A Statistical Analysis of Radiolabeled Gene Expression Data

Rafael A. Irizarry[*], Giovanni Parmigiani[†][*], Mingzhou Guo[‡], Tatiana Dracheva[§], and Jin Jen[§]

March 29, 2001

Abstract

This paper considers statistical issues in the analysis of a designed experiment to investigate differential gene expression in colon cancer and normal colon tissue. In this experiment gene expression is measured using radiolabeling–based array filters. Specific statistical issues arise in connection with radiolabeling technology, because of the absence of direct control, which are replaced by empty spots on the filter, and with designed experiments, because of the opportunity to systematically quantify important sources of random variation. Here we consider three aspects in detail: normalization of expression intensities; shrinkage estimates of intensity ratios between cancer and normal tissue; and ranking of genes by the strength of the evidence that they are differentially expressed. We propose a robust and simple–to–implement procedures for normalization and shrinkage, that addresses in a technology–specific way the problem of estimating ratios in presence of small and noisy denominators. We also discuss a graphical display to rank genes using a metric based on quantiles of a null distribution obtained by replicating the array experiment in normal tissue.

1 Introduction

Radiolabeling based gene expression measurements are useful for cancer research because they can be carried out using small amounts of biological materials. Statistical issues are different from fluorescence expression data, because radiolabeling gives absolute intensities that reflect gene expression and there is no internal control. See Ramaswami et. al. (1999) for more details.

The data-set described in this paper was obtained to identify genes that may be associated with lung cancer. Lung cancer tissue was obtained from 5 subjects. Normal tissues from the same type of cells was obtained from those same 5 subjects. From each of these tissues 2 samples were prepared using 2 different isotopic batches. Each of these 4 samples were hybridized with a filter spotted with cDNA from many genes in a 48 × 24 grid. We refer to these spotted filters as arrays. Each of these arrays were scanned to produce an image file which was then analyzed with specialized software that produced an intensity level for each grid point or *spot* on the array. Another set of four arrays was constructed for another subject in the same way except that both samples came from normal tissue. As we will describe later, these intensities will be useful to create reference distributions, and we will call these the *reference arrays.*

Not all the values read from the arrays are associated with genes. There were 207 spots where no cDNA was spotted. They were left empty. Because there is *non-specific* binding between the samples and the filters, positive values are obtained from these empty spots. The intensities read from these empty spots provide direct evidence about measurement error associated with the system. Spots associated with genes that are not expressed will also have intensities due to non-specific binding. Furthermore, there are 8 spots that have the same "control" gene.

Histograms of the raw intensities of empty spots for the 4 reference arrays suggest a strong batch effect and a relatively strong filter effect. Also, the ranges of the histograms are constant in a log scale, suggesting that we should consider log intensities.

We use x to denote log intensity levels from the arrays containing normal tissue and y to denote the log intensity levels from arrays containing the tumor tissue. Specifically we use

$$x_{g,i,j} \text{ and } y_{g,i,j},\ g = 1, \ldots, G,\ i = 0, 1, \ldots, I, \text{ and } j = 1, 2$$

[*] Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205, rafa@jhu.edu

[†] Department of Oncology, Johns Hopkins University

[‡] Department of Otolaryngology, School of Medicine, Johns Hopkins University

[§] National Cancer Institute

with $G = 1152$ the number of spots on the array, $I = 5$ the number of subjects and $i = 0$ representing the reference arrays, and j indexing the two replicates coming from the 2 different batches. We also denote as $\mathcal{E}$ the subset of $\{1, \dots, 1152\}$ containing the indices representing the spots left empty. Similarly, we denote $\mathcal{C}$ the subset representing the control genes. We denote the rest of the indices as $\mathcal{G}$.

The two questions we try to answer through the analyses presented in this paper are 1) can we rank genes by differential expression between cancer and normal tissues in each subject? 2) can we rank genes by differential expression in all subjects?

To answer the first question a naive approach is to look for the g giving the highest log ratios $z_{g,i,1} = y_{g,i,1} - x_{g,i,1}$ and/or $z_{g,i,2} = y_{g,i,2} - x_{g,i,2}$. We would do this for each i. However, looking at the data suggests there are problems with this approach.

One problem is filter/batch effects. The analysis of variance (ANOVA), described in the next section, shows that there is a strong batch effect and a relatively strong filter effect. This means that if a particular subject happens to be assigned to a highly expressed filter for the cancer tissue then the zs will be big for all genes. This would make it hard to interpret values of z across subjects. Even after removing filter effects, we still must ask ourselves: What value do the zs have to reach for us to say it is “differentially expressed”? There will always be a highest value!

