

Is log ratio a good value for identifying differential expressed genes in microarray experiments?

Alfred Ultsch

Databionics Reseach Laboratory, Department of Computer Science

University of Marburg, D-35032 Marburg, Germany

ultsch@informatik.uni-marburg.de

Motivation: This paper investigates the log ratio calculations for DNA array experiments using two color dyes. It is shown that the calculation of logarithms of the ratio of the two color intensities (LogRatio) has several disadvantages. The effects of numerical instabilities and rounding errors are demonstrated on published data. As an alternative to log ratio calculation, relative differences (RelDiff) are proposed. The stability against numerical and rounding errors of RelDiffs are demonstrated to be much better than for LogRatios. RelDiff values are linear proportional to LogRatios for the range where genes are not differentially expressed. Relative differences map differential expression to a finite range. For most subsequent analysis this is a big advantage, in particular for the search of similar expression patterns. A compensation for variance is difficult for LogRatios and much easier for RelDiff.

Results: This paper shows that RelDiff is a much better value for DNA microarray analysis than log ratio calculation.

Contact: ultsch@informatik.uni-marburg.de

Supplementary information: <http://www.informatik.uni-marburg.de/~databionics>

1 Introduction

In complementary DNA (cDNA) microarray experiments the data for each gene (spot) of are two fluorescence intensity measurements (Quackenbusch 2001, Dudoit 2002). The measurements are produced by a mixture of two portions of mRNA labeled with two different fluorescent color dyes. One portion of the mRNA is labeled by the dye Cy5 producing a red fluorescence color (R), the other is marked by the dye Cy3 producing a green fluorescence color (G). The predominance of one of the colors indicates the relative abundance of the corresponding DNA sequence. This means the over expression of a particular gene. Equal intensities in red and green fluorescence at a spot indicate no particular over- or

under expression of the corresponding gene. In most publications on microarray data the (binary) logarithm of the ratio R/G (LogRatio) is used (Quackenbush 2001). In this paper we argue that this is not a particular good choice for several reasons: LogRatios are numerically instable when R and G values are small and almost equal. Rounding errors are important and negative values are not tolerated. The range of LogRatio is unbound. This leads to arbitrarily big values which disturb subsequent calculations, for example, the estimation of variances. In this work a better value for the indication of over- or under expression of a gene, the relative difference (RelDiff) of R and G is proposed. The properties of RelDiff in comparison to LogRatio are evaluated.

2 Data Sets and basic Definitions

One of the cDNA data sets used consists of microarray experiments of *Saccharomyces Cerevisiae* (DeRisi 1997). The focus of the experiment is the shift from anaerobic (fermentation) to aerobic (respiration) metabolism. At 7 time points during this diauxic shift the expression of 6153 genes is measured. We call this data the “DiauxicShift” data. The data is publicly available from the website <http://cmgm.stanford.edu/pbrown/explore/index.html>. Another data set was published by Eisen (Eisen et al. 1998). The data consists of a set of 2465 gene expressions of yeast in 79 different experiments. The data is available from the website <http://www-genome.stanford.edu>. We refer to this data set as the “Yeast” data.

Let R denote the intensity measurement of the red fluorescent dye, G denote the intensity measurement of the green fluorescent dye. R and G values are typically positive values calculated as the pixel intensities inside the spot R^* (G^*) minus the background luminescence R_b (G_b): $R = R^* - R_b$; $G = G^* - G_b$. If spot intensities are smaller than the surrounding intensities R and G may become negative. The distribution of R and G is typically severely skewed. I.e. there are many small values and few very big values. The values for R and G , for example in the DiauxicShift data, range from 50 to 50.000. An impression of the probability density distribution of the R and G data is given by the PDEplot in Figure 1.

