

A Bayesian approach to the analysis of cDNA microarray data

M.A. Black^{1,2}, B.A. Craig^{1,2}, M. Tanurdzic^{2,3} & R.W. Doerge^{1,2,4}

¹Department of Statistics, Purdue University, ²Computational Genomics, Purdue University
³Purdue Genetics Program, Purdue University, ⁴Department of Agronomy, Purdue University

Abstract

The recent explosion of interest in microarray technology has resulted in it becoming the preferred methodology for conducting gene expression experiments. Although the ability of an array experiment to simultaneously examine the expression of thousands of genes gives a previously unheard of level of insight to researchers, it also raises a plethora of statistical questions involving both the sheer volume of data being produced, as well as the variability inherent in this technology. In this paper we present statistical methods based on Bayesian linear models to investigate the various sources of variability present in array experiments. Data from a previously published cDNA microarray experiment is used to illustrate this methodology.

A major goal of genomic research involves the determination of gene function, the discovery of which ultimately gives investigators fundamental insight into the ways in which genes act to affect the traits exhibited by an organism. The ability of a gene to influence an organism's characteristics is the result of the manufacture (expression) of proteins by that gene, the identity of which is determined by the gene's genetic sequence (DNA). By observing the expression patterns of genes (i.e., the extent to which genes are turned on or off) under various treatment conditions, important clues about gene function can be obtained. Studies which investigate such patterns are termed *gene expression experiments*.

A common method for determining the level of expression is to measure the amount of mRNA (an intermediate product in protein production) being produced by that gene. Although not exact, there is relatively high degree of correspondence between the volume of mRNA present, and the amount of protein produced. The ability to measure mRNA levels is not new, however, microarray technology^{1,2} provides an avenue to perform such experiments on a grand scale. Specifically, these methods allow researchers to obtain expression data from thousands of genes under various treatment conditions in a single experiment. This wealth of data has led to the proposal of many different analytic approaches, with an early preference for clustering (i.e., grouping similar gene expression profiles)³⁻⁷ gradually giving way to studies of experimental variability^{8,9}, and methods for determining significant differential gene expression across treatments¹⁰⁻¹³.

A microarray experiment consists of mRNA extracted from cells under different treatment conditions, and glass slides (the microarrays) to which spots of genetic material are attached. In the case of Affymetrix arrays², this material comprises oligonucleotides (short genetic sequences) which are synthesized on the array itself, while for cDNA arrays¹ the material is complementary DNA (cDNA), which is printed onto the array by a robotic tool. In both cases, a single spot on the array comprises thousands of strands of identical cDNA which represent the sequence of

a single gene, with thousands of spots (and thus thousands of genes) able to fit on a single array. Although the two technologies share many similarities, the focus here is on the cDNA array.

In its most simple form, the aim of a cDNA microarray experiment is to measure the fold change in mRNA expression between two different treatment conditions for a collection of genes. This is accomplished by tagging the extracted mRNA from the two treatment conditions with fluorescent labels (often green (*Cy3*) for the control condition, and red (*Cy5*) for the treatment condition), mixing the mRNA samples together, and then placing the combined mRNA on the microarray to allow *hybridization* to occur. That is, allowing the single-stranded mRNA to bond with the cDNA segments attached to the microarray. Since the sequence for each gene is unique, each gene's mRNA will only bond with the unique complementary sequence *from that gene*, which means that each spot only collects mRNA that was produced by the gene it represents. Using a laser scanner the labeled mRNA can be fluoressed, thus providing an estimate of the amount of mRNA from each treatment condition that has hybridized at each spot on the array. The set of intensity signals from each of the red and green labels are often referred to as *channels* in array experiments.

In addition to determining the red and green signal intensities, the software also calculates a *background intensity* for each label color at each spot. The background intensity aims to remove factors affecting the intensity levels (such as reflective glare from the array surface) directly associated with the presence of hybridized mRNA at the spot. This practice is referred to as *background correction*. Similar corrections must also be made for differing average signal intensities between the two channels, and across multiple slides. Such *normalization* procedures generally consist of standardizing the intensities from each channel to a median of one (i.e., by dividing through by the median), thus putting all intensities of the same scale¹⁴. For an array with only a single spot per gene, the ratio of the two fluorescence intensities at each spot gives an indication of the relative abundance of mRNA being produced by each gene under each treatment condition, allowing the identification of genes whose mRNA expression levels differ between the two treatments. Such an approach, however, assumes that the experimental process is relatively free of variability, and that all differences in fluorescence intensity are the result of changes in expression level caused by differences in mRNA concentrations in each treatment condition. In reality this is usually not the case, with variability introduced by the chance nature of hybridization, and intensity changes affected by the fluorescent labels, the arrays, the treatment conditions, and the genes themselves. In order to overcome these problems, Kerr *et al.*¹² proposed a linear models approach for the analysis of gene expression data. This allowed experimental effects to be incorporated as model parameters, thus accounting for the influence on fluorescence intensity, and excluding them from the estimation of the random error component. The use of a linear model in conjunction with techniques to account for multiple comparisons then makes it possible to detect which genes in an experiment display changes in expression level between treatment conditions.

