Bayesian hierarchical error model for analysis of gene expression data

HyungJun Cho* and Jae K. Lee

Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, University of Virginia School of Medicine, Hospital West Complex, Room 3181, P.O. Box 800717, Charlottesville, VA 22908-0717, USA

Received on June 10, 2003; revised on February 17, 2004; accepted on February 27, 2004

Advance Access publication March 25, 2004

ABSTRACT

Motivation: Analysis of genome-wide microarray data requires the estimation of a large number of genetic parameters for individual genes and their interaction expression patterns under multiple biological conditions. The sources of microarray error variability comprises various biological and experimental factors, such as biological and individual replication, sample preparation, hybridization and image processing. Moreover, the same gene often shows quite heterogeneous error variability under different biological and experimental conditions, which must be estimated separately for evaluating the statistical significance of differential expression patterns. Widely used linear modeling approaches are limited because they do not allow simultaneous modeling and inference on the large number of these genetic parameters and heterogeneous error components on different genes, different biological and experimental conditions, and varying intensity ranges in microarray data.

Results: We propose a Bayesian hierarchical error model (HEM) to overcome the above restrictions. HEM accounts for heterogeneous error variability in an oligonucleotide microarray experiment. The error variability is decomposed into two components (experimental and biological errors) when both biological and experimental replicates are available. Our HEM inference is based on Markov chain Monte Carlo to estimate a large number of parameters from a single-likelihood function for all genes. An F-like summary statistic is proposed to identify differentially expressed genes under multiple conditions based on the HEM estimation. The performance of HEM and its F-like statistic was examined with simulated data and two published microarray datasets—primate brain data and mouse B-cell development data. HEM was also compared with ANOVA using simulated data.

Availability: The software for the HEM is available from the authors upon request.

Contact: hcho@virginia.edu; jaeklee@virginia.edu

*To whom correspondence should be addressed.

INTRODUCTION

Microarray technology enables us to measure and examine expression patterns of a large number of genes. In microarray data heterogeneous biological and experimental error variability is observed among different genes, and even for the same gene under different experimental/biological conditions. However, the number of replicated arrays per condition is often small due to the high costs associated with microarray experiments and the limited number of biological samples. Despite such a small number of replicated arrays, it is important to estimate separately the effects of a large number of genes under multiple biological conditions in order to analyze microarray data successfully.

Fold-change approaches identify differential expression by selecting genes with, e.g. a 2- or greater-fold change in expression values regardless of variability. Such a fold-change discovery often results in the preference for genes with larger variation in low-level log-intensities (Lee, 2001). In order to evaluate the statistical significance of differentially expressed genes, the two-sample t-test is often used under two comparing conditions (see Jin et al., 2001); however, p-value adjustment procedures are required in the large-screening microarray studies to control the random chance that a large number of genes with no differential expression are falsely detected as significant. The Westfall-Young step-down method is proposed to tightly control the family-wise error rate (FWER) under a desired level, while the dependency structure of microarray genes is preserved by using a permutation resampling (Dudoit et al., 2002). The significance analysis of microarray (SAM); Tusher et al., 2001 procedure controls the false discovery rate (FDR), i.e. the expected proportion of false positives among all significant hypotheses.

FDR adjustment is also used for the local pooled error (LPE) test that greatly increases the statistical power by pooling variability of genes with similar expression levels when the number of replicates is small (Jain et al., 2003). Bayesian approaches are also used for differential expressions: Newton et al. (2001) use Gamma–Gamma and Gamma–Gamma–Bernoulli distributions for modeling

Several approaches are developed through consideration of a mixture distribution of two or more different distributional structures that depend on each gene’s (over-, under-) differential or non-differential expression status. Efron et al. (2001) and Efron and Tibshirani (2002) introduce a simple nonparametric mixture model for two populations. Lee et al. (2002) propose three-component normal mixture models for a summary statistic to identify differentially expressed genes between two conditions in both Bayesian and frequentist frameworks. Broet et al. (2002) propose a multiple-component normal mixture model for the difference between two conditions in a Bayesian framework, while Allison et al. (2002) propose a multiple-component Beta mixture model for $p$-values arising from testing the difference between two conditions. Newton et al. (2001) suggest that a semiparametric hierarchical mixture model is sensitive and flexible in detecting differential expression with heterogeneous variability, and Kendziorski et al. (2003) elegantly extend this to parametric mixture models for comparing multiple conditions.

ANOVA modeling approaches have been proposed both for two-channel cDNA microarray data (Kerr et al., 2000; Kerr and Churchill, 2001; Wolfinger et al., 2001) and for single-channel oligonucleotide data (Chu et al., 2002; Hsieh et al., 2003). These ANOVA approaches, however, are limited for simultaneously modeling and inferring the effects and their heterogeneous errors of tens of thousands of genes on microarrays. A considerable number of array replicates per condition is required for estimation of these genetic and variance parameters whereas most microarray studies are conducted with a limited number of replicates due to costly microarray experiments and limited amount of biological samples. Furthermore, a gene often has heterogeneous error variability under different conditions, which fact is not considered by these ANOVA approaches. Global ANOVA modeling that simultaneously estimates the effects of all genes, conditions, and their interactions in a single model is extremely difficult or computationally infeasible due to the large number of parameters to be estimated. In practice, ANOVA inference is performed as within-gene modeling that separately fits each ANOVA model one gene at a time. The error variability in this within-gene ANOVA modeling varies on different genes but is constant across the different conditions (and gene-condition combinations) of each gene (Chu et al., 2002; Hsieh et al., 2003). The ArrayTools of the NCI Biometry Research Branch, which has many useful tools for microarray data analysis, is also based on such within-gene ANOVA modeling for identifying microarray expression patterns under multiple conditions (http://linus.nci.nih.gov/BRB-ArrayTools.html). Note that these within-gene ANOVA approaches are computationally convenient to estimate heterogeneous variances of arrays and genes, but as mentioned in Kerr (2003), global ANOVA modeling is more desirable because it can borrow information across genes and conditions for estimation.