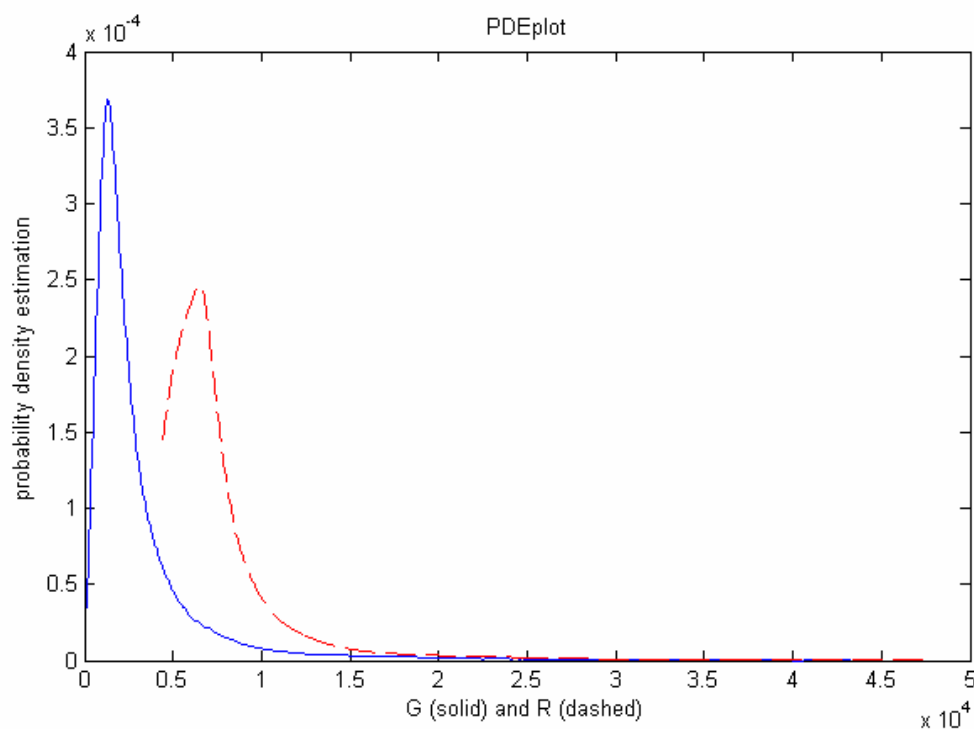


Figure 1: PDEplot of R and G of the DiauxicShift data

PDEplots are empirical probability density estimations based on information optimal sets. These are optimal density estimations for DNA microarray data. For details and availability of software for PDEplots see (Ultsch 2003) respectively <http://www.informatik.uni-marburg.de/~databionics>.

The LogRatio, used in most two color microarray publications is defined as

$$\text{LogRatio}(R, G) = \text{ld}\left(\frac{R}{G}\right)$$

where ld is the logarithm for basis two.

If R is equal to G, then LogRatio equals 0. If LogRatio is greater than 0 (less than 0), then R is greater than G (R less than G). The distribution of LogRatios is centered around zero, but has a substantial number of values which are greater than one in absolute value. In the PDEplot of Figure 2 the distribution of all 194.893 values of the Yeast data can be compared with a Gaussian normal distribution $N(m, s)$, where m is the mean and s^2 the variance of the data (empirical Gauss).