Another problem is small denominators. Many of the genes are not expressed in the tissues we are studying. Because there is non-specific binding some value for intensity will be read. The data seems to suggest that for genes that are not expressed we could have that the intensities in the normal tissue are close to 0 by chance making z artificially high.

In the remainder of this paper we offer relatively straight forward solutions that we have found to work well at answering the above questions and that have motivated development of statistical methods.

2 Normalizing the data

Normalizing microarray data is a subject of current interest, see for example Yang et. al (2001). ANOVA has been used as a first step in understanding microarray data by Kerr and Churchill (2000). We can think of various things that could have an effect on intensity level, for example batch (replicate), filter, subject, genes, and cancer status. Table 1 shows the results of an ANOVA. Let A denote gene effect, B denote the cancer effect, C the subject effect, and D the batch effect.

	Genes				Empties				Control			
	Comparisons		Reference		Comparisons		Reference		Comparisons		Reference	
Effect	DF	$\sqrt{\text{MS}}$	DF	$\sqrt{\text{MS}}$	DF	$\sqrt{\text{MS}}$	DF	$\sqrt{\text{MS}}$	DF	$\sqrt{\text{MS}}$	DF	$\sqrt{\text{MS}}$
A	911	7.240	911	2.87	206	2.01	206	0.693	8	2.15	8	0.958
B	1	18.70	1	17.70	1	12.70	1	14.300	1	2.09	1	0.534
C	4	18.700	NA	NA	4	7.26	NA	NA	4	1.37	NA	NA
D	1	149.000	1	117.000	1	76.60	1	60.50	1	19.60	1	13.90
AB	911	0.615	911	0.249	206	0.608	206	0.416	8	0.144	8	0.196
AC	3644	0.379	NA	NA	824	0.537	NA	NA	32	0.245	NA	NA
AD	911	0.967	911	0.699	206	0.59	206	0.421	8	0.304	8	0.304
BC	4	15.000	NA	NA	4	7.70	NA	NA	4	1.79	NA	NA
BD	1	4.30	1	18.90	1	1.99	1	2.03	1	0.797	1	1.76
CD	4	45.00	NA	NA	4	17.80	NA	NA	4	4.61	0	NA
BCD	4	6.52	NA	NA	4	1.97	NA	NA	4	0.784	0	NA
Total	18240	2.150	3648	2.480	4140	1.520	828	2.21	180	1.730	36	2.38

Most of the variability in the data is from the batch. Because each cancer and normal comparison comes from the same batch, this effect cancels out when computing the zs. There is also a substantial subject heterogeneity as well as an overall effect of cancer, as expected. The variability explained by the AC interaction is small, indicating that in this technology the genetic signal is not the primary source of variation, and that identification of subtle expression changes is difficult and unlikely to be reliably carried out for a large set of genes.

The BCD interaction in the comparisons and BD for the reference, which represent the filter effect not accounted for by the subject and batch effect, shows a relatively strong effect. This effect does not cancel out when computing the zs so to be able to interpret the ratios we need to remove it.

A common procedure in microarray data analysis is to simply normalize the filters by subtracting the mean of each filter from each value, i.e. consider $y_{g,i,j}^{(normalized)} = y_{g,i,j} - \bar{y}_{\cdot,i,j}$ and similarly for the xs. The danger with doing this is that many of the genes spotted on the arrays are usually selected because researchers consider them likely to be over-expressed. This means that the mean of the ys should be larger than the xs and this difference in mean is confounded with the difference in filter effect. By subtracting means we would be subtracting out some of the differential expression between cancer and normal tissues.

In Figure 1 we plot the ratio of the intensities vs. the product of the intensities in a log scale, i.e. $y - x$ vs. $x + y$, for the two replicates of subject 1. Notice that the *filter effect* seems to change with the total intensity of a particular spot.

