

Assessing the Information Content of Microarray Time Series

E. Yang

Rutgers University, USA

I.P. Androulakis

Rutgers University, USA

INTRODUCTION

While the rise of microarrays has heralded a new era in molecular biology with its ability to measure the expression level of thousands of genes at once, the usefulness of microarrays is exigent upon the ability to obtain accurate gene expression data for the individual genes (Bowtell, 1999; Brown & Botstein, 1999; Cheung, Morley, Aguilar, Massimi, Kucherlapati, & Childs, 1999). However, there has been significant criticism as to how meaningful the information derived via microarrays is. In cases where one has attempted to find genes that correlated to types of cancer or survival rate, it was found that different analysis techniques would often times yield radically different set of genes, calling into question the validity of the overall experiment itself (Dupuy & Simon, 2007). It is our contention that part of the problem associated with microarrays is that there does not exist a coherent method for dealing with data quality, and if a coherent method for dealing with data quality existed, many of the criticisms of microarrays could be addressed.

BACKGROUND

“Fishing Expedition” is normally used in a negative connotation in the legal field. The negativity has carried over to the scientific field and describes an experiment in which the researcher does not know precisely what one is looking for. However, the promise of microarrays is the fact that they allow for just such experiments, and coupled with different algorithms allow for a data driven approach to science (Nakai & Vert, 2002). By identifying the possible targets of gene regulation, researchers can then formulate more specific experiments to validate such a hypothesis. While the technology behind microarrays consistently advances in terms of

the density of microarrays as well as the repeatability between each individual microarray, there still exists the major issue of noise. Whilst the technological improvements themselves are able to minimize the technical noise, there still exists significant *biological noise* which for complex multitissue organisms cannot be easily overcome with technology. For instance, despite the standardization of rat/mice lines, there still exists significant variation in any reading taken from a population of animals. Due to this noise, it is difficult to identify the genes which actually respond to a given treatment, and those genes whose fluctuations are due to random noise.

Therefore the most important algorithms for the processing of microarray data are those that select meaningful genes from the thousands that are measured via the microarray. The most common metric used by these algorithms is that of *statistical significance* (Smyth, Yang, & Speed, 2003). There is one caveat with the use of statistical significance primarily in the fact that not all genes that are statistically significance are biologically relevant. The set of biologically relevant genes is dependent wholly upon the biological ground truth, whilst the set of statistically significant genes are dependent upon the number of replicates, the quality of the microarray platform, inherent SNR, as well as biological significance. This is a recognized problem which has been addressed by many researchers as a *feature selection* problem under the assumption that the set of biologically relevant genes ought to be able to work as classifiers between the different states being tested (Wu, 2005).

While not all genes that are statistically significant are biologically relevant and vice versa, there does exist a tendency for biologically relevant genes to be statistically significant as well. Therefore, by selecting statistically significant genes, one increases the likelihood of identifying biologically relevant genes

as well. However for one to have confidence in these initial results, care must be taken in the selection of statistically significant genes, paying special attention to normalization and the setting of statistically significant cutoffs.

STATISTICAL SELECTION OF GENES FROM MICROARRAYS

Normalization

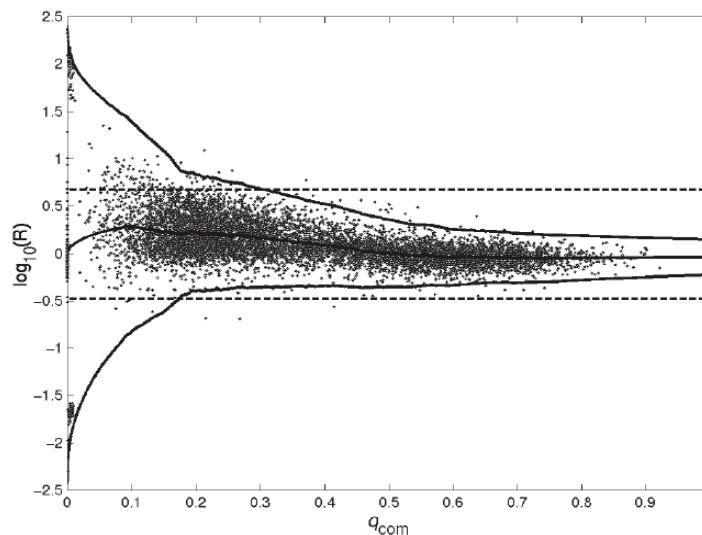
The use of normalization is important because many of the selection algorithms that look for over/under expressed genes in a two state experiment base their assumption off of the fact that the initial distribution log is normalized and compute their confidence intervals according to that distribution. Therefore by transforming the data so it does conform to the log-normal distribution, one is able to use standard statistical tests such as the t-test to ascertain whether or not the variations are due to noise or due to some intrinsic change in the expression level of the gene.