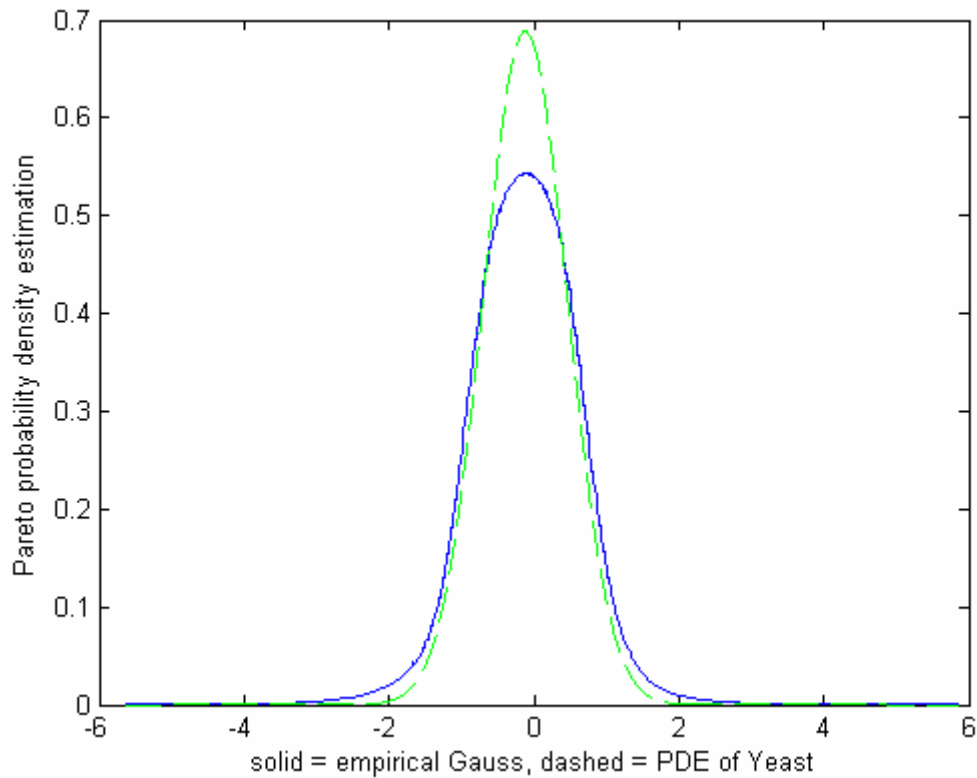


Figure 2: Distribution of LogRatio values of the Yeast data

In the Yeast data set 2800 values i.e. 1.2% of LogRatios are greater than ± 2 , 113 values i.e. 0.06% are greater than ± 4 .

In this paper we propose another value for the indication of differentially expressed genes, the relative difference (RelDiff).

The relative difference is the ratio of the difference (R-G) to the mean intensity of the spot. RelDiff is defined as follows:

$$\text{RelDiff}(R, G) = \frac{R - G}{\frac{1}{2}(R + G)} = 2 \frac{R - G}{R + G}$$

If R is equal to G, then RelDiff equals 0. If RelDiff is greater than 0 (less than 0), then R is greater than G (R less than G).

The relative difference may also be measured in percent. This leads to the definition:

$$\text{RelDiff}\%(R, G) = \frac{R - G}{R + G} \cdot 200 \quad [\%]$$

3 Interpretability

The direct interpretation of LogRatio values is difficult except for powers of two. It is easy to see that a LogRatio of 1 means that the expression level of a particular gene is two fold. It is, however, not straight forward to see that a LogRatio value of 1.58 corresponds to an threefold, a LogRatio value of 3.322 to a ten fold over expression rate. For an interpretation of such LogRatio values one must be familiar with dual logarithms. The numerical values of RelDiff and in particular RelDiff% have a straight forward interpretation. Even an “odd” value of, for example 22.12%, for RelDiff has a direct interpretation. Such a value means that there is 22.12 percent more red color in the average luminosity of the particular spot.

4 Numerical stability

In many two color experiments almost all of the thousands of measured genes have an identical level of R and G. In order to investigate the numerical properties of the formulas above we assume that R is equal to G plus some small measurement error ε . I.e. we assume $R = G + \varepsilon$ for some small error ε . For LogRatio we obtain :

$$\text{LogRatio}(G + \varepsilon, G) = \text{ld}\left(\frac{G + \varepsilon}{G}\right) = \text{ld}\left(1 + \frac{\varepsilon}{G}\right)$$

This term results in very large negative values the closer ε gets to $-G$. For small values of G this might be the case. This means that there might be numerical instable LogRatio calculations. Furthermore for G close to zero the LogRatio may become very big.

For RelDiff on the other hand we obtain:

$$\text{RelDiff}(G + \varepsilon, G) = 2 \frac{G + \varepsilon - G}{G + \varepsilon + G} = \frac{2\varepsilon}{2G + \varepsilon} = \frac{\varepsilon}{G + 0.5\varepsilon}$$

In this case ε must become as big as $-2G$ in order to cause numerical problems. It can be concluded that RelDiff is twice as numerically stable as LogRatio. For G approaching zero, the RelDiff values approach 2. This means that the error for RelDiff is bound:

$$|\text{RelDiff}(G + \varepsilon, G)| \leq 2.$$

For the DiauxicShift data we have observed the numerical properties empirically. R was set to $G + \text{EPS}$. The measurement error ε (EPS) was varied in the interval $[0, 10\%]$ relative to the maximal value ($\max(G)$) occurring in G. Figure 3 shows the resulting LogRatio and RelDiff values.