We consider a Bayesian hierarchical error model (HEM) to overcome these restrictions and to analyze simultaneously a large number of genes and heterogeneous error components in an oligonucleotide microarray experiment. In particular, when both biological and experimental replicates are available, HEM is constructed with two hierarchical layers: one for experiment error variability and the other for biological error variability; when one of either biological or experimental replicates is unavailable, HEM is reduced into an one-layer model. Estimation of a large number of genetic parameters in HEM is performed by Markov chain Monte Carlo (MCMC). An $F$-like summary statistic is also proposed to identify differentially expressed genes among multiple conditions based on the HEM estimates. The performance of HEM is examined with simulated data, primate brain data and mouse B-cell development data. HEM is also compared with a within-gene ANOVA modeling approach using simulated data.

**METHODS**

Hierarchical error model with replicates

Suppose that there are $G$ genes on each microarray chip at each of $C$ conditions (e.g. tissue types or samples). We observe the $l$-th (replicated) gene expression value $y_{i,j,k,l}$ of the $i$-th gene for a particular $k$-th individual with the $j$-th condition, where $i = 1, \ldots, G; j = 1, \ldots, C; k = 1, \ldots, m_{i,j}; l = 1, \ldots, n_{i,j,k}$. Assume that observations were log transformed with base 2 and normalized as described in Dudoit et al. (2002).

We consider a Bayesian HEM with two layers of error. The first layer of HEM is for experimental/instrumental error:

$$y_{i,j,k,l} \mid \{x_{i,j,k}, \sigma^2_{i,j,k,l}\} = x_{i,j,k} + e_{i,j,k,l}, \quad (1)$$

where $x_{i,j,k}$ is a gene expression intensity free from experimental error, and the experimental error effect $e_{i,j,k,l}$ is assumed to be $i.i.d. N(0, \sigma^2_{i,j,k,l})$. Note that the experimental error $e_{i,j,k,l}$ can vary for gene $i$, condition $j$, individual $k$ and replicate $l$ even though we here assume a constant variance; the extension to non-constant variances is discussed later. If both biological and experimental replicates are available, this error term can be estimated differently for each combination of gene and condition. In the next layer, expression intensity is decomposed into several additive effects of each gene and condition:

$$x_{i,j,k} \mid \{\mu, g_i, c_j, r_{i,j}, \sigma^2_{b_{i,j,k}}\} = \mu + g_i + c_j + r_{i,j} + b_{i,j,k}, \quad (2)$$

where $\mu$ is the parameter for the grand mean; $g_i$ and $c_j$ are the parameters for the gene and condition effects, respectively; $r_{i,j}$ is the parameter for the interaction effect of gene and condition; and $b_{i,j,k}$ is the error term for the biological variation.
assuming \(i.i.d. N(0, \sigma^2_{\text{b}_i})\). The parameter \(\sigma^2_{\text{b}_i}\) is allowed to be heterogeneous for each combination of gene \(i\) and condition \(j\) (but not for individual \(k\) and replicate \(l\)), i.e. the variance is gene-condition dependent.

After eliminating non-biological experimental variation, HEM estimates the main and interaction effects of gene and condition, as well as heterogeneous biological error variation. Computation for estimating a large number of parameters can be performed by MCMC as described in the next subsection.

**Computation**

For the two-layer hierarchical error model above, the joint probability of the observed and unobserved variables is as follows:

\[
Pr(y, x; \theta) = \prod_{i,j,k,l} \phi \left( \frac{y_{i,j,k,l} - x_{i,j,k,l}}{\sigma_e} \right) \times \prod_{i,j,k,l} \phi \left( \frac{x_{i,j,k,l} - \mu - g_i - c_j - r_{i,j}}{\sigma_{\text{b}_i}} \right),
\]

where \(\theta = (\mu, g, c, r, \sigma^2_{\mu}, \sigma^2_e, \sigma^2_{\text{b}_i})\) and \(\phi\) is the density function of the standard normal distribution.