One of the challenges with analyzing microarray data is the problem of translating the recorded intensity level obtained by the detection equipment and determining the true expression value of the given probe. Generally, for genes which show a high intensity value, there exists a good correlation between the intensity between the two

samples follows a roughly linear trend. However, at low intensity levels, the correlation between the two samples deviates from this (see Figure 1). The justification for the majority of genes being linearly correlated is that under most situations, only a small fraction of genes are responding to the overall treatment and even with the addition of noise, they should be consistent over multiple chips in *temporal gene expression* experiments. The small fraction of genes that do deviate from this linear relationship are then the ones that deviate by a given statistically significant level. Techniques such as the LOESS, dCHIP, and PDNN (Cleveland, 1979; Li & Hung Wong, 2001; Millenaar, Okyere, May, van Zanten, Voesenek, Peeters, 2006; Nielsen, Gautier, & Knudsen, 2005), attempt to normalize the data in such a manner in which the correlation between two samples becomes consistent, thereby allowing for easier identification of statistical outliers.

LOESS (Cleveland, 1979) is a local nonparameter method which attempts to fit a low order polynomial, normally linear or quadratic, to the scatter-plot attempting to minimize the random variations in the data. It is most often used for the *normalization* of two dye experiments, primarily to account for the slight difference in affinity between the two dyes at low expression levels, but can be used generally to correct for the nonlinearities found at lower intensity levels. It is similar to a nonlinear regression fit, except it performs a local regression upon blocks of data. The blocks of

Figure 1. The deviation from the log normal distribution at low intensity levels. The LOESS curve centers the distribution and forces log-normality (Wang, Hessner, Wu, Pati, & Ghosh, 2003)



overlapping data used are determined via a smoothing parameters which is set by the user under the constraints given in Equation 1. The size of the data block used for each local fit is then given as nq .

$$\frac{(d+1)}{n} \leq q \leq 1 \quad (1)$$

where d is the degree of the polynomial used.

Since LOESS is a local fitting method, the number of points used in each fit is nq , where q is the selected smoothing factor which satisfies the constraint in Equation 1 and n is the total number of points in the dataset. There is an additional weighting function given in Equation 2 which forces points closer to the point of estimation to contribute more to the local fit than those that are further away. The distance of the points to the point of estimation are calculated, and then normalized, so the data-point at the point of estimation has a value of 0, and the point furthest away has a value of 1. The primary advantage of loess is the fact that it is a non-parametric normalization and henceforth can be used when an *explicit model* of the data is not available.

$$w(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases} \quad (2)$$

While the LOESS method is a nonparametric normalization of the data, dCHIP and PDNN (Position Dependent Nearest Neighbor), on the other hand, are model based approaches. In dCHIP (Li & Hung Wong, 2001), the primary normalization method creates a model of chip variability through the use of invariant marker genes, whilst PDNN (Nielsen et al., 2005) uses a nearest neighbor model to describe the physical binding of mRNA to the probes. The different relative binding of each probe to its exact match is then used to normalize the intensity value to obtain an expression level given the intensity. This method has the advantage of normalizing multiple probes for the same gene to roughly the same expression level. This is important because it was observed that oftentimes for time series, probes that had a different sequence that targeted the same gene had significantly different expression levels, despite being well correlated with each other. The problem with this issue is that oftentimes, one probe set will

show statistically significant differential expression, whereas the other probe-set would not.

