

Gene Expression Patterns During Somatic Embryogenesis in Maize Tissue Culture: A Bayesian Approach

Tanzy Love, Alicia Carriquiry, Ping Che, Stephen Howell, Bronwyn Frame, and Kan Wang
Department of Statistics, Iowa State University
Ames, IA 50011-1210

1 Introduction

Large cDNA microarray studies are carried out to investigate complex processes in cyclic and developmental behavior. Typically, genes from biological materials that have been subjected to different treatments or that arise from different tissues or from the same tissue at different stages of development are spotted on different arrays (or slides). The objective is to draw inferences about differential gene expression levels across treatments, tissues or developmental stages. Gene expression levels can only be compared across different slides after the appropriate background cleaning and normalization procedures have scaled the data to the same range. There are several important sources of variation in gene expression measurement that must be accounted for in statistical analyses, and much of this variation is slide specific.

The data generated from cDNA microarray experiments are obtained by combining two types of images of the microarray slide. The two images are obtained while the slide is excited with a laser tuned to Cy5 and to Cy3 fluorescent dyes, respectively. Different laser strengths and the sensitivities of the camera result in different images. While a setting for laser and camera may produce a large number of saturated spots on the slide image, other settings may result in too many spots with measured expression below the minimum that can be captured by the instruments. The laser strength and the sensitivity of the camera to light can be adjusted by the operator to find a ‘best’ picture, one where most of the spots show some measurable expression and where very few of the spots reach saturation. Once the best image is obtained, the spots of cDNA must be found, a process called segmentation, and the background of each spot must be calculated. Finally, before true analysis of the data can be underway, each slide must be normalized within itself to recover from any systematic dye bias (usually Cy3 is stronger) and all the slides must be normalized together to make them comparable [1].

In this work, we focus on the measurement error that is introduced when scientists vary the strength of the laser and the sensitivity of the camera used to record expression levels and propose a statistical approach that allows incorporating multiple readings of each slide into the analysis. We show that under relatively lax model assumptions, expression levels can be estimated with significantly lower bias and higher precision when combining multiple readings for each gene into the statistical analysis than when choosing only the ‘best’ reading.

This paper is organized as follows. In Section 2 we describe in more detail the process by which multiple slide readings are obtained and hypothesize about the association between various slide readings. Section 3 develops the Bayesian hierarchical model for jointly analyzing multiple cDNA slide scans. Section 4 applies the proposed approach in a maize embryogenesis experimental dataset. The proposed approach is compared to other approaches published in the literature and that address this type of measurement error. Section 5 discusses the implications of incorporating multiple slide scans into the statistical analyses of microarrays.

2 Multiple Laser and Sensor Settings

Different laser and sensor settings can be used to read a cDNA microarray slide. Stronger laser settings create more fluorescence and stronger sensor settings pick up more signal. There is a balance to be struck between picking up signal from the lowly fluorescing spots and over-exposing the highly expressing genes. There is an upper limit of 65535 to the measurement of fluorescence; readings of spots which are brighter are truncated. Over-exposing the high intensity spots will cause them to be artificially near other high expression values. Correspondingly, low signals will be artificially assigned to 0 if the laser and sensor settings are too low.

The expression estimates used are background corrected average pixel intensities. An overexposed spot will have variation in its pixels due to inconsistencies in spot printing and irregular spot shape. Further, background correction will reduce the measured expression value so that 65535 is not the true point of truncation. An underexposed spot will also have variation in its pixels for the same reasons. Background correction, however, will create expression estimates that are negative or near zero. The truncated values in the figure correspond to those genes that are not expressed in that particular slide, but also to those genes which may have exhibited measurable expression levels had the spot been more exposed. Negative expression measurements are routinely set to zero, however, often the true point of truncation is not zero.