The prior distributions are assumed to be uniform prior on \(\mu\) and normal priors on \(g_i, c_j\) and \(r_{i,j}\) with mean zero and variance \(\sigma^2_g, \sigma^2_c\) and \(\sigma^2_r\), respectively, for mathematical convenience. For variance parameters \(\sigma^2_{\text{b}_i}\) and \(\sigma^2_e\), we use gamma priors with parameters \((a_{\text{b}_i}, b_{\text{b}_i})\) and \((a_e, b_e)\), respectively. Our choices and further recommendation of hyperparameters for these prior specifications will be described in the Discussion section. The posterior distribution \(\pi(x, \theta | y)\) of the unobserved data \(x\) and the parameters \(\theta = (\mu, g, c, r, \sigma^2_{\mu}, \sigma^2_e, \sigma^2_{\text{b}_i})\), given the observed data \(y\), is proportional to

\[
Pr(y, x; \theta) \times \prod_i \phi \left( \frac{g_i}{\sigma_g} \right) \times \prod_j \phi \left( \frac{c_j}{\sigma_c} \right) \times \prod_{i,j} \phi \left( \frac{r_{i,j}}{\sigma_r} \right) \times \prod_{i,j} \Gamma \left( \frac{\sigma^2_{\text{b}_i}}{a_{\text{b}_i}}, b_{\text{b}_i} \right) \times \Gamma \left( \frac{\sigma^2_e}{a_e}, \frac{b_e}{\sigma^2_e} \right),
\]

where \(Pr(y, x; \theta)\) is the joint probability in (3) and \(\Gamma(\ast; \alpha, \beta)\) is the density function of a Gamma distribution with mean \(\alpha/\beta\) and variance \(\alpha/\beta^2\).

Direct estimation of these parameters is restricted because there are a large number of parameters and missing data in the complex model. The Gibbs sampling (Besag et al., 1995), a stochastic resampling technique, is utilized for computation and inference. This technique enables us to sample the parameters and missing data from the posterior distribution of a parameter or missing data one at a time, given the other parameters and missing data. The posterior conditional distributions derived from the above posterior distribution \(\pi(x, \theta | y)\) can be found in Appendix 1. The samples of parameters and missing data from the Gibbs sampler or MCMC are collected for inference on their posterior distributions.

**Summary statistic**

Estimates of parameters in the hierarchical error model are denoted by \(\tilde{\mu}, \tilde{g}_i, \tilde{c}_j, \tilde{r}_{i,j}, \tilde{\sigma}^2_{\text{b}_i}\) and \(\tilde{\sigma}^2_e\), since they are the averages of MCMC samples. It follows that an expression intensity estimate for gene \(i\) at condition \(j\) is \(\hat{\mu}_{i,j} = \tilde{\mu} + \tilde{g}_i + \tilde{c}_j + \tilde{r}_{i,j}\). Even though these MCMC estimates can be used for inference on each of the model parameters, one still needs a summary statistic to identify significant differential expression patterns of each gene across multiple conditions. In ANOVA modeling this is often evaluated by an \(F\)-statistic.

For our summary statistic, we also employ a similar form to ANOVA \(F\)-statistic utilizing estimates of parameters in the hierarchical error model as follows:

\[
F_i = \frac{\sum_{j=1}^C m_{i,j}(\tilde{\mu}_{i,j} - \bar{\mu}_i)^2}{M_i(\tilde{\sigma}^2_{\text{b}_i} + \tilde{\sigma}^2_e)},
\]

where \(\bar{\mu}_i = \frac{\sum_{j=1}^C \tilde{\mu}_{i,j}}{C}\) and \(M_i = \sum_{j=1}^C m_{i,j}\). Like the \(F\)-statistic of ANOVA, the numerator measures the difference of expression intensities between conditions and the denominator measures variation. However, when the assumption of constant variances for all conditions is violated, the \(F\)-statistic of ANOVA may falsely be dominated by the effect of a condition with larger variation because it standardizes the sum of squares of expression differences by the sum of variations. Our proposed statistic sums up the ratios between each differential expression (for each pairwise contrast) and its corresponding magnitude of variation. More significantly expressed genes have larger HEM \(F\)-scores; hence, we select genes with large HEM \(F\)-scores for further biological investigation. The performance of HEM \(F\)-score and ANOVA \(F\)-statistic on the simulated data is compared and summarized in the Results section.

Our gene selection can be performed based on a posterior distribution derived from all relevant posterior samples. An extreme posterior probability (PP) of a gene \(i^*\), \(PP(F_i \geq F_{i^*})\), can be computed based on the sampling distribution of this HEM \(F\)-score. Larger HEM \(F\)-scores correspond to smaller PP values; however, PP values are ranged from 0 to 1. For example, if 50 genes out of 12 000 have equal or greater HEM \(F\)-scores, the PP values are 0.0042.

**Hierarchical error model with no replicates**

One of the biological and experimental replicates is often unavailable in microarray experiments. In such microarray data, the biological and experimental errors are confounded. HEM is then reduced into a model with one layer as follows:

\[
y_{i,j,k} | \{\mu, g_i, c_j, r_{i,j}, \sigma^2_{\text{e}_{i,j}}\} = \mu + g_i + c_j + r_{i,j} + \epsilon_{i,j,k},
\]

where \(\epsilon_{i,j,k}\) is the error term for the biological and experimental error variation, assuming \(i.i.d. N(0, \sigma^2_{\text{e}_{i,j}})\). Note that
mi

numbers of biological replicates are
σ

row shows the plots of the estimated
replicates are
ni

observed variables is as follows:
model.
l
the

Fig. 1. Estimates of \( \mu_{i,j} \) and \( \sigma_{\nu_{i,j}}^2 \). The top row shows the plots of the estimated \( \mu_{i,j} \) versus the true \( \mu_{i,j} \) for each condition, and the bottom row shows the plots of the estimated \( \sigma_{\nu_{i,j}}^2 \) versus the estimated \( \mu_{i,j} \). The solid line represents the true relationship between \( \sigma_{\nu_{i,j}}^2 \) and \( \mu_{i,j} \). The numbers of biological replicates are \( m_{i,1} = 2, m_{i,2} = 5 \) and \( m_{i,3} = 10 \) for conditions 1, 2 and 3 respectively, and the numbers of experimental replicates are \( n_{i,j,k} = 2 \) for all \( i, j, k \).