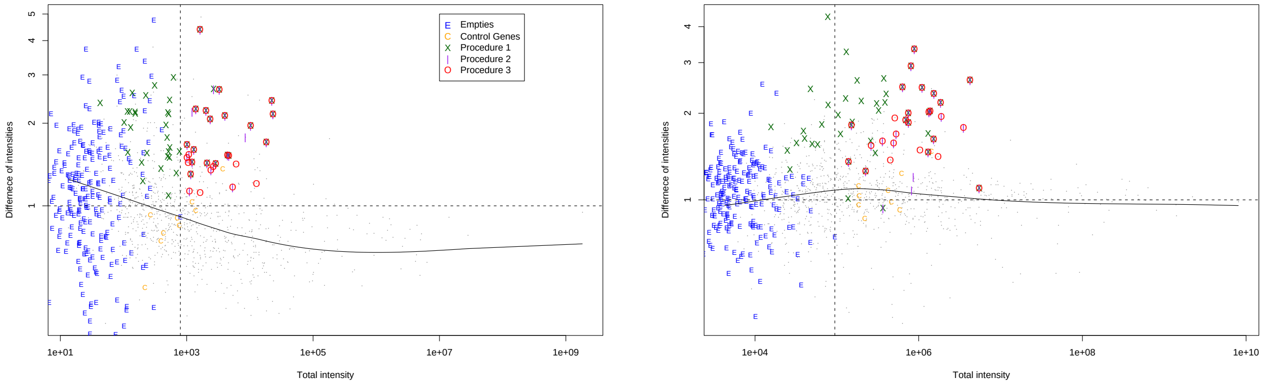

Figure 1: Ratios of intensities plotted versus total intensity in a log scale. The vertical line represents the highest total intensities among the empty spots. The horizontal line represents a ratio of 1.

For this reason using medians or trimmed means to remove the filter effect is not a good solution. If we model x and y as random variables then we have that the expected filter effect depends on the total intensity, i.e. $\mathrm{E}(y - x|x + y)$ is not constant. This arises because specific binding and non-specific binding are two different natural processes. Because we have no way of knowing which points represent non-specific binding and which represent specific binding we cannot normalize by just estimating two means. Rather, we estimate $\mathrm{E}(y - x|y + x)$ using a nonparametric regression approach, in this case Clevaland's (1979) loess. It is critical to use a robust loess, so that large differences do not affect the fit too much. We can then obtain a normalized difference and redefine $z_{g,i,j} = y_{g,i,j} - x_{g,i,j} - \hat{\mathrm{E}}_{i,j}(y_{g,i,j} - x_{g,i,j}|x_{g,i,j} + y_{g,i,j})$ for $g \in \mathcal{G}$, $i = 0, \dots, 5$, and $j = 1, 2$.

3 What is a large difference?

We will refer to a *score* as a value assigned to each gene which is interpreted as *evidence* of over expression. An example of a score is the average of the zs defined above over the two batches, that is $\bar{z}_{g,i,\cdot} = (z_{g,i,1} + z_{g,i,2})/2$. We may then suggest that genes with "large" values of $\bar{z}$ are differentially expressed. But how do we define "large"? The reference arrays provide a way of constructing a reference distribution for the "typical average difference between two samples when there is no gene effect", specifically we construct an empirical null cumulative distribution $\hat{F}(z)$ using $\bar{z}_{g,0,\cdot}$'s. We can then assign a value, in probability scale, to each $\bar{z}_{g,i,\cdot}$ by looking for the tail probability in the null distribution. In Figure 1 we denote with an X points related to genes with $\hat{F}(\bar{z}_{g,1,\cdot}) = 1$. We call this procedure 1 in the legend of Figure 1.

In Figure 1 we see the that for some genes, total raw intensities where lower that the total raw intensities of the measurements taken from the empty spots. We denote this set of genes with with $\mathcal{N}_{i,j} \equiv \{g \in \mathcal{G} : x_{g,i,j} + y_{g,i,j} \leq \max_{g \in \mathcal{E}}(x_{g,i,j} + y_{g,i,j})\}$. These genes are unlikely to be expressed. However, notice that many of the large ratios are obtained for genes that are in this set. This is because these have very small xs. A first attempt at correcting this is taking an average, as done by the $\hat{z}$s. However, even after taking this average we see that many of the Xs in Figure 1 are associated with genes in $\mathcal{N}$ or to the left to the vertical line. Now we propose two simple, alternative procedures to address this problem.

The second is simply to remove all genes in either $\mathcal{N}_{i,1}$ or $\mathcal{N}_{i,2}$ for each subject i. We do this for each subject separately, since different subjects can have different expressed genes. We then define $e_{g,i} = 0$ if $g \in \mathcal{N}_{i,1}$ or $g \in \mathcal{N}_{i,2}$ and equal to 1 otherwise and consider $e_{g,i}\bar{z}_{g,i}$ as the score. We perform this procedure on the control array data to obtain a null distribution and obtain scores in probability scale. In Figure 1, we denote with | points related to genes that received a score of 1 under this procedure. Notice we still see many genes "close" to the vertical line. This is probably due to the fact that some genes that are not expressed are randomly falling barely to the right of the vertical line. It seems that further shrinkage is needed.

