[125h]

A New Probabilistic Framework for cDNA Microarray Data Analysis

Babatunde A. Ogunnaike, Claudio A. Gelmi and Jeremy S. Edwards

Department of Chemical Engineering
University of Delaware
Newark, DE 19716-3110
ogunnaik@che.udel.edu

Prepared for Presentation at the 2003 Annual Meeting, San Francisco, CA, Nov. 16-21

July 2003

Extended Abstract

One of the primary goals of functional genomics is to provide a quantitative (as opposed to qualitative) understanding of the functions of genes, how they influence and are influenced by proteins and the environment, and how they regulate the function of complex living organisms from the cellular level all the way to the physiological level. The success of this enterprise, however, depends significantly on the quality of the data upon which such quantitative understanding is to be based—highlighting the importance of the generation of high quality gene expression data and the development of effective techniques for data analysis.

Gene expression studies, as currently done with printed arrays or pre-fabricated gene chips, generate large quantities of data from which one may determine the relative expression levels of literally thousands of genes in a cell. Many techniques have been proposed for determining estimates of differential expression level for each gene of interest, each incorporating varying degrees of statistical rigor, from some of the earliest (Chen, 1997; Audic and Claverie, 1997) with the most rudimentary attention to statistics, to the more recent contributions (e.g. Newton, et al, 2001; Rocke and Durbin, 2001; Olshen and Jain, 2002; Ibrahim *et al*., 2002) where serious attempts have been made to confront the daunting statistical problems in a more systematic fashion. Nevertheless, the fundamental problems associated with these techniques for generating gene expression data still remain: there are so many sources of systematic as well as random errors that ultimately, without performing a prohibitively large number of carefully replicated experiments, the experimental information will be insufficient to allow adequate estimation of the desired expression levels with sufficient confidence. In the final analysis, experimental designs and statistical analysis for microarray technology are generally recognized as still evolving, with no consensus yet on effective methodologies (see reviews by Nadon and Shoemaker 2002, and Sebastiani *et al.*, 2003).

We present in this paper a novel approach to the analysis of cDNA Microarray data. It centers on a theoretical framework that recognizes the intrinsic characteristics of microarrays and the mechanisms behind the data generation. Specifically, we have developed a theoretical basis for the following tasks: (i) Representing microarray data appropriately as an ensemble; (ii) Characterizing this ensemble; (iii) Analyzing microarray data sets on the basis of this ensemble characterization; and (iv) Drawing realistic inferences.

Representation

Let I_{i1} and I_{i0} be the fluorescence signal intensities measured from each spot i on a microarray, with I_{i1} obtained from the gene in question under test conditions, and I_{i0} from the gene under control conditions. We have established that the original data in the form of the ordered pair of signal intensities (I_{i1}, I_{i0}) should be converted from this "cartesian" representation to the corresponding "polar" form (r_i, x_i), where r_i is the usual vector magnitude (in this case, $\sqrt{(I_{i1}^2 + I_{i0}^2)}$), the "intensity magnitude"), and x_i , the "fractional intensity" (the ratio $I_{i1} / (I_{i1} + I_{i0})$), is related to the vector angle. Several advantages accrue from this coordinate transformation: (i) It is *more natural* (because of the inherent heteroskedasticity of microarray data—the variance increases with increasing intensity measurement); (ii) It normalizes the data *more efficiently* because none of the original information is lost. Observe that while x_i is dimensionless and naturally scaled between 0 and 1, r_i retains the information about intensity magnitude, thus preserving the two-dimensional character of the original data. (Standard techniques to date are based solely on intensity ratios, losing all information about intensity magnitude. (cf. Newton, *et al.* 2002)). Most importantly, (iii) it *enables statistical rigor and endows the problem with analytical tractability* not possible with either raw intensity data or the popular intensity ratios. This is because r_i

provides a rational metric for partitioning the data, (allowing us to separate low magnitude data from higher magnitude data), and the probabilistic characteristics of x_i *can be derived from first principles* as indicated in the following key result:

Key result: Based on a formal derivation by which we establish that, under very mild and reasonable assumptions, the signal intensities possess a Gamma distribution, we show that the random variable X_i (from which the observed fractional intensity x_i is but a single realization) possesses a Beta (α_{i1}, α_{i0}) distribution where the indicated parameters arise directly from the Gamma distributions of the contributing intensity signals. Hence:

> *Casting microarray data in the form of a histogram of fractional intensities x_i therefore yields a representation whose underlying theoretical probability distribution is a mixture of overlapping Beta density functions.*

Characterization, Analysis and Inference

We may thus characterize the histogram of fractional intensity data as a mixture of *at least* three Beta distribution functions, $f_1(x)$ for the data from the collection of genes showing lower differential expression, $f_0(x)$ for those showing no differential expression, and $f_2(x)$ for those showing higher differential expression. Thus by fitting a mixture of Beta distributions to histogram data, we obtain a theoretical probabilistic characterization of the data where each contributing Beta distribution constitutes the population description of a category of genes with common differential expression attributes. We are then in a position to use this ensemble beta density description to assign probabilities to each gene of belonging to any one of the categories identified from the data histogram. The details of how this characterization and subsequent analysis are carried out in practice (for various data classifications based on values of r_i, the associated intensity magnitudes) will be discussed in the presentation. Also as a natural consequence, the method allows us to correct for the bias due to the differences between the dye labeling.