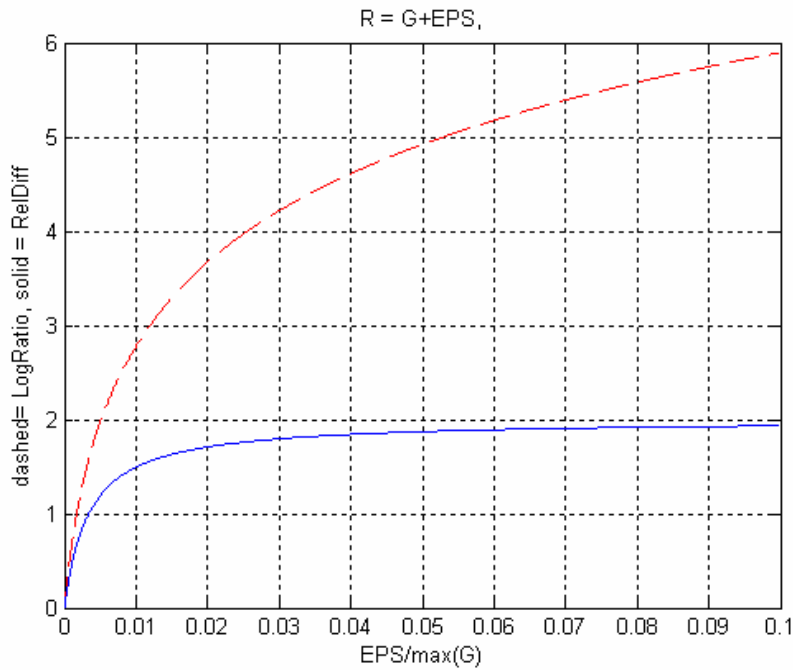


Figure3 : LogRatios and RelDiff values for R=G+EPS in the DiauxicShift data

The erroneous RelDiff values are consistently lower than the LogRatio errors. For all ϵ values RelDiff was more than 40% lower than LogRatio. For $\epsilon / \max(G) > 1\%$ RelDiff was more than 40% lower. For $\epsilon / \max(G) > 4\%$ RelDiff was more than 100% better. Figure 3 also shows that the RelDiff errors are bound by 2 while the LogRatio error may become arbitrarily big.

In many two dye microarray experiments most measured intensities are very small (see Figure 1). To analyze the numerical situation in this case we assume $R = G + \epsilon$ and $G \approx \epsilon$. This gives:

$$\text{LogRatio}(\epsilon + \epsilon, \epsilon) = \text{ld}\left(\frac{\epsilon + \epsilon}{\epsilon}\right) = \text{ld}(2\epsilon) - \text{ld}(\epsilon)$$

This is numerical instable! For ϵ close to zero LogRatio values explode. For RelDiff on the other hand we get:

$$\text{RelDiff}(\epsilon + \epsilon, \epsilon) = 2 \frac{\epsilon}{\epsilon + \epsilon + \epsilon}$$

Even if ϵ gets very close to zero, the denominator of the fraction in this term is always three times the numerator. This means:

$$\text{RelDiff}(\epsilon + \epsilon, \epsilon) \cong \frac{2}{3} = 0.67$$

The clear conclusion is that log ratios are numerically unstable in particular for small intensities with equal values in the red and green intensities. Relative differences on the other hand are numerically stable with a maximum error value of 0.67, if both intensities are small and practically equal.

5 Rounding Errors

The Yeast data is published as LogRatios with a precision of two digits after the decimal point. We examine the errors with respect to rounding by the exact calculation of LogRatio and RelDiff values, then round the values to a precision of one hundredth. From these rounded values the original value of G is reconstructed using the correct value of R. This gives a value G'. The difference between the value G' and the true value of G leads to the relative error (err) measured in percent of the true value. I.e.:

$$\text{err} = \frac{G' - G}{G} \cdot 100 \quad [\%]$$

Figure 4 shows the relative error for the reconstructed values of G in the diauxic shift data. The left side of Figure 4 shows the error for LogRatio, the right side for rounded RelDiff.