Multiple readings of the microarray slides can be taken for both fluorescence channels. Since all of the readings at different settings attempt to capture true expression levels for the genes on the slide, it is reasonable to assume that all readings contribute useful information about true expression levels and to think of combining the multiple readings into one estimate of gene expression for each spot. If the readings at different settings contain information about the true expression of the gene, then the variance in estimated gene expression that is due to the experimental process should be reduced in the estimate that is based on all available readings.

Several aspects of the measurement process of gene expression create challenges for statistical modeling. As discussed earlier, many microarray experiments include pseudo-replicates, which we define as the multiple readings of the slide under different laser and sensor settings. Generally, settings for different slides are very different because of the large experimental variation between slides. That is, one slide may result in a good reading at low laser and sensor settings while another may require higher settings to reduce the number of expression levels below threshold while keeping the number of overexposed spots to a minimum. Because of this practice, we are typically unable to assume that the settings act as blocks in a traditional experimental design. However, since the settings to read the two channels are almost always picked independently across slides, we can model each slide/dye combination separately. In what follows, we consider an arbitrary slide and dye channel in the experiment and propose a hierarchical model for estimating gene expression levels that permits incorporating multiple measurements for each gene into a single analysis.

3 Bayesian Hierarchical Gamma Model

In order to estimate gene expression, we propose a Bayesian hierarchical model. This model incorporates all slide scans into one estimate of expression per spot. To formulate the model, we rely on the natural ordering of slide readings. For instance, if we have two readings with the same sensor setting and different laser settings, the measurements on the reading with the higher laser setting will tend to be larger. Dudley et al. (2002) discuss gene expression and its dependence on changes in one of the experimental settings. Here we consider multiple settings simultaneously and to do so order the slides from smallest to largest median reading. Clearly,

the median-based ordering is subject to some uncertainty because of the measurement error in observed gene expressions.

3.1 Likelihood Function

Suppose that there are $m + 1$ readings taken on n spots for each combination of a slide and a dye. For a given gene i , we use $S_{i1}, \dots, S_{i(m+1)}$ to denote the $m + 1$ ordered signal measurements after background correction. Here S_{i1} is the gene expression measurement from the smallest (in median expression units) slide and $S_{i(m+1)}$ denotes the reading for gene i on the largest slide.

We assume that all readings measure the same quantity – actual gene expression – with error. Therefore, under suitable scaling the readings would be identically distributed. We assume that the strictly positive scaled readings can be represented by a Gamma distribution. The Gamma has support in the positive real line and depending on parameter values, exhibits noticeable skewness. Therefore, we can model the background corrected signals for each gene i across the $m + 1$ in the following way:

$$S_{i1} * \chi_1 = S'_{i1} \sim \Gamma(a, \psi_i) \quad (1)$$

$$S_{i2} * \chi_2 = S'_{i2} \sim \Gamma(a, \psi_i) \quad (2)$$

$$\vdots \quad \vdots \quad (3)$$

$$S_{i(m+1)} * \chi_{m+1} = S'_{i(m+1)} \sim \Gamma(a, \psi_i) \quad (4)$$

where $\chi_1, \dots, \chi_{m+1}$ are constant for all genes in a given slide and dye combination. This amounts to assuming that the changes in laser and sensor settings affect the amount of each spot's fluorescence equivalently.

As formulated, the model is not identified in that there is no way to estimate the parameters ψ_i directly. Thus, we do not attempt to estimate $\chi_1, \dots, \chi_{m+1}$ and instead focus on estimating $\theta_i = \chi_{m+1} \psi_i$ instead. We choose one of the $m + 1$ readings as a reference reading and scale all other readings to that level. By scaling all readings upwards to the highest one we are increasing the effective range of gene expression measurement. This does not limit the usefulness of the model in any way, because all measures of gene expression are relative and normalization is performed on the expression estimates.