PDNN works by assuming that the observed signaling intensity comes from three sources, Binding, Non-Specific Binding, and background. It calculates the expected intensity via Equation 3

$$\bar{I}_{ij} = \frac{N_j}{1 + e^{E_{ij}}} + \frac{N^*}{1 + e^{E^*_{ij}}} + B \quad (3)$$

Where I is the expected signal intensity, i is the probe index, j is the gene index, N_j is the number of mRNA copies in the sample, N^* is the number of mRNA copies that contribute to Nonspecific Binding, E_{ij} is the free energy of formation for the specific binding and E^*_{ij} is the average non-specific free energy, and B is the background intensity.

E_{ij} and E^*_{ij} are calculated as the sum of the stacking energy for the binding which is calculated via the nearest neighbor model given in Equation 4.

$$E = \sum w_k e(b_k, b_{k+1}). \quad (4)$$

The w_k represents the individual weights for each nearest neighbor pair given by ϵ . All of the parameters are computed by minimizing the following Equation 5.

$$F = \sum_i \sum_j (\ln \bar{I} - \ln I)^2 \quad (5)$$

The normalized expression for each gene then consists of N_j which would then be the predicted number of mRNA copies given the intensity values. This value of N_j is useful because it provides a method for reconciling probe sets which target the same gene, but due to variation in their mRNA sequence show different intensity profiles. This allows for the acquisition of a more accurate measure of mRNA activity than by just relying on intensity. Furthermore, genes represented by multiple probe-sets can have their predicted gene expression averaged to further reduce the contribution of noise in the overall signal, something which cannot be done when using intensity data only.

Comparisons to the different normalization methods (Millenaar et al., 2006; Ryden et al., 2006) have

suggested that while they show some difference, it was found that the set of genes that were selected as differentially expressed show a significant level of concordance.

Selection of the Statistical Cutoff

In the selection of genes that show a statistically significant chance of being non-random, there is the question of p-values. Many researchers utilize a p value of $p < .05$, or $p < .01$ (Lee, Kuo, Whitmore, & Sklar, 2000). However, it is important to consider the overall sample size. In a representative array, the Affymetrix RAE230A, there are over 15K probe spots on the array. With a p-value of .05, it still leads to the possible random selection of 750 genes, which represents a significant reduction in the number of genes being analyzed but may lead to results which would be difficult to justify statistically. Instead, the p-value ought to be set a p-value of $p < 1/N$ where N is the number of probes on the chip. For microarrays, we term this the “Natural” p-value since it directly relates the p-value to the sample size.

Gene Selection

The simplest filtering technique which researchers use to select for statistically significant genes is the 2-fold test. This test essentially selects for genes that have shown either a two fold up/down regulation, and is widely used because of its inherent simplicity. This test is essentially an approximation of the t-test with three samples at a confidence interval of $p < .05$. It tends to work poorly at low expression levels due to the greater variability vs. the mean, that is, a lower signal to noise ratio (Novak, Sladek, & Hudson, 2002) as well as the fact that the log normal distribution breaks down at the tail end corresponding to low expression levels. Therefore, before the use of the 2-fold selection method, there is usually a normalization step performed such as LOESS, as well as a filter for sample data quality to filter out the gene with a low expression level. However, despite the mathematical transformations to correct for the deviations from the log-normal distribution as well as the loss of the SNR, the 2-fold test represents a very weak filtering method because there is no explicit control as to what the confidence level of the selected genes are.

The problem with utilizing any sort of fold measure whether it be 2 fold, or n fold for greater or lesser

stringency is that one is unable to report the confidence level of the selection because of the dependence of the confidence interval upon the variation between the different replicates. As a general rule, there are two competing trends with utilizing this selection:

1. The repeatability of the gene chips is increasing due to improvements in quality and technological improvements such as the use of 60 base pair recognition sequences (Hardiman, 2004), making the 2-fold test more stringent now than in the past.
2. The number of spots per array has increased, thereby increasing the natural p-value needed to separate out random changes in expression levels.