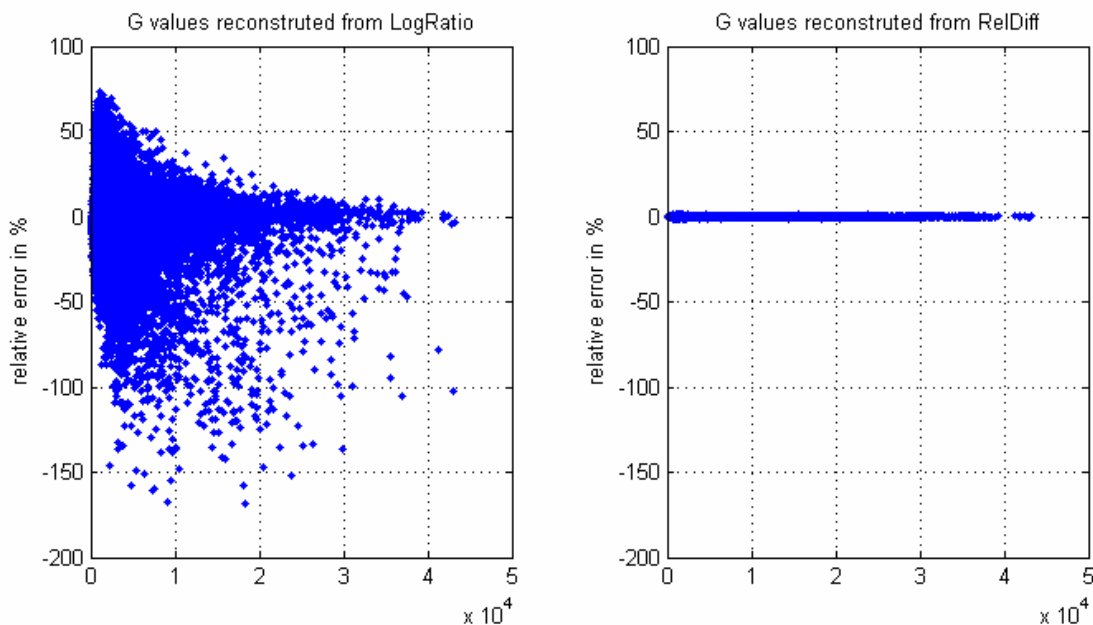


Figure 4: Errors produced by rounding values to two decimal digits

It can be seen that the error level may become very extreme for LogRatios. The conclusion is that rounding LogRatios to is considerable more critical than a rounding of RelDiff. To round LogRatios to two decimal digits, as done in Eisen et al's data is critical and may influence severely subsequent calculations. The same rounding has much less effects for RelDiff data. A very natural precision setting for the RelDiff data is percentage values rounded to two decimal digits. In this case there are five meaningful digits within the range of -200.00 to 200.00. The precision of such values is enough for subsequent calculations and can be directly understood (see Chapter 3).

6 Negative and zero values

In some two color microarray experiments the raw values measured for the intensities are close to the background values. Sometimes there is even more background intensity encountered than intensity inside a spot. The difference between spot measurements R^* (G^*) and the background luminescence R_b (G_b) become in these cases negative. For such cases the logarithm is undefined. This may lead to unwanted numerical errors or imaginary results in LogRatio. For RelDiff negative values are no problem. If one of the intensities is zero, this causes a numerical error for LogRatios. If R is zero, the logarithm gets arbitrarily big. If G is zero, the denominator of R/G causes an error. For RelDiff zero values in R or G are uncritical. The result is a meaningful value of RelDiff. If both intensities are zero, RelDiff is undefined. In this case, however, no intensity at all is measured in both colors. This case can be treated properly either by ignoring the measurement or setting the resulting RelDiff value to zero.

The difficulties in the calculation of zero or negative logarithms might be the reasons for the 3760 undefined values in the published Yeast data set.

7 In the Limit

We will now investigate the properties of LogRatio and RelDiff for very large differences in color intensities. Let $R \gg G$ such that $R \pm G \cong R$ resp. $\text{ld}(R) \pm \text{ld}(G) \cong \text{ld}(R)$. We obtain

$$\text{LogRatio}(R, G) = \text{ld}\left(\frac{R}{G}\right) = \text{ld}(R) - \text{ld}(G) \rightarrow \text{ld}(R) ;$$

This means in particular, that there is no theoretical limit to the LogRatio values. Under the same assumptions for RelDiff holds:

$$\text{RelDiff}(R, G) = 2 \frac{R - G}{R + G} \rightarrow 2 \frac{R}{R} = 2$$

In the same manner for $G \gg R$ Log Ratio goes to $-\ln(G)$ and RelDiff approaches -2. This means that RelDiff has a limited value range.

