Bayesian Markov Chain Monte Carlo and Restricted Maximum Likelihood Study of Gene Expression Patterns Across Time

Feng Hong[1], Sandra L. Rodriguez-Zas[1,2]
Department of Statistics, University of Illinois at Urbana Champaign[1]
Department of Animal Sciences, University of Illinois at Urbana Champaign[2]

Abstract

The performance of Restricted Maximum Likelihood (REML) and Bayesian Markov Chain approaches to study gene expression trends across time were compared. Measurements of performance included the consistent identification of cDNAs differentially expressed and not differentially expressed and estimates of changes in cDNA expression across age was investigated. The normalized observations were assumed to have a Gaussian distribution in both approaches and two sets of prior distributions for array and residual variances with different levels of information were considered in the Bayesian approach. One set of prior distributions were non-informative (uniform) while the other set of prior distributions (Log-Normal) were more informative and based on the distribution of the variances across multiple (all) cDNAs. The identification of differentially expressed cDNAs was based on a combination of maximum fold-change among any pair of ages and P value in the REML approach or Bayesian Factors (BF) in Bayesian approaches. A total of 437 cDNAs were declared differentially expressed based on P value $< 10^{-4}$ and maximum fold change between ages greater than 2 in the REML approach. Of the 437 cDNAs, 409 and 423 cDNAs had BF > 1800 and 216 (comparable to approximate P value $< 10^{-4}$ and $< 10^{-3}$, respectively) when non-informative prior distributions were used and 429 and 434 cDNAs had BF > 1800 and 216 respectively, when informative prior distributions were used. There results suggest that for relatively small microarray data sets comparable to that studied here, the use of information from multiple cDNAs improves the ability to detect differential expression. Out of 500 cDNAs not differentially expressed in REML (P values >0.1), 458 had BF <3.8 (comparable to approximate P value >0.1) when non-informative prior distributions were used. The correlations of the maximum fold change estimates from the REML and Bayesian non-informative approaches were 0.995, 0.996 and 0.9956 for the 437, 409 and 423 cDNAs previously characterized. The differences between the REML and Bayesian approach with non-informative prior may be due to the low information contained in the data and impact of the prior distribution on the posterior density estimates.

1 Introduction

Microarray technology provides the opportunity to measure the gene expression of thousands of genes simultaneously. However, for each gene, the number of measurements are often limited due to the limited number of arrays available. Typically the analysis is done on a by-gene basis and the microarray may have multiple sources of technical noise, which may limit the power of the analysis. The objective of this study was to explore the perforBayesian and REML analyses to overcome this performance limitation. The incorporation of priors in the Bayesian approach can overcome some of the limitations. The Bayesian Markov Chain Monte Carlo (MCMC) algorithms facilitate the evaluation of fitting more flexible models.

2 Material and Methods

2.1 Experimental Design and Data Preprocessing

The gene expression of Apis mellifera mellifera honeybees was measuredin a 20 cDNA microarrays experiment with a loop design. On each of 5 ages (day 0, 4, 8, 12 and 17 after adult emergence) three nurse bees were sampled, and on day 17 after emergence three forager bees were sampled. The expression of genes from individual brains was assessed using the double-spotted Apis mellifera brain 9K version 3.0 cDNA microarray using the protocols described by Whitfield et al. (2003), Grozinger et al. (2003) and Cash et al. (2005). On each array, there are 8887 reporter cDNAs and each cDNA has 2 duplicated spots.

The filtering and analysis procedures were conducted using R (R Development Core Team, 2006). Feature intensities were filtered when: the spots pertain to controls or other sequences (e.g. virus, suspected to be contaminated or present in high levels in hypopharyngeal glands) also excluded in Cash et al. (2005); and the spots were deemed of bad quality (and assigned a -100 flag) by the image analysis software (GenePix Pro 5.0; www.moleculardevices.com). After filtering, 7605 cDNAs were left for the analysis. The duplicated spots on the same microarray were then combined into one value, the average of the two spots when available or the value of a single spot remaining after filtering. The $\log_2$ intensity values were normalized using a lowess transformation (Wu *et al.* , 2002) and centered.