For the one-layer HEM above, the joint probability of the observed data \( y \) is as follows:

\[
Pr(y, \theta) = \prod_{i,j,k} \phi \left( \frac{y_{i,j,k} - \mu - r_{i,j} - c_{j} - b_{i,j,k}}{\sigma_{\nu_{i,j}}} \right).
\]

(6)

where \( \theta = (\mu, g, c, r, \sigma_{\nu}^2) \). For variance parameter \( \sigma_{\nu_{i,j}}^{-2} \), we use gamma prior with parameters \( (\alpha_{\nu}, \beta_{\nu}) \). The posterior distribution \( \pi(\theta | y) \) of the the parameters \( \theta = (\mu, g, c, r, \sigma_{\nu}^2) \), given the observed data \( y \), is proportional to

\[
Pr(y, \theta) \times \prod_{i} \phi \left( \frac{\bar{y}_{i,k} - \mu_{i,k} - b_{i,j,k} + c_{j} - r_{i,j}}{\sigma_{\nu_{i,k}}} \right) \times \prod_{j} \phi \left( \frac{c_{j}}{\sigma_{c}} \right) \times \prod_{i,j} \phi \left( \frac{r_{i,j}}{\sigma_{r}} \right)
\]

\times \prod_{i,j} \Gamma(\sigma_{\nu_{i,j}}^{-2}, \alpha_{\nu}, \beta_{\nu}).

(7)

The posterior conditional distributions derived from the above posterior distribution \( \pi(\theta | y) \) can be found in Appendix 2. It follows that the summary statistics is

\[
F_i = \sum_{j=1}^{C} m_{i,j} \frac{(\bar{\mu}_{i,j} - \bar{\mu}_{i})^2}{M_{i} \sigma_{\nu_{i,j}}^2}.
\]

RESULTS

Simulated data

For investigating the performance of HEM estimation, we generated simulated data from the following linear model:

\( y_{i,j,k} = \mu_{i,j} + b_{i,j,k} + e_{i,j,k} \), where the parameter \( \mu_{i,j} \) is the true gene expression value of gene \( i \) at condition \( j \), and the error terms \( b_{i,j,k} \) and \( e_{i,j,k} \) are for biological variation and experimental variation, respectively, assuming that \( b_{i,j,k} \sim i.i.d.N(0, \sigma_{b_{i,j,k}}^2) \) and \( e_{i,j,k} \sim i.i.d.N(0, 0.1) \). The biological variation, \( b_{i,j,k} \), varies on different expression levels, which closely assimilates heterogeneous (gene-condition dependent) variances in practical microarray data (see the solid line in the bottom row of Fig. 1). The line interpolated approximately equal-spaced 100 expression values and corresponding biological variances obtained by applying the LPE baseline variance estimator (Jain et al., 2003) to a real microarray dataset.
Based on this relationship, biological variances corresponding to randomly chosen expression values were determined. For example, suppose a randomly selected expression value for a combination of gene 1 and condition 1 is $\mu_{1,1} = 7.1$. The corresponding biological variance is $\sigma^2 = 0.078$; thus, $y_{i,j,k,l} = \mu_{i,j} + \mu_{k,l} + \epsilon_{i,j,k,l}$, where $\mu_{i,j}$ and $\epsilon_{i,j,k,l}$ for all $i,j,k,l$ were randomly sampled from $N(0,0.078)$ and $N(0,0.1)$, respectively. We assumed that there were $G = 1000$ genes and $C = 3$ conditions. For individual and experimental replicates, $m_{i,1} = 2, m_{i,2} = 5, m_{i,3} = 10$ and $n_{i,j,k} = 2$ for all $i,j,k$.

It is shown that expression intensities were well estimated and the estimates for the condition with more chips were only slightly better (see the top row of Fig. 1). The bottom row of Figure 1 shows that estimation of heterogeneous biological variances improves as the number of individual replicates increased. This confirms that when the replicate number is small (e.g. $m_{i,j} = 2$), it is difficult to detect true variation and Bayesian estimation considerably depends on its prior specification of Gamma($\alpha$, $\beta$). Extension to non-constant parameters for the prior and empirical prior specification are discussed later.

For comparison of HEM and ANOVA, we generated another simulated data from the above model with $G = 1000$ and $C = 3$. For individual and experimental replicates, (1) $m_{i,j} = 2$ and $n_{i,j,k} = 2$ for all $i,j,k$ or (2) $m_{i,j} = 5$ and $n_{i,j,k} = 2$ for all $i,j,k$. Especially, in this case 50% of 1000 genes were generated with varying degrees of differential expression. For the differentially expressed genes (positives), the expression values were randomly selected from approximately equal-spaced 100 expression values, and the expression value of one condition differed from those of the other conditions. Because biological variances were determined according to the chosen expression values, at least two conditions had different expression values and biological variances. In contrast, for the indifferentially expressed genes (negatives) those of all the conditions were the same. Because global ANOVA modeling is computationally restricted for a large number of genes, we followed the within-gene ANOVA approach of Hsieh et al. (2003) without the effect of probes, i.e. $y_{i,j,k,l} = \alpha_j + I_k(\alpha_j) + \epsilon_{j,k,l}$ for gene $i$, where $\alpha_j$ is the fixed effect of condition, $I_k(\alpha_j)$ is the random effect of individual nested in condition and $\epsilon_{j,k,l}$ is the error term. We used the Proc Mixed procedure (SAS Institute Inc., 1999) to fit the above model one at a time for each of the 1000 genes. Note that even though the simulated data assumed gene-condition dependent variances to be close to practical microarray data, this simulation was more preferable to within-gene modeling because the simulated data were generated one gene at a time.

