)
- Makes normalized log ratio independent of either Cy3 or Cy5 signal
- Tightens up overall distribution of log ratios, as measured by SD of all log ratios
- The magnitude of the effect is greater for weakly expressed genes
- Improvement in the accuracy of ratios calculated for known spiked mRNAs (data not shown)

The magnitude of the effect varies from slide to slide, with the maximum magnitude of the difference between the two methods varying from 1.2-fold to 2.8-fold in these data sets

One advantage of this approach is that the normalization is based on all the data, not just a single control gene or a series of external spikes. A possible limitation of this approach is that it may not work as well with highly selected gene sets, or if controls make up a large portion of the data.

The molecular mechanism for this distortion in the ratios is not clear. It may be due to differences in the physicochemical properties of the cDNAs labeled with the two dyes. Another mechanism which has been proposed is that the distortion is the result of non-equilibrium kinetics and mis-matched molar concentrations of the two labeled cDNAs.

Summary of Exponential Normalization Method

The method employed involves the following steps:

1. Calculate log ratio for all spots (excluding control, empty, and those flagged for poor visual quality or failed PCR);
2. Calculate exponential decay fit to log ratio (y) vs. log(Cy5 signal) (x) according to the equaton y = a + b*exp(-cx); and
3. Calculate a normalized log ratio for each spot by subtracting the fitted log ratio based on step 2 from the observed log ratio for that spot.

Conclusions

An exponential fitted normalization provides improved results in two-color gene expression analysis.

This normalization method is available in Microarray ScoreCard™, a product combining control spotting samples, mRNA spikes, and analysis software for verifying the accuracy of gene expression experiments.

Contact: Thomas M. Houts, Molecular Dynamics, 928 E. Arques Avenue, Sunnyvale, CA 94085
tom.houts@am.apbiotech.com

This poster was presented at CAMDA 2000, "Critical Assessment of Techniques for Microarray Data Mining"
Duke University Medical Center, December 18-19, 2000

[1] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998 Dec;9(12):3273-97

molecular dynamics
part of Amersham Pharmacia Biotech