2.2 Bayesian Linear Model for Gene Expression Data

We use a linear mixed effect model to describe the $\log_2$ normalized cDNA expression measurements. Suppose y_{gijk} is the measurement corresponding to the gth cDNA, ith array, jth dye and kth age. The model for it is

$$y_{gijk} = mu_g + A_{gi} + D_{gj} + T_{gk} + e_{gijk},$$
$$e_{gijk} \sim N(0, \sigma^2_{err,g}),$$

where μ_g is cDNA specific overall mean, A_{gi} is the effect of the ith array,$D_{gj}$is the effect of the jth dye, and T_{gk} is effect of the kth age, and σ^2_{err} denotes the error variance. The age effects T_{gk} are of biological interests. The aim is to find those cDNAs whose expression changes during the life stages of honeybees. For each gene g, the hypothesis for age effects is tested, which is equivalent to choose between the reduced model

$$y_{gijk} \sim N(\mu_g + A_{gi} + D_{gj}, \sigma^2_{err,g}),$$

and the full model

$$y_{gijk} \sim N(\mu_g + A_{gi} + D_{gj} + T_{gk}, \sigma^2_{err,g}).$$

Bayes Factors were used to assess the differential expression across ages. The likelihood harmonic mean approach (Kass & Raftery, 1995) was used to estimate the Bayes Factors. cDNAs with Bayes Factors greater than 1800 and 216 were approximated to classical significance P values $< 10^{-4}$ and 10^{-3}, respectively.

The dye effects are considered fixed. The vague, nearly flat, noninformative prior distributions were used:

$$D_{gj} \sim N(0, 10^6).$$

For identifiability purposes, a sum-to-zero constraint was imposed on the dye effects.

The array effects are considered random and two level hierarchical prior distributions are used. In the first level the array effects are described with a Normal distribution centered at zero and with variance σ^2_A.

$$A_{gi} \mid \sigma_{A,g} \sim N(0, \sigma^2_{A,g}),$$

where σ_A is the hyper parameter which has its own prior.

For the error and array variances two sets of prior distributions were evaluated. One set of prior distributions consisted of vague uniform distributions on square root of the variances, encompassing a wide range of variances within the parameter space.

$$\sigma_{err,g} \sim U(0, 100),$$
$$\sigma_{A,g} \sim U(0, 100).$$

The other set of prior distributions have log-normal (LN) distributions:

$$\sigma^2_{err,g} \sim LN(\mu_1, \sigma^2_1),$$
$$\sigma^2_{A,g} \sim LN(\mu_2, \sigma^2_2).$$

We used Empirical Bayes approach by using point estimates of the location(μ_1, $\mu 2$) and dispersion (σ_1, σ_2) hyperparameters. One way is to obtain the estimated hyper parameters based on the REML estimates from all the cDNAs studied. In this way the information is borrowed from other cDNAs and more robust estimates of the variance components, less sensitive to the structure and information content of the data are used. The impact of these set of prior distributions on the identification of cDNAs with significant, borderline significant and non-significant differential expression across ages was studied.

The selection of prior distribution for age effects was more sophisticated because we are used Bayes factors to test the significance of the age effects. The usual noninformative flat prior distribution cannot be used because improper prior distributions would lead to unidentifiable Bayes factors, which could only be obtained up to a constant . The nearly flat vague prior may not be appropriate for the calculation of the Bayes factors either (Kass & Raftery, 1995). To use proper prior distributions while without giving subjective information other than what is given by data, we assumed the normal prior distributions for the age effects,

$$T_{gk} \sim N(\mu_{Tk}, \sigma^2_{Tk}),$$

and the hyperparameters μ_{Tk} and σ^2_{Tk} were estimated by using the estimates of the age effects obtained by noninformative prior distributions.

Although the noninformative nearly flat prior distributions for age effects are not appropriate for the calculation of BF, they are valid for the estimation of the age effects. It is possible that after the identification of the significant cDNAs, we can use the noninformative prior distributions to obtain the estimates of the age effects and compare them with the estimates obtained by the empirical Bayes method.

The Monte Carlo Markov Chain Gibbs sampler was used to draw samples from the conditional distributions of the unknown parameters (Normal for dye and age effects, Gamma for array and error precision) and obtain posterior density estimates of the parameters of interest. The Gibbs sampling was implemented in WinBUGS(Spiegelhalter et al 2003) and the posterior densities were based on a chain of length 10000 after removal of the first 5000 samples and all Markov chains were inspected for convergence.