We computed false positive rates (FPR) and false negative rates (FNR) by increasing threshold values of HEM-$F$ and ANOVA-$F$, where FPR for a threshold value $c$ is the rate of negatives with $F$-values equal to or greater than $c$, and FNR is the rate of positives with $F$-values less than $c$. The ROC plot of 1−FNR versus 1−FPR (Fig. 2) shows the difference in the power curves (at each false positive error rate) of HEM and ANOVA approaches by varying their decision cutoff values (or critical values) of HEM $F$-like scores and ANOVA $F$-statistics, respectively. In this figure, it can be seen that the HEM power curves (at the same level of false positive error) are superior to those of ANOVA for both cases with two or five biological replicates. A similar result was obtained when one of either biological or experimental replicates was not available (data not shown).

**Fig. 2.** ROC curves from HEM (solid lines) and ANOVA (dotted lines) models with two and five replicated arrays. The horizontal axis is $1 − \text{FPR} = 1 - \Pr(\text{positive} \mid \text{array})$ and the vertical axis is $1 − \text{FNR} = 1 - \Pr(\text{negative} \mid \text{positive})$.

**Primate brain data**

Humans and chimpanzees share many identical genomic sequences, but may have different expression patterns among these genes relevant to their morphological, behavioral and cognitive differences. Enard et al. (2002) studied gene expression of humans and chimpanzees to identify intra- and interspecific variation in primate gene expression patterns using microarrays (Affymetrix U95A Gene Chip™). The frozen brains of three humans and three chimpanzees were used to take the postmortem tissue samples, and two independent tissue samples for each individual were taken, i.e. for HEM with two layers, $m_{i,1} = 3, m_{i,2} = 3$ and $n_{i,j,k} = 2$, where $i = 1, \ldots, 12600, j = 1, 2, k = 1, \ldots, m_{i,j}$ and $l = 1, \ldots, n_{i,j,k}$. We applied the two-layer HEM to the data and found that many well-known genes were identified with large HEM $F$-scores. For identifying differentially...
expressed genes, we computed the extreme posterior probabilities (PP) of the HEM $F$-scores using its (posterior) sampling distribution. It was found that many well-known genes had large HEM $F$-scores (and small PP values). For example, gene 32798_at with the largest $F$-score is the human class mu GSTM3 glutathione transferase gene; the glutathione transferases are a family of phase II detoxication enzymes catalyzing conjugation reactions between glutathione and electrophilic compounds (Patkovsky et al., 1999). Gene 41155_at with the second largest $F$-score is involved in the encoding of a cadherin-associated protein and determination of its primary structure and chromosomal localization (Furukawa et al., 1994); we listed the top five genes with the largest $F$-scores, i.e. the smallest posterior probabilities (Table 1).

**Mouse B-cell development data**

We now demonstrate the use of one-layer HEM with the mouse B-cell development data set, which consists of gene expression of the five consecutive stages (pre-B1, large pre-B2, small pre-B2, immature B and mature B cells) of mouse B-cell development. The data were obtained with high-density oligonucleotide arrays, Affymetrix Mu11k Gene Chip™, from flow-cytometrically purified cells. Each of six sample replicates for pre-B1 cell and each of five sample replicates for the other conditions was hybridized on a chip, but there was no replicate for an identical sample condition (i.e. $m_{11} = 6$, $m_{12} = m_{13} = m_{14} = m_{15} = 5$). We estimated expression intensities of genes at each stage and discovered differentially expressed genes among the five consecutive stages. Out of 13 027 probe sets, we examined particularly the 18 target genes that were selected by a non-parametric test in Hoffmann et al. (2002).

Table 2 and Figure 3. The plots show various expression patterns; e.g. $X_{68670_s_at}$ in Table 2 had the largest $F$-score among all the genes and thus had the smallest PP. Figure 3 shows that the first stage (pre-B1) of the gene was extremely over-expressed compared with the other stages. On the other hand, gene $m_{62553_s_at}$ had similar expression estimates and large variances and thus had a small $F$-score of 0.56 and a large PP of 0.1021, as seen in Table 2 and Figure 3.

For comparison, we also performed the SAM analysis (Tusher et al., 2001), developed for the significance test of gene expression under two or more conditions. SAM scores and ranks of the known targets are summarized in Table 2. The SAM results were consistent with the HEM results, but different from those reported in Hoffmann et al. (2002). The four genes ($X_{13450_s_at}$, $X_{20327_s_at}$, $B_{32798_at}$, and $m_{62553_s_at}$) at the bottom were found to have small SAM scores, and they also had small HEM $F$-scores. In fact, as shown in Figure 3, their mean expressions had very weak or no differential expression patterns although they had small $p$-values by the Kruskal–Wallis test, a non-parametric test for multiple conditions, in Hoffmann et al. (2002). Note that we used version 5.0 of Affymetrix software (Microarray Suite) to obtain intensity values instead of version 4.0 used in Hoffmann et al. (2002).