This paper extends the work of Kerr *et al.*¹² and Kerr and Churchill¹³ by taking a Bayesian approach to the analysis of gene expression data. In this work

Markov chain Monte Carlo (MCMC) techniques were utilized to obtain the joint posterior distribution of the linear model parameters, and based on these results, a Bayesian stepwise selection method is proposed for detecting a maximal number of genes which have undergone differential expression, while maintaining a suitably low posterior probability of error.

Methods

Analysis of variance

The major strength of the ANOVA approach is its ability to take all experimental effects into account, whether or not they are of interest to the investigator. This allows the systematic variation due to these factors to be separated from the true random error that is present in the experimental process.

Let X_{ijkgr} represent the channel j background subtracted intensity for replicate r of gene g under treatment k on array i . Kerr *et al.*¹², proposed the following fixed effects ANOVA model,

$$Y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + VG_{kg} + \epsilon_{ijkgr} \quad (1)$$

where $Y_{ijkgr} = \log(X_{ijkgr})$, μ represents the overall mean expression intensity, A is an effect due to array, D is an effect due to the fluorescent tag (dye) used in the labeling process, V is a variety (treatment) effect, G is a gene effect, and VG is the effect of an interaction between variety and gene. In other works this basic model has been extended to include various interactions between the main effects^{13,15}, as well as the treatment of A as a random effect^{16,17}. Regardless of the form of the model, however, the basic idea of the ANOVA approach to the analysis of gene expression data is to isolate the intensity changes which are due only to differences in the response of individual genes to the treatment conditions.

In assessing the presence of differential expression, the variables of interest are the treatment by gene interaction terms. By formulating contrasts between these variables for each gene, it can be determined whether the expression levels of each gene are changing between treatment conditions. Such comparisons can be dealt with via hypothesis testing, where the following hypothesis is tested for each gene across the range of treatment conditions

$$H_0 : VG_{kg} = VG_{k'g} \text{ versus } H_1 : VG_{kg} \neq VG_{k'g} \quad (2)$$

For an experiment involving arrays of n genes and t treatment conditions, this results in $nt(t-1)/2$ hypothesis tests, making multiple comparisons a major issue even in modest experiments. For example, a reasonably small scale study involving 1000 genes and 5 treatments requires 10,000 hypothesis tests. In order to deal with this issue, various multiple comparison procedures (e.g., adjusted p-values¹⁸, or control of the false discovery rate¹⁹) have been proposed, although in at least one case the techniques employed were criticized for their conservatism²⁰.

An alternative to the frequentist multiple comparison approach is to use Bayesian methodology to investigate the joint posterior distribution of the treatment by gene interaction terms. By manipulating this distribution, inferences can be made about the *joint* probability of non-zero differences occurring between treatment conditions across each gene in the experiment. This allows a probability statement to be made

about the joint distribution of the treatment by gene contrasts, while at the same time taking into account the correlation structure of the model parameters.

Bayesian analysis

The Bayesian approach to linear models is well established^{21,22} and has become a popular alternative to traditional analysis methods with the advent of user-friendly Markov chain Monte Carlo techniques^{23,24} and software packages²⁵. The fact that the absence of informative prior information generally provides results very similar to those of a standard ANOVA analysis means that the main advantage of the Bayesian approach is the ability to make probabilistic statements about the model parameters, and thus the probability of differential expression.

The form of the Bayesian linear model used here is identical to the fixed effects model shown in (1), however the inclusion of additional parameters (either as fixed or random effects) is easily accomplished. Diffuse independent non-informative priors were used for each of the the fixed effects parameters, while a normal distribution centered at zero with an unknown variance component was used for the error distribution. For the variance component a diffuse non-informative prior was again used.

Multiple comparisons

In the standard least squares solution to the fixed effects ANOVA model (or, equivalently, the REML solution to the mixed model), a single “best” estimate is produced for each model parameter. In order to detect differential expression, one must determine which differences between VG terms are statistically significant, with a gene considered to have undergone differential expression if at least one pairwise treatment contrast for that gene is significant. Using this definition, the size of the linear model used here requires the comparison of many linear combinations of model parameters, creating multiplicity issues on a grand scale, and necessitating the use of multiple comparisons procedures.