The results from the Bayesian implementations were compared to equivalent linear mixed effects model evaluated in a Restricted Maximum Likelihood (REML) classical framework. Significance P-values $< 10^{-4}$ and maximum fold change between any two ages >2 were used to identify the cDNAs with differential expression across ages. The REML approach was implemented using MAANOVA (Wu et al. 2002).

3 Results

3.1 Test the Age effects

In REML, a total of 437 cDNAs exhibit significant variation in expression across ages using a threshold P-value $< 10^{-4}$ and maximum fold change between pairwise age comparisons greater than 2. Figure 1 presents the plot of the $\log_{10}$ P-values versus the $\log_2$ maximum fold change between ages and the horizontal and vertical lines mark the thresholds used to assess statistical significance. The 437 cDNAs with significant variation across ages are distributed in the upper left and right regions of the "volcano" plot and the labels associated with each cDNA denote the ages exhibiting the maximum change in expression levels. Labels 1 to 6 denote day 0, 4, 8, 12 and 17 nurse and day 17 forager ages, respectively. Differences between gene expression at the start and end of the maturity period considered (days 0 and 17) account for the vast majority of the significant differential expression observed.

In the Bayesian analysis using non-informative prior distributions, 2254 out of 7605 cDNAs had BF > 1800. The Bayesian analysis using non-informative prior confirmed 409 cDNAs with > 1800 (comparable to approximate P-value $< 10^{-4}$) out of the 437 detected in the REML analysis. Likewise, the Bayesian analysis using informative prior distributions for the residual and array variance confirmed 423 cDNAs with BF > 1800.

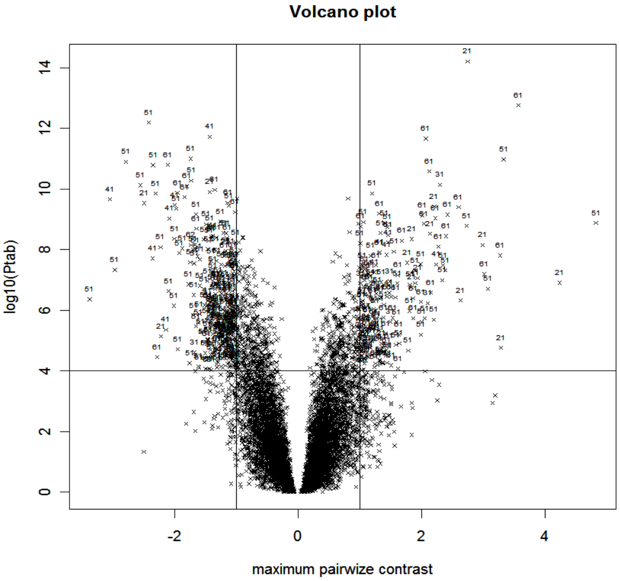


Figure 1: Plot of log$_1$0(P-values) versus log$_2$(maximum fold change between ages). The horizontal and vertical lines demark thresholds for statistical significance.

For a random sample of 500 non-significant cDNAs with P-value $> 10^{-1}$, 458 cDNAs had BF < 3.8(comparable to approximate P-value $> 10^{-1}$).

Figure 2 depicts the scatter plot of $\log_e$(BF) versus $\log_{10}$(P-value) for the 7605 cDNAs. There is a clear linear association between BF and P-values for the set of cDNAs with significant age effect that is not present in the non-significant set of cDNAs. Few cDNAs show non-significant BF values associated with significant REML P-values. In these cases, the impact of the prior on the posterior density estimates and the behavior of the Markov chain must be further evaluated.