**DISCUSSION**

Each gene on microarray data often presents heterogeneous biological variability on different biological conditions. In addition to such biological variability, array instrumental variability, which is also found to be heterogeneous on varying intensity ranges, occurs from several experimental sources. If the experimental error can be distinguished and separately estimated from the biological error, we can make better inferences for the analysis of microarray expression patterns. Standard linear modeling approaches, such as split plot designs, may capture multiple layers of error. However, these are still limited to the within-gene modeling and inference (that fits each separate model for one gene at a time) because of the computational difficulty in the global ANOVA modeling (that simultaneously fits the genetic parameters of all genes under a single likelihood function).

We here propose a Bayesian HEM that simultaneously models and infers the effects of all genes and their interactions and heterogeneous variances for different combinations of genes and conditions. HEM can efficiently estimate this large number of model parameters utilizing Bayesian hierarchical modeling and MCMC techniques. HEM also efficiently factors out biological variability from experimental errors.
Fig. 3. Expression intensity estimates on the scatter plots of expression values (with log base 2) against stages for the target genes in mouse B cell development. The solid line links the HEM estimates.
are often eliminated or are imputed based on the handling of missing data. Genes with missing values are frequently analyzable; however, the combination of these values may be more useful in identifying the biological and experimental error components to identify differentially expressed genes with improved biological relevance. For example, Ibrahim et al. (2002) also use a two-layer Bayesian model, but its first layer is considered for a threshold of small expression values and the second layer for error variability of mean expression; thus, these errors are more relevant to preprocessing issues of microarray data rather than different sources of error components. On the contrary, HEM decomposes total expression variability into experimental and biological error components, together with the additive effects of gene, condition and interaction.

In this study, we have attempted to decompose only two error components. HEM, however, can easily be generalized to the cases with more than two layers of error. For example, in addition to an experimental error component, there are (random) subgroups within each biological group, such as different time points of biological stimuli. In such a case, if replicated arrays are provided for these components, we may have more than one biological error component, e.g. one for the biological group and the other for the time point.

The prior distributions are assumed to be $g_i \sim N(0, \sigma^2_i)$, $c_j \sim N(0, \sigma^2_j)$, $r_{i,j} \sim N(0, \sigma^2_{r_{i,j}})$, $\sigma^{2}_{b_{i,j}} \sim \Gamma(\alpha_b, \beta_b)$ and $\sigma^{2}_{e^{2}} \sim \Gamma(\alpha_e, \beta_e)$ for the one-layer HEM. In this study, we used the following hyperparameter specifications: $\sigma^2_g = \sigma^2_c = \sigma^2_1 = 1$, $\alpha_b = \alpha_e = 1$ and $\beta_b = \beta_e = 0.5$ (or $\alpha_e = 1, \beta_e = 0.5$). These hyperparameter values were chosen based on the observation that variances of log-expression values are less than 2.0 in most practical microarray data sets. However, as can be seen in the Appendix, the effects of these prior specifications are weak on their corresponding posterior distributions if informative observed data are available for estimation of each parameter. To improve the convergence rate of MCMC runs, our software program uses initial values as follows: $x_{i,j,k} = \sum_y y_{i,j,k} / n_{i,j,k}$, $\mu = \bar{y} = \sum_{i,j,k} y_{i,j,k} / \sum_{i,j,k} n_{i,j,k}$, $r_{i,j} = \sum_k x_{i,j,k} / n_{i,j,k} - \bar{y}$, $g_i = \sum_{j} r_{i,j} / C - \bar{y}$, $c_j = \sum_{i} r_{i,j} / G - \bar{y}$, $\sigma^{2}_{b_{i,j}} = 1 / \Gamma(1,0.5)$ and $\sigma^2_e^2 = 1 / \Gamma(1,0.5)$. With such reasonable initial values, our MCMC samples converged quickly within 300–1500 runs, so that a moderate number of MCMC samples were discarded for our MCMC estimation; we conservatively used 1500 burn-in iterations in this study. In order to obtain accurate estimates for this evaluation. This would first require generalizing the concept of FDR to the HEM $F$-like statistic under the Bayesian paradigm. Then, if a null distribution of microarray data (i.e. array data without differential expression) can be generated by a method such as simulation or resampling, we can calculate FDR for the HEM $F$-like statistic at its varying cutoff values. This study based on various resampling techniques is currently in progress.

Our HEM is one of the emerging Bayesian hierarchical modeling approaches in microarray data analysis; however, it differs from the others in that it explicitly decomposes biological and experimental error components to identify differentially expressed genes with improved biological relevance. For example, Ibrahim et al. (2002) also use a two-layer Bayesian model, but its first layer is considered for a threshold of small expression values and the second layer for error variability of mean expression; thus, these errors are more relevant to preprocessing issues of microarray data rather than different sources of error components. On the contrary, HEM decomposes total expression variability into experimental and biological error components, together with the additive effects of gene, condition and interaction.