We now have the following model:

$$\begin{aligned} S_{i1} * \delta_1 &= S'_{i1} \sim \Gamma(a, \theta_i) \\ S_{i2} * \delta_2 &= S'_{i2} \sim \Gamma(a, \theta_i) \\ &\vdots \quad \vdots \\ S_{im} * \delta_m &= S'_{im} \sim \Gamma(a, \theta_i) \\ S_{i(m+1)} &\sim \Gamma(a, \theta_i) \end{aligned} \quad (5)$$

where $\delta_1, \dots, \delta_m$ are constant for all genes in a given slide and dye combination. For notational convenience, let $\delta_{m+1} = 1$. The unknown parameters in this model are $a, \theta_1, \dots, \theta_n$, and $\delta_1, \dots, \delta_m$.

The mean of each of the Gamma distributions for the i th spot is a/θ_i . Within a classical framework, an estimate of expression level for the i th gene would be based on the corresponding mean or perhaps on a suitable

function of the $m+1$ means. These estimates still require normalization so that expressions observed for different slide/dye combinations can be compared.

3.2 Prior Distributions

We adopt a Bayesian approach to estimating the parameters in the model. In order to do so, we must complete the specification of the model by assigning prior distributions to each parameter. We restrict our attention to proper prior distributions to guarantee integrability of the posterior, and within the family of proper distributions we focus on the conjugate or semi-conjugate families to attempt to simplify computations wherever possible. If the prior distribution for the parameters is conjugate, then the posterior will have the same form as the likelihood function.

We assume that the scale parameters, $\theta_1, \dots, \theta_n$, arise from a common population distribution independently (a priori) of the scaling parameters $\delta_1, \dots, \delta_m$. Let

$$p(\theta_1, \dots, \theta_n, \delta_1, \dots, \delta_m) = p(\theta_1, \dots, \theta_n) \times p(\delta_1, \dots, \delta_m)$$

represent a joint prior distribution that for now will remain unspecified. We derive a posterior distribution for the vectors $\theta = (\theta_1, \dots, \theta_n)$ and $\delta = (\delta_1, \dots, \delta_m)$ and then determine the form of the prior distribution $p(\theta, \delta)$ that would be conjugate for the likelihood.

However, this distribution is difficult to interpret from a biological viewpoint and further, implies a prior dependency between θ and δ which we cannot justify. Thus, the conjugate prior option, while convenient from a mathematical viewpoint appears to be unsuitable from a biological viewpoint. We consider instead independent Gamma prior distributions for each of the $n + m$ parameters. Gamma distributions can be justified from a biological point of view because typically genes spotted on a slide exhibit low expression levels and only some of them exhibit high levels of expression. The Gamma distribution would appear to be an appropriate model for the population distribution because the expression values of the genes, estimated by a/θ_i , will be skewed. Thus

$$\theta_i \sim \Gamma(a_0, \nu)$$

for $i = 1, \dots, n$. The Gamma model may also be reasonable for the strictly positive scaling parameters, so that

$$\delta_j \sim \Gamma(\alpha_1, \alpha_2)$$

for $j = 1, \dots, m$. The joint Gamma prior has the form

$$p(\theta, \delta) \propto \prod_{i=1}^n \theta_i^{a_0} \prod_{i=1}^m \delta_j^{\alpha_1} e^{-\nu \sum_{i=1}^n \theta_i - \alpha_2 \sum_{j=1}^{m+1} \delta_j}. \quad (6)$$

The conditional posterior distributions of $\theta|\delta$ and $\delta|\theta$ are Gamma distributions under this prior, but the joint posterior of (θ, δ) is not. Therefore, the prior in (6) is a semi-conjugate prior distribution.

The hyperparameters in the model are $\eta = (a, a_0, \nu, \alpha_1, \alpha_2)$. We must either specify prior distributions for these hyperparameters or fix the parameters at some appropriate value. The hyperparameters α_1 and α_2 are both chosen to be 10 to create a relatively noninformative hyperprior on the δ 's. Specifying a value of the

other hyperparameters a , a_0 and ν , however, requires some thought since these parameters can have a significant effect on the estimates of expression levels.