3.2 The variance component estimates

The histograms (Fig. 3) of the log of the variance component estimates for all the cDNAs obtained by REML show that the distribution of these variance component estimates can be approximated by a log-normal distribution, and the location and dispersion hyperparameters in the log-normal distribution can be estimated by sample mean and variance of the log of the

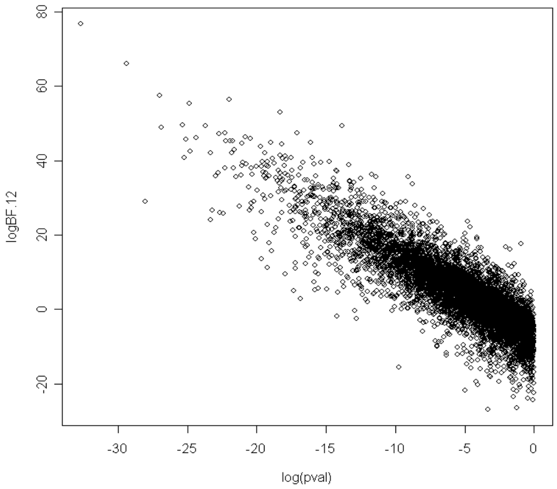


Figure 2: Plot of $log_e(BF)$ versus $log_{10}(P - value)$ for all the 7605 cDNAs

estimates. There were 391 cDNAs for which REML gives zero estimates for the array variance, and these estimates were excluded for calculating the hyperparamters. We used the approximate log-normal distributions as the informative prior for the variance of the array effects (log-normal(-2.7, 0.93)) and the variance of residual (log-normal(-3.56, 0.88))

Figure 4 presents the histogram of posterior distribution of the median array and residual standard deviations for the significant cDNAs corresponding to the non-informative and informative Bayesian analyses. The posterior distributions of the array and residual standard deviations from the informative Bayesian analysis using log-normal prior distributions are more concentrated in the middle values than the posterior distribution from non-informative uniform prior distributions. This difference is due to the impact of the information contained in the log-normal prior distributions that resulted in less extreme variance component estimates than with the uniform prior distributions.

3.3 The Age Effect Estimates

The expression profiles age effect for the significant cDNAs were of main interest. It is biologically meaningful to cluster the cDNAs with similar age effect profiles together and study the relationship of the cDNAs in the same cluster. We compare the estimates of the age effect (with sum-to-zero constraint) From Bayesian analysis and REML analysis for 423 significant cDNAs (Figure 5). It is noticeable that Bayesian and REML estimates are consistent, especially for the estimates

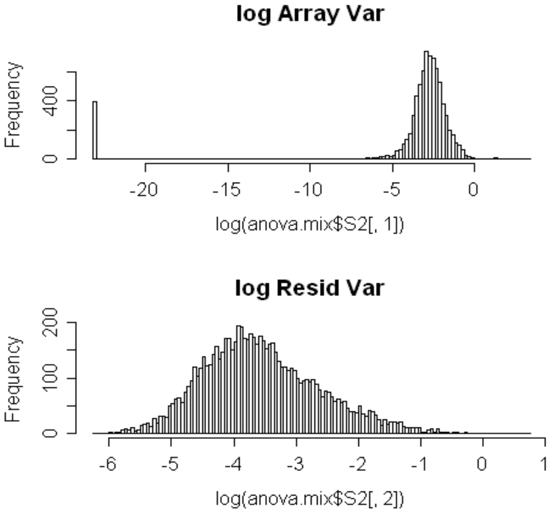
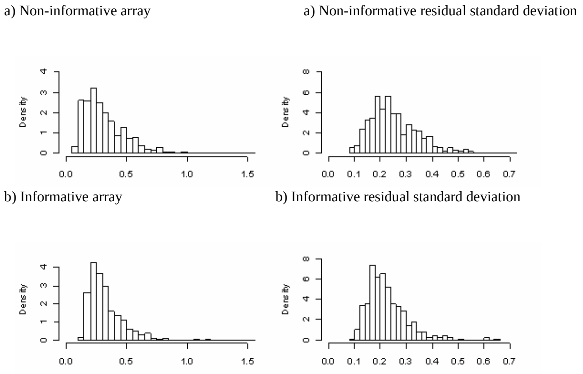


Figure 3: histograms of log array variance and log error variance estimates


a) Non-informative array

a) Non-informative residual standard deviation

b) Informative array

b) Informative residual standard deviation

Figure 4: Histograms of Bayesian median estimates of array and residual standard deviation using a) non-informative uniform and b) informative log-normal prior distributions.

for ages 0, 4, and 8. Although there are slight difference between two methods for the estimates of ages 12, 17 and forager age 17, for some cDNAs, overall the estimates for these ages are also consistent. The slightly lower consistency of the estimates in the final ages may be due to the inherently higher variability of the expressions at later stages of maturation.