We believe that the main advantage of our Bayesian HEM approach is to provide more accurate estimation of biological and experimental errors by utilizing information from other genes and conditions, which consequently provides significantly improved statistical inference for the discovery of differential expression patterns. Another advantage of HEM is the handling of missing data. Genes with missing values are often eliminated or are imputed based on ad hoc methods, e.g. replacing a missing value with the mean of other observed values. HEM handles missing values naturally by randomly sampling them from their posterior distributions as shown in Appendix 1.

Using the MCMC samples of HEM, we introduce an $F$-like summary statistic and its extreme posterior probability to identify significantly differentially expressed genes among multiple biological and experimental conditions in a microarray experiment. Statistical significance of this extreme posterior probability is conceptually different from that of a classical $p$-value and can be viewed as a relative measure based on the empirical posterior distribution (of tens of thousands of genes); thus, classical family-wise error correction is not directly relevant in this case. Like the SAM procedure, however, the random chance of false positives among tens of thousands of candidate genes, the so-called, multiple comparisons issue, may also need to be considered variability, when both biological and experimental replicates are available. HEM differs from the ANOVA model in that it constructs a complete likelihood function for all genes and conditions whereas the latter uses a likelihood function for each gene at a time.

### Table 2. Mouse B-cell development data; target genes selected in Hoffman et al. (2002)

<table>
<thead>
<tr>
<th>Probe set ID</th>
<th>HEM</th>
<th>SAM</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>X68670_s_at</td>
<td>28.04</td>
<td>0.0001</td>
<td>0.943</td>
</tr>
<tr>
<td>V008002_s_at</td>
<td>15.02</td>
<td>0.0014</td>
<td>0.685</td>
</tr>
<tr>
<td>Msa.803.0_s_at</td>
<td>11.29</td>
<td>0.0025</td>
<td>0.565</td>
</tr>
<tr>
<td>xo5556_s_at</td>
<td>9.90</td>
<td>0.0034</td>
<td>0.560</td>
</tr>
<tr>
<td>Msa.2379.0_at</td>
<td>9.87</td>
<td>0.0035</td>
<td>0.567</td>
</tr>
<tr>
<td>m25324_s_at</td>
<td>8.85</td>
<td>0.0043</td>
<td>0.517</td>
</tr>
<tr>
<td>M29475_s_at</td>
<td>7.74</td>
<td>0.0058</td>
<td>0.675</td>
</tr>
<tr>
<td>m94351_at</td>
<td>5.38</td>
<td>0.0105</td>
<td>0.587</td>
</tr>
<tr>
<td>X04123_s_at</td>
<td>4.87</td>
<td>0.0117</td>
<td>0.408</td>
</tr>
<tr>
<td>J03857_s_at</td>
<td>2.56</td>
<td>0.0246</td>
<td>0.337</td>
</tr>
<tr>
<td>V00790_at</td>
<td>2.46</td>
<td>0.0256</td>
<td>0.264</td>
</tr>
<tr>
<td>y00864_s_at</td>
<td>1.98</td>
<td>0.0321</td>
<td>0.267</td>
</tr>
<tr>
<td>m83312_s_at</td>
<td>0.92</td>
<td>0.0637</td>
<td>0.271</td>
</tr>
<tr>
<td>M64796_s_at</td>
<td>0.56</td>
<td>0.1021</td>
<td>0.117</td>
</tr>
<tr>
<td>x05556_s_at</td>
<td>0.63</td>
<td>0.0910</td>
<td>0.117</td>
</tr>
<tr>
<td>Msa.2376.0_at</td>
<td>0.63</td>
<td>0.0910</td>
<td>0.117</td>
</tr>
<tr>
<td>m25553_s_at</td>
<td>0.56</td>
<td>0.1021</td>
<td>0.117</td>
</tr>
</tbody>
</table>
and desired confidence levels of all HEM parameters, we also evaluated the sample size of a MCMC sample adjusted for its dependency (the so-called effective sample size), which is equivalent to the sample size of its corresponding independent sample (Geyer, 1992). In this paper, we found that using 4000 iterations (after 1500 burn-ins), we obtained at least 100 effective sample size for all MCMC samples of the model parameters.

For mathematical convenience we here used normal, gamma or uniform priors for the model parameters. Prior specification can be defined differently. For example, biological variance parameter, $\sigma_{b_j}^2$, was experimented with using a gamma prior of two parameters ($\alpha_{b_j}, \beta_{b_j}$) and these prior parameters were again specified with their hyperpriors. This modification is more flexible in accommodating heterogeneous variances and is less dependent on the choice of prior parameters. We found that this hyper prior specification provided a similar result to the results in our current applications. Another promising prior specification is an empirical Bayes approach, which has also been successfully applied in a number of recent microarray studies (Efron et al., 2001; Efron and Tibshirani, 2002; Newton and Kendziorski, 2003). Assuming heterogeneity of experimental error variation, i.e. $e_{i,j,k,l} \sim N(0, \sigma_{b_j}^2)$, we can empirically estimate the variance parameter by using a baseline variance estimation method such as LPE (Jain et al., 2003). This empirical Bayes study is in progress.

ACKNOWLEDGEMENT

We thank Drs Paabo and Hoffmann for providing us their primate and mouse array data, and the reviewers for their helpful comments and suggestions. This study was supported by the American Cancer Society grant RSG-02-182-01-MGO.