The third procedure is similar to the first, except we multiply the average by $s_{g,i}$, a shrinkage coefficient inversely proportional to the empirical signal to noise ratio. Shrinking expression ratios has been found to be useful in improving the understanding of gene expression levels, see Newton et. al (2001). Our score is now $s_{g,i}e_{g,i}\bar{z}_{g,i}$. Again we form a null distribution with the reference genes and obtain a scores in probability scale. In Figure 1, we denote with O the points related to genes that had a score of 1. Notice that this procedure eliminates some of the genes chosen by procedure 2 that are close to the vertical line. It also includes some genes that were consistently large and not included by procedure 2. However, it still seems like there are two many choices close to the vertical line. Methods for finding shrinkage coefficients is subject of current work.

4 Genes differentiable expressed in all subjects

To construct a total score for all subjects we simply add the scores in probability scale of each subject separately giving a score of 0 to genes that were removed for not being expressed. We can now rank all the genes by this score. But what is a “large” score? We can no longer construct a reference distribution from the reference arrays because we only have one subject. Obtaining control arrays for 5 subjects would almost double the cost of the experiment.

To get a sense for what is a large score we use a bootstrap procedure (Efron (1979)) in the following way. Under the null hypothesis that no gene is differentially expressed, and that the difference in expression is independent from subject to subject then the total score should not change if for each replicate of each subject we re-shuffle the genes and perform the entire procedure again. We construct 100 bootstrap total scores by randomly re-shuffling the pairs of raw intensity levels of each subject and applying the procedure to the re-shuffled data. In Figure 2, we plot the score obtained with the original data as a solid black line and the 100 bootstrap scores as yellow lines. Notice that the largest 11 scores are larger that the maximum of the 100 bootstrap maximum scores. This is strong evidence against the null hypothesis is not true. We can compare the k-th highest score with the bootstrap average k-th highest score to assess how “significant” it is. However, it could be the independence assumption that is not true. Constructing a bootstrap score that take correlation into account is subject of current research.

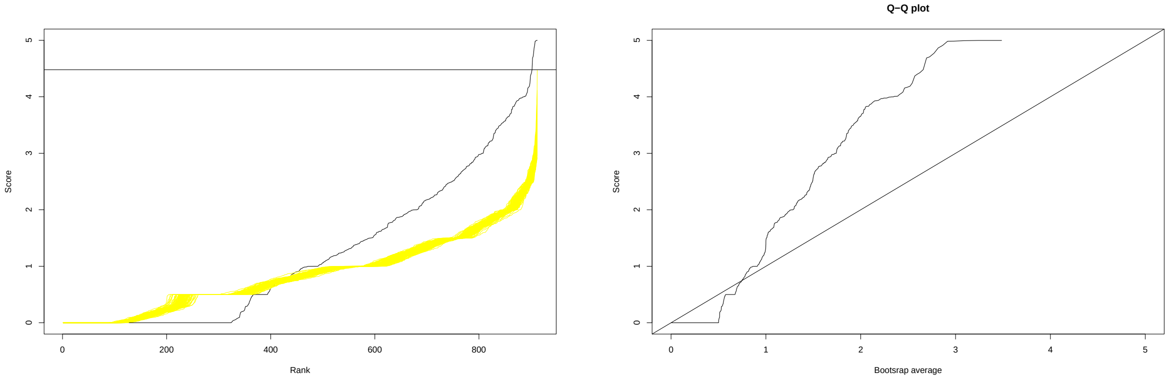

Figure 2: The black line represents the ranked total scores for all genes. The yellow lines are bootstrap total scores obtained under the null hypothesis of no differential expression.

We have compared our results with those obtained using SAGE libraries and found that the genes ranked high by our procedure show high expression levels as measured by the SAGE counts.

References

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Kerr, K. and Churchill, G. (2000). Experimental design for gene expression microarrays. *Biostatistics*. To appear.

Ramaswami, A., Tihan, T., Bornman, D., Johnston, J., Saltz, J., Weigering, A., Piantadosi, S., and Gabrielson, E. (1999). Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Res*, 59(20):5119–5122.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cdna microarray data. Manuscript in preparation.