One approach to obtaining values for hyperparameters is to find the values $(\hat{a}, \hat{a}_0, \hat{\nu})$ that maximize the marginal likelihood of the parameters (MMLEs, e.g., Carlin and Louis, 2003). The marginal likelihood $p(a, a_0, \nu | S)$ is obtained by integrating (δ, θ) out of the joint likelihood function as follows (the complete derivation of $p(a, a_0, \nu | S)$ is presented in the Appendix):

$$\begin{aligned}
p(a, a_0, \nu | S) &= \int \int p(a, a_0, \nu, \theta, \delta | S) d\delta d\theta \\
&= \Gamma(a)^{-n(m+1)} \Gamma(n(a-1) + \alpha_1)^m \prod_{i=1}^n \prod_{j=1}^{m+1} S_{ij}^{a-1} \\
&\quad \int \prod_{i=1}^n \theta_i^{a(m+1)+a_0-1} e^{-\nu \sum_{i=1}^n \theta_i} \prod_{j=1}^m \left(\sum_{i=1}^n \theta_i S_{ij} - \alpha_2 \right) d\theta. \tag{7}
\end{aligned}$$

This marginal distribution is not analytically tractable. However, we could integrate δ out analytically if instead of conditioning on S we were to derive the marginal likelihood given only the largest reading, $S_{\bullet(m+1)}$. Here, we have chosen to use the largest reading arbitrarily. Any reading could be used as the standard and the model would be adjusted to estimate $\theta = \chi_h \psi$. The subsequent normalization that expression estimates must undergo makes any choice of standard reading equivalent. In this case,

$$p(a, a_0, \nu | S_{\bullet(m+1)}) = \int \int p(a, a_0, \nu, \theta, \delta | S_{\bullet(m+1)}) d\delta d\theta \tag{8}$$

$$\propto \int \int p(S_{\bullet(m+1)} | a, a_0, \nu, \theta) p(\theta | a_0, \nu) p(\delta) d\delta d\theta \tag{9}$$

$$= \int p(S_{\bullet(m+1)} | a, a_0, \nu, \theta) p(\theta | a_0, \nu) d\theta \tag{10}$$

$$\begin{aligned}
&= \int \Gamma(a)^{-n} \prod_{i=1}^n \theta_i^a S_{i(m+1)}^{a-1} \exp(-\theta_i S_{i(m+1)}) \\
&\quad \Gamma(a_0)^{-n} \nu^{na_0} \prod_{i=1}^n \theta_i^{a_0-1} \exp(-\nu \theta_i) d\theta \tag{11}
\end{aligned}$$

$$\begin{aligned}
&\propto (\Gamma(a)\Gamma(a_0))^{-n} \prod_{i=1}^n S_{i(m+1)}^{a-1} \nu^{na_0} \\
&\quad \int \prod_{i=1}^n \theta_i^{a+a_0-1} \exp\left(-\sum_{i=1}^n \theta_i (S_{i(m+1)} + \nu)\right) d\theta \tag{12}
\end{aligned}$$

$$= (\Gamma(a)\Gamma(a_0))^{-n} \prod_{i=1}^n S_{i(m+1)}^{a-1} \nu^{na_0} \prod_{i=1}^n \Gamma(a+a_0) (S_{i(m+1)} + \nu)^{-a-a_0} \tag{13}$$

$$= \left(\frac{\Gamma(a+a_0)}{\Gamma(a)\Gamma(a_0)} \right)^n \prod_{i=1}^n S_{i(m+1)}^{a-1} \nu^{na_0} \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-a-a_0}. \tag{14}$$

The resulting expression can now be maximized with respect to a , a_0 and ν using standard nonlinear optimization techniques.