REFERENCES


Bayesian hierarchical error model

APPENDIX 1
Posterior conditional distributions for HEM with replicates

The posterior conditional distributions, which can be derived from the posterior distribution $\pi(x, \theta | y)$ of the parameters and the unobserved data for the hierarchical error model, are as follows:

$$\pi(\mu | \text{rest}) = N \left[ \sum_{i,j} \frac{\xi_{i,j,k} - g_i - c_j - r_{i,j}}{\Lambda \sigma^2_{b_{i,j}}}, \Lambda^{-1} \right],$$

$$\pi(g_i | \text{rest}) = N \left[ \sum_{j} \frac{\xi_{i,j,k} - g_i - c_j - r_{i,j}}{\sigma^2_g + \Lambda \sigma^2_{b_{i,j}}}, \left(\frac{1}{\sigma^2_g} + \Lambda^{-1}\right) \right].$$

$$\pi(c_j | \text{rest}) = N \left[ \sum_{j} \frac{\xi_{i,j,k} - g_i - c_j - r_{i,j}}{\sigma^2_c + \Lambda \sigma^2_{b_{i,j}}}, \left(\frac{1}{\sigma^2_c} + \Lambda^{-1}\right) \right].$$

$$\pi(r_{i,j} | \text{rest}) = N \left[ \frac{\sigma^2}{\sigma^2 + \frac{1}{m_{i,j}}} \times \sum_{j} \frac{\xi_{i,j,k} - g_i - c_j - r_{i,j}}{m_{i,j}}, \left(\frac{1}{\sigma^2} + \frac{m_{i,j}}{\sigma^2_{b_{i,j}}}ight)^{-1} \right],$$

$$\pi(x_{i,j,k} | \text{rest}) = \begin{cases} N \left[ \frac{\sigma^2_{b_{i,j}}}{\sigma^2_{b_{i,j}} + \frac{1}{n_{i,j,k} + \sigma^2_{b_{i,j}}}} \right], & \text{if } n_{i,j,k} > 0, \\ N \left( \mu + g_i + c_j + r_{i,j}, \sigma^2_{b_{i,j}} \right), & \text{if } n_{i,j,k} = 0, \end{cases}$$

$$\pi(\sigma^{-2}_{b_{i,j}} | \text{rest}) = \text{Gamma} \left[ \alpha_b + \frac{m_{i,j}}{2}, \beta_b + \frac{\sum_{k} \left( \xi_{i,j,k} - g_i - c_j - r_{i,j} \right)^2}{2} \right],$$

$$\pi(\sigma^{-2}_e | \text{rest}) = \text{Gamma} \left[ \alpha_e + \frac{\sum_{i,j,k} n_{i,j,k}}{2}, \beta_e + \frac{\sum_{i,j,k} \left( y_{i,j,k} - x_{i,j,k} \right)^2}{2} \right],$$

where rest indicates the parameters and the unobserved data except for that of interest, $\Lambda = \sum_{i,j} (m_{i,j} / \sigma^2_{b_{i,j}})$, and Gamma$(\alpha, \beta)$ represents a gamma distribution with mean $\alpha / \beta$ and variance $\alpha / \beta^2$.

APPENDIX 2
Posterior conditional distributions for HEM with no replicates

The posterior conditional distributions, which can be derived from the posterior distribution $\pi(\theta | y)$ of the parameters for the hierarchical error model, are as follows:

$$\pi(\mu | \text{rest}) = N \left[ \sum_{i,j} \frac{y_{i,j,k} - g_i - c_j - r_{i,j}}{\Lambda \sigma^2_{b_{i,j}}}, \Lambda^{-1} \right],$$

$$\pi(g_i | \text{rest}) = N \left[ \sum_{j} \frac{y_{i,j,k} - g_i - c_j - r_{i,j}}{(\sigma^2_g + \Lambda \sigma^2_{b_{i,j}})^{-1}}, \left(\frac{1}{\sigma^2_g} + \Lambda^{-1}\right) \right].$$

$$\pi(c_j | \text{rest}) = N \left[ \sum_{j} \frac{y_{i,j,k} - g_i - c_j - r_{i,j}}{(\sigma^2_c + \Lambda \sigma^2_{b_{i,j}})^{-1}}, \left(\frac{1}{\sigma^2_c} + \Lambda^{-1}\right) \right].$$

$$\pi(r_{i,j} | \text{rest}) = N \left[ \frac{\sigma^2}{\sigma^2 + \frac{1}{m_{i,j}}} \sum_{j} \frac{y_{i,j,k} - g_i - c_j - r_{i,j}}{m_{i,j}}, \left(\frac{1}{\sigma^2} + \frac{m_{i,j}}{\sigma^2_{b_{i,j}}}ight)^{-1} \right],$$

$$\pi(\sigma^{-2}_e | \text{rest}) = \text{Gamma} \left[ \alpha_e + \frac{m_{i,j}}{2}, \beta_e + \frac{\sum_{i,j,k} \left( y_{i,j,k} - x_{i,j,k} \right)^2}{2} \right].$$

where rest indicates the parameters except for that of interest, $\Lambda = \sum_{i,j} (m_{i,j} / \sigma^2_{b_{i,j}})$, and Gamma$(\alpha, \beta)$ represents a gamma distribution with mean $\alpha / \beta$ and variance $\alpha / \beta^2$. 

2025