The final outcome of our proposed analysis technique will therefore be an ordered pair of results for each gene: a raw computed fractional (or relative) change in expression level, and an associated probability that this number indicates lower, higher, or no differential expression between the test and reference conditions (a “category-membership probability”).

The problem of statistical inference in this framework therefore becomes one of assigning to each gene the probability that it belongs in the category of those showing lower, higher, or no differential expression. From here, the researcher is then able to combine these assigned probabilities with any domain knowledge to identify a subset of genes for more precise study.

Additional details about the theory and implementation of the technique (including how it allows us to correct for the well-known existence of die bias) will be discussed in the presentation.

Application results

Experimental data from gene expression studies in *Deinococcus radiodurans* following DNA damage (carried out by Dr. John Battista of LSU) have been analyzed using this technique.

Applying the traditional fold-change criteria (greater than 2.0 or smaller than 0.5) to the raw *D. radiodurans* data results in the identification of 743 genes (or 8% of the 9237 genes) as being differentially expressed (598 genes down-regulated and 145 genes up-regulated). Upon compensating *explicitly* for the inherent dye bias as prescribed by our methodology, (but still using the fold-change criteria for decision-making) these results change significantly in three important ways:

1) the total number of genes considered to be differentially expressed is reduced from 743 to 429 (a 42.3% reduction);
2) furthermore, this net reduction of 314 genes is composed of 402 "deletions" and 88 new "additions"; in other words, 402 of the original 743 genes are eliminated as "false positives", while 88 of the remaining 8494 genes are "false negatives" that had gone unidentified;
3) the new total of 429 differentially expressed genes are distributed as follows: 196 genes are up-regulated and 233 are down regulated. (Compare with the distribution under the traditional fold-change criteria where there are 4 times as many down-regulated genes as there are up-regulated ones).

We recall now that the true essence of our proposed method is to assign to each gene the probability that it is down-regulated, up-regulated, or not differentially regulated at all. A full analysis of the data from this perspective therefore results in a list of all the genes along with the associated probability of differential expression; the researcher is then able to make decisions regarding which genes to pursue for further studies using these probability values. However, for the purpose of comparing and contrasting this probability-based approach to the (bias-corrected) fold-change decision criteria, we present the following sample set of results:

1) A total number of 1701 genes have a probability of 0.5 or greater of being differentially expressed (representing 18.4% of the total number of genes);
2) To obtain a list of genes that encompasses those identified using the (bias-corrected) fold-change criteria requires a threshold of 0.74 for the probability of differential expression. Specifically, a total of 1206 genes have a probability of 0.74 or greater of being differentially expressed—a list that includes *all* the 429 genes identified using the bias-corrected fold-change criteria.
3) When the threshold of the probability of differential expression is increased to 0.99, the number of implicated genes is 434, which is comparable to the 429 identified using the bias-corrected fold-change criteria, except that these two lists share only 244 genes in common.

Some important implications of this final result are as follows: (i) of the 429 genes identified by the (bias-corrected) fold-change criteria, only 244 have a probability of 0.99 or higher of being differentially expressed; the others have a lower probability of differential expression. Also, there are 190 genes (434 less 244) identified by our probability-based technique with associated probability of differential expression of 0.99 or higher that were *not* identified as differentially expressed by the (bias-corrected) fold-change criteria.

Future work

Future work will include developing a strategy within this framework for obtaining estimates of the percent increase (or decrease) in gene expression level (and measures of associated variability), and then developing a similar methodology for Affymetrix GeneChip arrays.

References

Audic, S. and Claverie J. M. (1997). The significance of digital gene expression profiles. *Genome Res*. 7, 986-995.

Chen, Y., E. R. Doherty, and M.L. Bottner. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, **2** 364-374.

Ibrahim, J.G., M. H. Chen, and R. J. Gray. (2002) Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97, 88-99.

Nadon, R. and J. Shoemaker, (2002). Statistical Issues with microarrays: processing and analysis. Trends in Genetics, 18, **5**, 265-271

Newton, M.N., C.M. Kendrziorski, C.S. Richmond, F.R. Blattner and K.W. Tsui (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8, 37-52.

Olshen, A. B., and A. N. Jain. (2002). Deriving quantitative conclusions from microarray expression data, *Bioinformatics*, 18, 961-970.

Rocke, D. M. and B. Durbin, (2001). A Model for measurement error for gene expression analysis. *Journal of Computational Biology*, 8, 557-569.

Sebastiani, P., E. Gussoni, I Kohane, and M. Ramoni. (2003). Statistical Challenges in Functional Genomics, *Statistical Science*, 18, 1, 35-70.