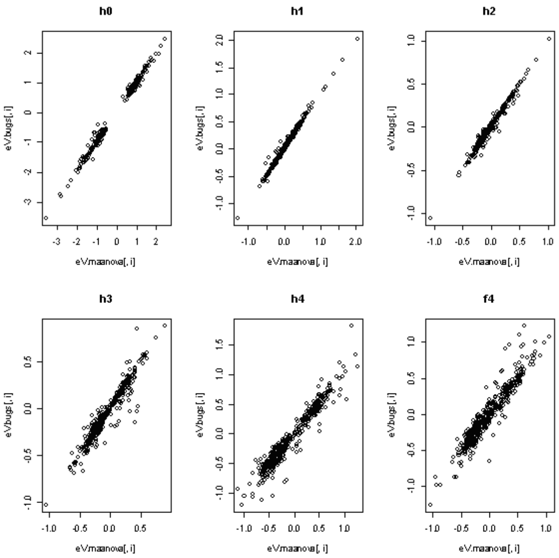


Figure 5: Comparing the age effect estimates for the 423 significant cDNAs by Bayesian method and REML method. y-axis indicates estimates given by bayesian method, and x-axis inicates the esimates given by REML.

We clustered the profiles of significantly differentially expressed cDNAs according to their Bayesian estimates (Fig. 6). Consensus kmeans method was used to assign each cDNA into one of the 8 groups. A total of 1000 starting points are random generated and in each iteration kmeans clustering was used to clustered the cDNAs into 8 groups. The cDNAs clustered into the same group over 60% of the time were assigned to the group. Otherwise the cDNAs were assigned to the "not-in-any-group" cluster. The study of functions of the cDNAs within the clusters support that cDNAs with similar neurological functions were assigned to the same cluster.

4 Acknowledgements

The support of NIH/NIGMS (Grant Number: 1R01GM068946-01) and of NSF/FIBR (Grant Number: 0425852) are greatly appreciated

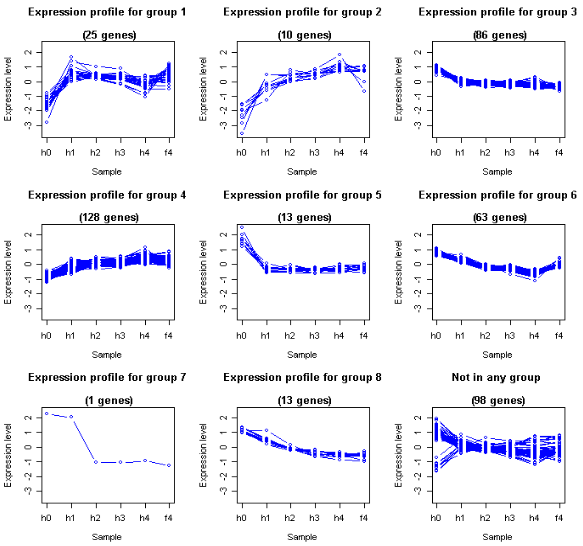


Figure 6: The kmeans clustering of the profiles of age effect estimates

References

Cash, A C, Whitfield, C W, Ismail, N, & Robinson, G E. 2005. Behavior and the limits of genomic plasticity: power and replicability in microarray analysis of honeybee brains. *Genes Brain Behav*, **4**(4), 267–71.

Grozinger, Christina M, Sharabash, Noura M, Whitfield, Charles W, & Robinson, Gene E. 2003. Pheromone-mediated gene expression in the honey bee brain. *Proc Natl Acad Sci U S A*, **100 Suppl 2**(NIL), 14519–25.

Kass, Robert E., & Raftery, Adrian E. 1995. Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Spiegelhalter, D, Thomas, A, Best, N, & Lunn, D. 2003. *WinBUGS User Manual version 1.4 [http://www.mrc-bsu.cam.ac.uk/bugs/]*.

Whitfield, Charles W, Cziko, Anne-Marie, & Robinson, Gene E. 2003. Gene expression profiles in the brain predict behavior in individual honey bees. *Science*, **302**(5643), 296–9.

Wu, H, Kerr, M K, Cui, X, & Churchill, G A. 2002. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments [http://www.jax.org/staff/churchill/labsite/pubs/Wu_maanova.pdf].