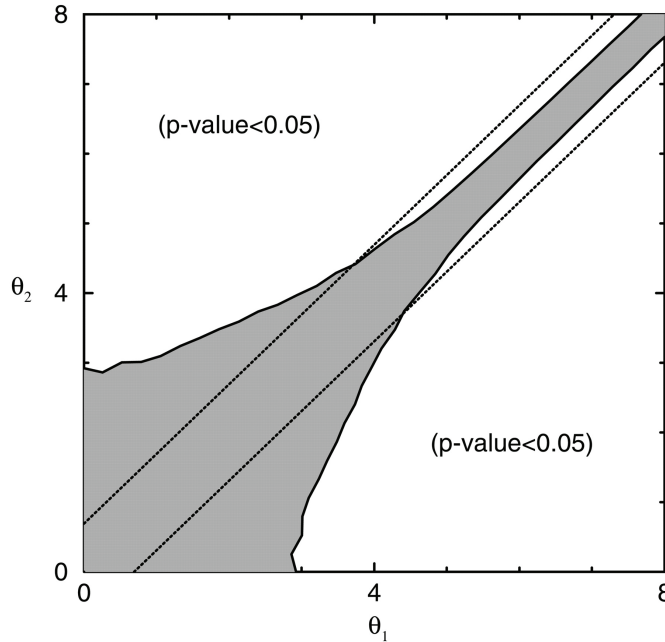
Both of these can be rectified with the use of the t-test which is give in Equation 6.

$$t = \frac{\overline{X}_a - \overline{X}_b}{\sqrt{\frac{\text{var}(X_a)}{\text{size}(X_a)} + \frac{\text{var}(X_b)}{\text{size}(X_b)}}} \quad (6)$$

The benefit of the utilizing the standard t-test is that the “Natural” p-value can be used as the cutoff for stringency as well as the fact that the metric is tailored for each individual gene with their attendant means and standard deviation. This method also has the added benefit of being able to handle the higher SNR found at the lower expression levels, Figure 2, because it does not assume the fact that there is a strict linear cutoff at all measured expression levels. Again as in the two-fold test, it is imperative that proper normalization take place before the use of the t-test, for without proper normalization, the any systematic errors that take place at lower expression levels tend to overwhelm the t-test.

A more sophisticated method for determining the quality of a gene’s expression data is the Significance Analysis of Microarrays (SAM) method (Tusher, Tibshirani, & Chu, 2001). This method utilizes similar statistical metric to the t-test Equation 7. The primary difference between SAM and the standard t-test is the presence of an S_0 term which is calculated form the data that is designed to minimize the coefficient of variance of $d(i)$ between the genes with low expression level and those with high expression levels. Additionally, instead

Figure 2. The variability between two chips. It is evident that as the expression level is lower, the linear trend between the two chips breaks down, and henceforth the 2-fold threshold does not become applicable. However, the t-test is able to correctly compensate for this change at lower expression level. Note however, that we do not advocate the use of the confidence interval of $p < .05$ for outlier detection (Tu, Stolovitzky, & Klein, 2002)



of focusing on a statistical distribution such as the one used for the t-test, the cutoff is determined arbitrarily by looking at a sorted plot of the $d(i)$ values, Figure 3. Despite the fact that it is slightly different than that of the t-test, it satisfies the same goal which is primarily the assessment as to whether or not the difference measured between two samples has been reliably captured. The primary advantage of this method is that due to the normalization term S_0 , there is increased sensitivity at lower expression levels, that is, it is less sensitive to SNR than the simple t-test.

$$d(i) = \frac{\bar{x}_l(i) - \bar{x}_u(i)}{\sqrt{a \left\{ \sum_m [x_m(i) - \bar{x}_l(i)]^2 + \sum_n [x_n(i) - \bar{x}_u(i)]^2 \right\} + s_0}} \quad (7)$$