If one is not concerned with multiplicity issues, the obvious way for the Bayesian to proceed would be to calculate the marginal probability that each contrast is non-zero, and conclude that any gene with at least one pairwise treatment contrast having a marginal posterior probability of at least $(1 - \alpha)$ underwent some change in expression level between treatment conditions. Such an approach is roughly equivalent to performing many frequentist hypothesis tests in the absence of a multiple comparisons correction.

The general idea of multiple comparison corrections is to protect against certain types of error, while taking into account the number of comparisons made. In the context of gene expression, the goal is to detect genes which exhibit changes in expression level between treatments, while protecting against the risk of falsely classifying unchanged genes as being differentially expressed. This is equivalent to trying to detect as many genes as possible which undergo changes in expression level between treatments, while still maintaining a joint posterior probability of $1 - \alpha$. Thus, to deal with the multiplicity issue in a Bayesian manner, one could find the largest set of genes whose joint probability of differential expression is greater than or equal to $1 - \alpha$.

Computational details

In order to obtain the joint posterior distribution for the parameters of the mixed model proposed in (1), a Markov chain Monte Carlo approach utilizing the Metropolis-Hastings algorithm^{26, 27} was used. The large number of observations associated with gene expression experiments meant that even the relatively simple models used here contained many parameters, necessitating the use of efficient computational methods. For this reason custom written Fortran code was created to implement the Metropolis-Hastings algorithm, rather than relying on pre-existing software.

Although the task of detecting a maximal number of genes whose joint probability of differential expression is at least $1 - \alpha$ is relatively simple in low dimensions, the size of gene expression experiments makes this a challenging problem. To overcome the difficulty imposed by the high dimensionality of the posterior distribution, a procedure similar to that employed in stepwise regression was used. To begin with, all genes whose marginal posterior probability of differential expression was greater than $1 - \alpha$ were considered as candidates for inclusion in the final joint probability statement. The joint posterior probability that *all members* of this set underwent differential expression was then calculated, and, if less than $1 - \alpha$, the posterior probability of the set *with a single gene removed* was calculated, for each gene in the set. The gene whose removal increased the joint posterior probability by the most was then discarded, and the process continued until a set of genes was found which satisfied the $1 - \alpha$ constraint. The genes in this set were then declared to have *all* undergone differential expression with joint probability $1 - \alpha$.

Results

Analysis of human liver/muscle tissue data

In order to compare the Bayesian analysis to the standard frequentist approach, the data set of Kerr *et al.*¹² was used. These data comprised 1286 distinct cDNA sequences (spots) on two slides, with the “treatment” conditions being mRNA extracted from either muscle or liver tissue. Dye-swapping was used to ensure that any dye effects were estimable. This resulted in a completely balanced experiment, with each treatment appearing on each slide, as well as being labelled with each dye, and each dye appearing on each slide.

Based on the results of their model, Kerr *et al.* detected 305 differentially expressed genes, with 201 down-regulated between liver and muscle, and 104 up-regulated between liver and muscle. In contrast, the Bayesian approach selected a set of 37 genes which were differentially expressed, with 27 up-regulated, and 10 down-regulated. These genes were a subset of those produced by the Kerr *et al.* methods, with the basic ordering in terms of magnitude of VG terms very almost identical for the two models.

Conclusions

In terms of linear model parameter estimates, the results of the Bayesian methods gave very close agreement with the results obtained by Kerr *et al.*¹², with both analyses also agreeing on the genes which undergo the largest changes in expression between treatment and control. This is not surprising, since the non-informative Bayesian analysis is roughly equivalent to a frequentist maximum likelihood ap-

proach. The biggest difference between the two analyses was in the area of determining which genes were *significantly* differentially expressed.

The approach of Kerr *et al.* was to use a per gene bootstrap confidence level of 99% (i.e., $\alpha = 0.01$) to detect significant differential expression. Using a Bonferroni-type calculation, this α level should yield a minimum family-wise error rate of $1 - 0.99^{1286} \approx 1$, which indicates that a list of significantly differentially expressed genes produced by such a method will almost certainly contain at least one type I error. If a Bonferroni correction were used an α level of $\alpha = 0.05^{1286} \approx 0.00004$ would be required to declare significance. In contrast to this, the Bayesian analysis produced a set of genes with joint posterior probability of differential expression of 0.95, and a minimum *marginal* probability of differential expression of 0.996 (i.e., a “p-value” of 0.004), making it a less conservative approach than the Bonferroni correction.

Discussion

The aim of this work was to provide a framework for the analysis of gene expression data from a Bayesian perspective. The methods presented here accomplish this goal by taking a linear models approach to the analysis, with MCMC methods used to obtain observations from the joint posterior distribution of the model parameters. By forming contrasts of interest from this distribution, a maximal set of genes for which the joint probability of differential expression is at least $1 - \alpha$ can be found using the stepwise selection technique.