Many two color DNA microarray experiments search for some similar expression of genes. For this a similarity measure is defined. Typical measures are Euclidian distances, correlation measures, Mahalanobis distances and others. Since the expression data of several experiments have to be compared, the variance of the data have to be taken into account. The limited scope of the RelDiff values is advantageous in this case since it limits the influence of very outlying values.

In the DiauxicShift data, for example, the range in LogRatios varies by a factor up to 3.3 for the different times the experiment was performed during the diauxic shift. On the other hand the maximal ratio of the ranges of RelDiff is only 2.3, i.e. 45% less. In Figure 5 the variances of LogRatio and RelDiff values of the microarrays measured at different time points during the diauxic shift can be compared. In Figure 5 the variances are normalized such that the smallest variance is 1.

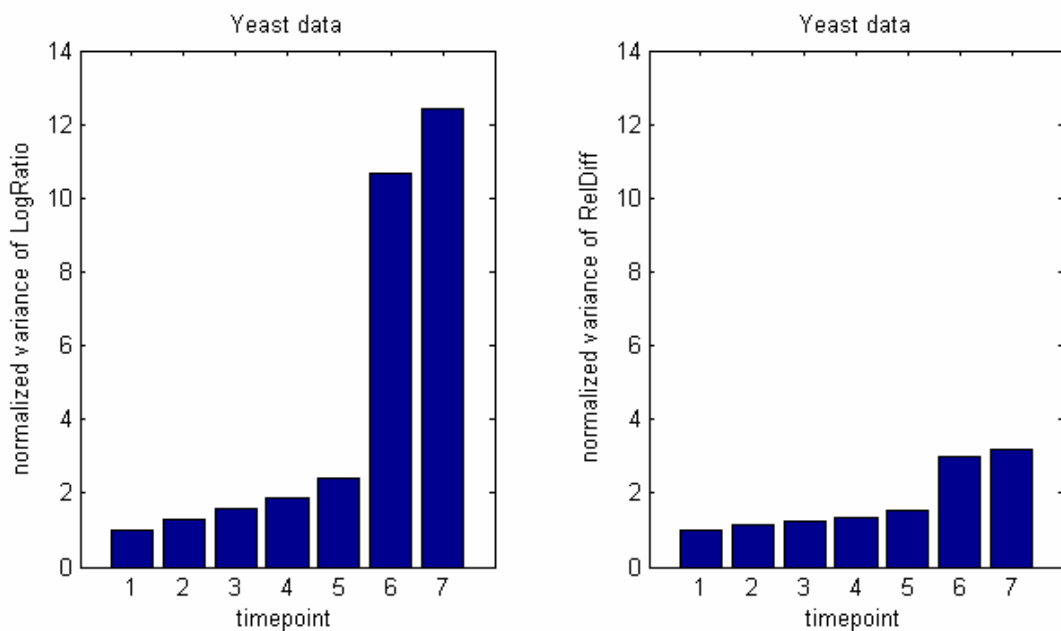


Figure 5: Comparison of the variances of LogRatio and RelDiff in the DiauxicShift data

Many authors use Euclidian distances to find similar expression patterns, for example (Torkkola et al 2001, Kaski et al 2001, Ultsch 2002). If the different variances are not compensated, the Euclidian distance is mainly determined by the experiments with largest variance. To normalize by the empirical variance is, however, a big problem, since the LogRatio values are not normally distributed. The “fat tails” of the distributions (see Figure 2) invalidate simple variance calculations. The same holds for many other distance calculations, e.g. for the correlation distances used in (DeRisi et al 1997).

For RelDiff values this normalization problem is alleviated due to the natural limitation of the range. From Figure 5 it can be concluded, that for RelDiff a compensation for variance is less critical. In this experimental setting the variances differ by a factor of 3 in RelDiff. Log Ratio variances differ by factor of more than 12 in the same experiment!

8 Proportionality

The relationship between LogRatio values and RelDiff values is considered. Within pen point size the LogRatio values are directly proportional in the range of $[-2;2]$. This can be seen in Figure 6 for the DiauxicShift data. This proportionality may be. Analytically the same can be deduced from approximations for logarithms.