FUTURE TRENDS

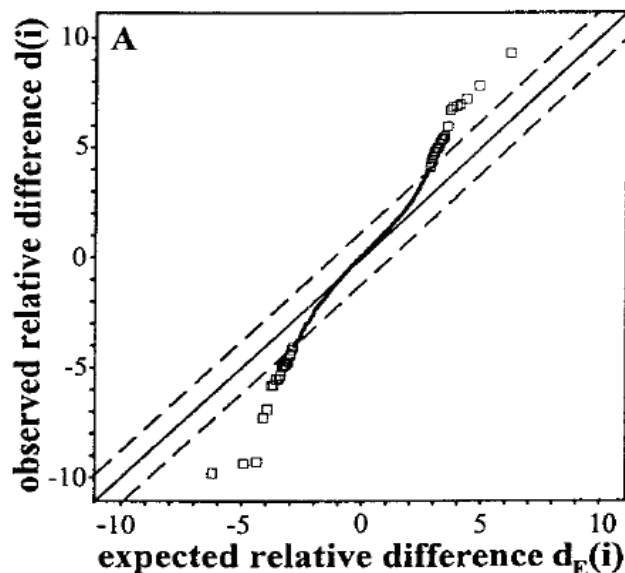
One of the major pushes recently with *microarrays* is the use of temporal expression data. On a fundamental level, the same problem applies, namely the selection

of statistically significant genes from the microarray. However, at this point it is not clear as to what a statistically significant temporal expression profile is. Some work has been done with the use of statistical over-representation of possible expression profile templates (Ernst & Bar-Joseph, 2006; Yang, Maguire, Yarmush, Berthiaume, & Androulakis, 2007) and future work has been proposed to make use of the replicates to identify those genes whose SNR in terms of *biological noise* is low enough where one can place high confidence in the overall shape of the obtained expression profile.

CONCLUSION

While microarrays have been hailed as a revolutionary advance in molecular biology, skeptics have pointed out the fact that there exists maddening inconsistencies in the results derived from microarray data (Kothapalli, Yoder, Mane, & Loughran, 2002). The primary problem with microarrays is the fact that at the same time, they are able to provide too much and not enough information. In the too much information realm, they provide the expression levels of thousands of genes at once, most

Figure 3. The threshold's for SAM allows for the differentiation between genes that show differential expression as compared to the base state



of which are not related to the underlying phenomenon being investigated. In the too few information realm, they are expensive enough where oftentimes there aren't sufficient replicates to deal with the inherent problem of technical and *biological noise*, making the selection of a set of genes questionable.

While technical noise is progressively being decreased through technological improvements via the makers of microarrays such as Affymetrix and Agilent, the problem of *biological noise* still remains. However, in either case, the ability to find experiments, genes, and probes whose intrinsic property remains despite the noise is still important. We feel that many of the criticisms of microarrays and the conclusions can be addressed with proper statistical evaluation of the data. This is not to say that conclusions that have not passed statistical muster do not contain important information; however statistical robustness adds another method of verification of the results. By utilizing experiment data quality assessments, one can determine whether an unknown process has indeed been captured in the context of both the biological response as well as technical issues such as the number of replicates, and the length of the time series. Proper analysis of the genes themselves will allow for the identification of genes that can be used for further analysis, and, finally, the analysis of the probes themselves ought to give

an insight into the overall confidence in the values obtained from the technology as well as providing information as to whether or not the gene is actually expressed in vivo.

REFERENCES

- Bowtell, D. D. (1999). Options available-from start to finish-for obtaining expression data by microarray. *Nat Genet*, 21(Suppl. 1), 25-32.
- Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet*, 21(Suppl. 1), 33-37.
- Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., & Childs, G. (1999). Making and reading microarrays. *Nat Genet*, 21(Suppl. 1), 15-19.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.
- Dupuy, A., & Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*, 99(2), 147-157.

Ernst, J., & Bar-Joseph, Z. (2006). STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7, 191.