3.3 Posterior distributions

We use Markov chain Monte Carlo (MCMC) methods to approximate the marginal posterior distributions of each of the $m + n + 1$ parameters in the model. To do so, we first derive the full conditional distributions for each of the parameters:

$$\delta_1 | \eta, \delta_{-1}, \theta, S \sim \Gamma(na + \alpha_1, \sum_{i=1}^n \theta_i S_{i1} + \alpha_2) \quad (15)$$

\vdots \vdots

$$\delta_m | \eta, \delta_{-m}, \theta, S \sim \Gamma(na + \alpha_1, \sum_{i=1}^n \theta_i S_{im} + \alpha_2) \quad (16)$$

$$\theta_1 | \eta, \delta, \theta_{-1}, S \sim \Gamma((m+1)a + a_0, S_{1(m+1)} + \sum_{j=1}^m \delta_j S_{1j} + \nu) \quad (17)$$

\vdots \vdots

$$\theta_N | \eta, \delta, \theta_{-N}, S \sim \Gamma((m+1)a + a_0, S_{n(m+1)} + \sum_{j=1}^m \delta_j S_{nj} + \nu) \quad (18)$$

$$\nu | \eta, \delta, \theta, S \sim \Gamma(na_0 + \beta_1, \sum_{i=1}^n \theta_i + \beta_2) \quad (19)$$

Notice that all full conditional distributions have standard form, and thus the Gibbs sampler can be used to sequentially draw parameter values from the conditionals. A point estimate for the expression of the i th gene is the posterior mean of a/θ_i . These estimates may be subsequently used as the expression values for further normalization.

4 Maize Embryogenesis Experiment

The program of gene expression associated with somatic embryo maturation and germination was examined in callus cultures from a regeneration-proficient hybrid line of *Zea mays*, Hi II. 12,060 element maize cDNA microarrays were used to generate gene expression profiles from embryogenic calli induced to undergo embryo maturation and germination.

4.1 Improvements over Single Readings

The all readings hierarchical model and the one reading hierarchical model were fit to one dye channel of one slide. Examining particular genes, we can see the improvement caused by using all readings over using one reading either alone or in a hierarchical model.

For this slide, we get the posterior estimates $\delta_1 = 1.908$ and $\delta_2 = 1.145$ ($1.9S_{i1} \approx 1.1S_{i2} \approx S_{i3}$). The first gene has a posterior estimate of 198.5. The three measurements were (51.3, 211.0, 227.3) When using only the highest slide, the estimate was 255.0, and the three slide estimate is within the 95% posterior probability interval

Gene 1735, which has a censored high reading, has a posterior estimate of 73.8. The three measurements were (59.9, 77.2, 0). When using only the highest slide, the estimate was 6.2, and the three slide estimate is not within the 95% posterior probability interval.

The variance of those estimates using three slides were 265 and 43. This is much less than 1260 and 224 when one slide was used. Therefore, the posterior distributions of the expression estimates are much wider when only one reading is used. The difference in value is a result of the subjective definition of the quantity being estimated. The order of the expression values is reasonably conserved and normalization will scale different repetitions to be comparable.

5 Discussion

The use of the multiple readings generally taken in cDNA microarray experiments is not part of the techniques usually recommended [1]. We have shown that these readings should be used whenever they are available. They substantially reduce the average variance of expression estimates. They also improve the dynamic range of gene expression estimation.

The use of multiple scans taken at the same laser and sensor settings have been proposed to reduce expression estimate variability [3]. However, this works to improve spot homogeneity and remove artifacts which should be possible with effective segmentation and background cleaning methods. Using multiple scans with the same laser and sensor settings does not increase the dynamic range of the technology. It should produce similar reductions in estimate variance to methods using multiple scaled scans at different settings. Also, different setting scans are routine. Therefore, we recommend using multiple scans at varying settings.

References

- [1] Smyth, G.K., Yang, Y.H., and Speed, T. (2002). *Statistical Issues in cDNA Microarray Data Analysis. Functional Genomics: Methods and Protocols* Humana Press, Totowa, NJ.
- [2] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8(1), 37–52.
- [3] Romualdi, C., Trevisan, S., Celegato, B., Costa, G., and Lanfranchi, G. (2003). Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Research* 31(23) e149
- [4] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis* Chapman & Hall, London