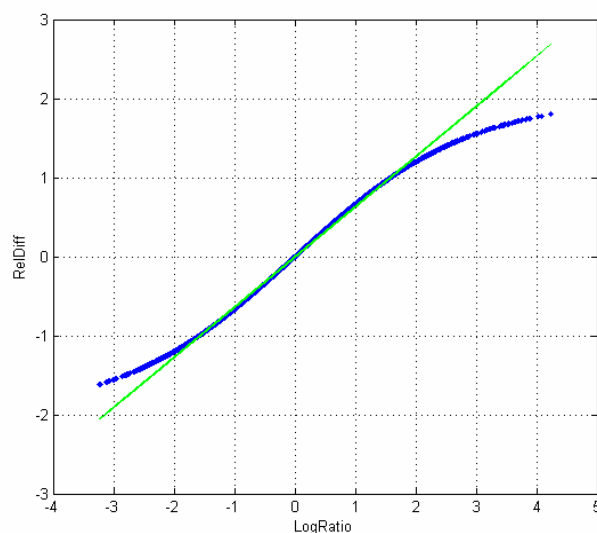


Figure 6: Proportionality of RelDiff and Log Ratio for DiauxicShift data.

Within a range of 1/4 to 4 fold of gene expression the values of LogRatio are practically proportional to RelDiff values. Outside of this range the RelDiff values become over proportionally smaller in absolute values. This means that all subsequent calculations that rely on the relative differences of LogRatio values are still possible with RelDiff values.

Differences in the analysis can only be expected, if the absolute values of extremely over - or under expression are important. The limitation to the range $[-2, 2]$ is mostly advantageous. In many publications gene expressions are depicted as color coded bars. For RelDiff data it is possible to assign a definite range. Extreme values are fixed to ± 2 for RelDiff. From the theoretical considerations above, it is also possible assign the range $[-0.67, 0.67]$ to a color indicating "hardly a differential expressed gene". A de facto convention for this color seems to be yellow (Eisen et al 2001, Quackenbush 2000).

9 Discussion

This paper investigates the log ratio calculations for DNA array experiments using two color dyes. The calculation of a logarithm of the ratio of the two color intensities has several disadvantages. The numerical stability is questionable for measurement errors and for small intensities. It has also been shown that rounding errors are important for LogRatio calculations. Both effects have been demonstrated to play a critical role for published data. As an alternative to LogRatios the calculation of relative differences (RelDiff) is proposed. These values are directly proportional to LogRatio values when the two color intensities are about identical. In contrast to LogRatio, however RelDiff values are bound. This is a big advantage if similarities on genes are calculated. A compensation for variance is difficult for LogRatios, and much easier for RelDiff. The numerical stability for RelDiff is much better than for LogRatio. For small and almost equal intensities in both colors the error of RelDiff is finite and can be calculated precisely. The error for LogRatio may, however, become arbitrarily big for such measurements. Rounding has a much smaller influence on the precision of the obtained values for RelDiff than for LogRatios. Finally the values of RelDiff have a direct interpretation while LogRatio values require the knowledge of binary logarithms. In summary this paper shows that RelDiff is much better for DNA microarray analysis than log ratio calculation.

10 References

- DeRisi et al., (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* 278, pp 680-686.
- Dudoit S. et al. (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data, *Journal of the American Statistical Association*, 97, 447, pp 77—87
- Eisen M. B. et al. (1998) Cluster analysis and display of genome-wide expression patterns, *PNAS*, vol95 pp14863-14868
- Kaski, S. et al. (2001) Analysis and visualization of gene expression data using self-organizing maps. *Proc Workshop on Nonlinear Signal and Image Processing.* Baltimore, Maryland, June 3-6.
- Quackenbush, J. (2001) Computational Analysis of Microarray Data. *Nat. Rev. Genet.* 2, pp. 418 -427.
- Torkkola K. et al (2001), Self-organizing maps in mining gene expression data, *Information Sciences I* 39, pp 79- 96.
- Ultsch, A. (2003) Maps for the visualization of high-dimensional data spaces, *Proc. Conf. on Self Organizing Maps*, Kyushu, Japan, pp 225 – 230.
- Ultsch, A., Eilers, M.: (2002) Databionic methods for the diagnosis of neuroblastoma tumors, *Cooperation in research and innovation*, Wiesbaden, Germany, pp 19 - 20.