Hardiman, G. (2004). Microarray platforms—Comparisons and contrasts. *Pharmacogenomics*, 5(5), 487-502.

Kothapalli, R., Yoder, S. J., Mane, S., & Loughran, T. P., Jr. (2002). Microarray results: How accurate are they? *BMC Bioinformatics*, 3, 22.

Lee, M. L., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA*, 97(18), 9834-9839.

Li, C., & Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol*, 2(8), RESEARCH0032.

Millenaar, F. F., Okyere, J., May, S. T., van Zanten, M., Voeselek, L. A., & Peeters, A. J. (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, 7, 137.

Nakai, K., & Vert, J. P. (2002). Genome informatics for data-driven biology. *Genome Biol*, 3(4), REPORTS4010.

Nielsen, H. B., Gautier, L., & Knudsen, S. (2005). Implementation of a gene expression index calculation method based on the PDNN model. *Bioinformatics*, 21(5), 687-688.

Novak, J. P., Sladek, R., & Hudson, T. J. (2002). Characterization of variability in large-scale gene expression data: Implications for study design. *Genomics*, 79(1), 104-113.

Ryden, P., Andersson, H., Landfors, M., Naslund, L., Hartmanova, B., Noppa, L., et al. (2006). Evaluation of microarray data normalization procedures using spike-in experiments. *BMC Bioinformatics*, 7, 300.

Smyth, G. K., Yang, Y. H., & Speed, T. (2003). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol*, 224, 111-136.

Tu, Y., Stolovitzky, G., & Klein, U. (2002). Quantitative noise analysis for gene expression microarray

experiments. *Proc Natl Acad Sci USA*, 99(22), 14031-14036.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9), 5116-5121.

Wang, X., Hessner, M. J., Wu, Y., Pati, N., & Ghosh, S. (2003). Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics*, 19(11), 1341-1347.

Wu, J., & Androulakis, I. P. (2005). *Selecting maximally informative genes: The interplay between accuracy and complexity*. Paper presented at the 18th International Conference on Systems Engineering.

Yang, E., Maguire, T., Yarmush, M. L., Berthiaume, F., & Androulakis, I. P. (2007). Bioinformatics analysis of the early inflammatory response in a rat thermal injury model. *BMC Bioinformatics*, 8(1), 10.

Zhang, L., Miles, M. F., & Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol*, 21(7), 818-821.

KEY TERMS

Biologically Informative: This is different from the notion of statistically significant because this set of genes is consistent over multiple experiments, replicates, and microarray platforms and reflects the underlying ground truth.

Locally Weighted Normalization of a Scatter Plot (LOESS; LOWESS): LOESS seeks to find a low order polynomial that best describes the overall variation in a scatter plot. This is used to normalize for the nonlinearities found in two state experiments.

Natural P-Value: The p-value a researcher should set in determining statistical significance. It is wholly reliant upon the number of samples in a given trial. Therefore, in a microarray, the natural p-value should be set to 1/N where N is the number of samples

Position Dependent Nearest Neighbor (PDNN) Model: A normalization technique by Zhang, Miles, and Aldape (2003), which makes the assumption that

the signal intensity is dependent on both the probe sequence being used and the number of mRNA copies. It performs the normalization by calculating the number of mRNA copies and an expected signal intensity by optimizing for various parameters such as base stacking energy.

Significance Analysis of Microarrays (SAM): A selection algorithm which is nominally very similar to that of the t-test. It is, however, more robust to mRNA signals of lower SNR and hence gives more reliable filtering for genes of low expression levels.

Statistically Significant: The ability for the variability of a sample to be attributed by a factor other than through random noise. This is dependent first upon the overall distribution of the samples, though most researchers assume that the random variations are gaussian. Due to systematic factors such as dye binding affinities as well as the nonlinear binding behavior in microarrays, normalization is required before the use of this gaussian assumption

Signal to Noise Ratio (SNR): In the context of microarrays, the noise comes from two sources, technical and biological. This is, however, the primary determinant of how many replicates are required but is complicated via the fact that different probes have different SNR.