The main advantage of this approach is that it allows general statements to be made about the joint probability of differential expression of groups of genes. This removes the need for formal multiple comparisons corrections, and presents results in a manner which are easily interpreted by non-statisticians. The biggest drawback to the the MCMC approach to fitting Bayesian linear models to gene expression data is the relatively high computational demands created by the number of model parameters required to represent the experimental setup. Although the authors feel that any such disadvantage is more than offset by the increased interpretability of the results, any methods for improving the speed of the MCMC algorithms used would certainly be worthwhile.

With this in mind the authors hope to continue to improve the methods presented here by both increasing the speed of their implementation, and by applying them to the increasingly complex models which are currently being proposed. It is hoped that these methods will provide a useful tool for those involved not only in gene expression analysis, but also for anyone attempting to deal with multiplicity issues in a Bayesian framework.

Acknowledgments

The authors thank Professor John Deely for his insightful comments on Bayesian multiplicity, and Professor Jeff Bennetzen for his ongoing support of the genetic experimentation associated with this work.

This work is funded by a USDA-IFAFs (00-52100-9615) grant to R.W. Doerge.

References

1. Schena M., Shalon D., Davis R., & Brown P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467–470.
2. Chee M., Yang R., Hubbell E., Berno A., Huang X., Stern D., Winkler J., Lockhart D., Morris M., & Fodor S. (1996) Accessing genetic information with high density DNA microarrays. *Science* **274**:610–614.
3. Eisen M., Spellman P., Brown P., & Botstein D. (1998) Cluster analysis of genome-wide expression patterns. *PNAS* **95**:14863–14868.
4. Tamayo P., Slonim D., Mesirov J. Zhu Q., Kitareewan S., Dmitrovsky E., Lander E., & Golub T. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hemotopoietic differentiation. *PNAS* **96**:2907–2912.
5. Törönen P., Kolehmainen M., Wong G., & Castrén E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Letters* **451**:142–146.
6. Brown P., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares M., & Haussler D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS* **97**:262–267.
7. Holter N., Mitra M., Maritan A., Cieplak M., Banavar J., & Fedoroff N. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *PNAS* **97**:8409–8414.
8. Lee M., Kuo F., Whitmore G., & Sklar J. (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *PNAS* **97**:9834–9839.
9. Black M. & Doerge R. (2001) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experimemnts. *Proceedings of the Conference on Applied Statistics in Agriculture* .
10. Chen Y., Dougherty E., & Bittner M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**:264–374.
11. Newton M., Kendzierski C., Richmond C., Blattner F., & Tsui K. (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**:37–52.
12. Kerr M., Martin M., & Churchill G. (2000) Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**:819–837.
13. Kerr M. & Churchill G. (2001) Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**:123–128.
14. Hegde P., Qi R., Abernathy K., Gay C., Dharap S., Gaspard R., Hughes J., Snesrud E., Lee N., & Quackenbush J. (2000) A concise guide to cDNA microarray analysis. *BioTechniques* **29**:548–562.
15. Kerr M. & Churchill G. (To appear) Analysis of variance for gene expression microarrays. *Biostatistics* .
16. Vitek O., Craig B., Black M., Tanurdzic M., & Doerge R. (2001) Designing microarray experiments: chips, dips, flips and skips. *Proceedings of the Conference on Applied Statistics in Agriculture* .
17. Wolfinger R., Gibson G., Wolfinger E., Bennett L., Hamadeh H., Bushel P., Afshari C., & Paules R. (In press) Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* .
18. Westfall P. & Young S. (1993) *Resampling-based based multiple testing: examples and methods for p-value adjustment*. Wiley, New York.
19. Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate: a prctical and powerful approach to multiple testing. *Journal of the Royal Statistcial Society, Series B* **57**:289–300.
20. Dudoit S., Yang Y., Callow M., & Speed T. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, Statistics Department, University of California at Berkeley.

21. Box G. & Tiao G. (1973) *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Company.
22. Broemeling L. (1985) *Bayesian Analysis of Linear Models*. Marcel Dekker, Inc.
23. Gilks W., Richardson S., & Spiegelhalter D. (eds.) (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall.
24. Carlin B. & Louis T. (2000) *Bayes and Empirical Bayes Methods For Data Analysis*. Chapman and Hall.
25. Lunn D., Thomas A., Best N., & Spiegelhalter D. (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**:325–337.
26. Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., & Teller E. (1953) Equation of state calculations by fast computing machine. *Journal of Chemical Physics* **21**:1087–1091.
27. Hastings W. (1970) Monte Carlo sampling methodss using Markov chains and their applications. *Biometrika* **57**:97–109.