Lee, ML, Kuo, FC, Whitmore, GA, and Sklar, J (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 97: 9834-9839.

Littell, RC, Milliken, GA, Stroup, WW, and Wolfinger, RD (1996). SAS system for mixed models (Cary, NC, SAS institute Inc.,).

Lönnstedt, I, and Speed, T (2002). Replicated Microarray Data. Statistica Sinica 12: 31-46.

Pan, W, Lin, J, and Le, CT (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol 3: research0022.

Pritchard, CC, Hsu, L, Delrow, J, and Nelson, PS (2001). Project normal: defining normal variance in mouse gene expression. Proc Natl Acad Sci U S A 98: 13266-13271.

Quackenbush, J (2001). Computational analysis of microarray data. Nat Rev Genet 2: 418-427.

Searle, SR, Casella, G, and McCulloch, CE (1992). Variance components, John Wiley and sons, Inc.).

Simon, R, Radmacher, MD, and Dobbin, K (2002). Designs of studies using DNA microarrays. Genet Epidemiol 23: 21-36.

Storey, J (2002). A direct approach to false discovery rates. J R Statist Soc B 64: 479-498.

Witkovsky, V (2002). MATLAB algroithm mixed.m for solving Henderson's mixed model equations. http://www.mathpreprints.com

Wolfinger, RD, Gibson, G, Wolfinger, ED, Bennett, L, Hamadeh, H, Bushel, P, Afshari, C, and Paules, RS (2001). Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8: 625-637.

Yang, YH, and Speed, T (2002). Design issues for cDNA microarray experiments. Nat Rev Genet 3: 579-588.

Zien, A, Fluck, J, Zimmer, R, and Lengauer, T (2002). Microarrays: how many do you need? Proc. RECOMB'02. 321-330.

IN: METHODS OF MICROARRAY DATA ANALYSIS III
EDITED BY KIMBERLY F. JOHNSON AND SIMON M. LEE
KLUWER ACADEMIC PUBLISHERS, BOSTON
2003
PP. 155 – 172.

10

BAYESIAN CHARACTERIZATION OF NATURAL VARIATION IN GENE EXPRESSION

Madhuchhanda Bhattacharjee[1], Colin Pritchard[2], Mikko J. Sillanpää[1] and Elja Arjas[1]

[1]Rolf Nevanlinna Institute, University of Helsinki, Finland [2]Division of Human Biology, Fred Hutchinson Cancer Research Centre, Seattle, WA.

Abstract: For gene expression data we propose a hierarchical Bayesian method of analysis using latent variables, wherein we have combined normalization and classification in a single framework. The uncertainty associated with classification for each gene can also be estimated based on the posterior distributions of the latent variables applied. The proposed models are implemented using the MCMC algorithm.

Key words: Bayesian latent class models, gene expression data, MCMC.

1. INTRODUCTION

We present a new latent-variable-based Bayesian clustering method for classifying genes into categories of interest. The approach is integrated in the sense that normalization and classification can be carried out jointly. This is done along with estimation of uncertainty that makes it unnecessary to test a large number of hypotheses. Possible distortion in measuring the actual expression level due to factors like environmental and experimental conditions, dye, etc. are incorporated into the normalization part of the model. The observed expression is treated as a "black box" for the different effects, which are considered jointly in a nested common structure. The adjusted expression ratios are then classified into different categories of interest. The approach is very general in the sense that it is easily customizable for different needs and can be modified with the availability of additional information.

Preliminary and extended versions of the model were applied to the expression data provided by Pritchard *et al.* [2001]. The classification categories of interest are variational categories of genes in normal circumstances. Our findings support the hypothesis that, apart from the fact that there are several sources of variation affecting the observed expression of the genes, some genes by nature exhibit highly varied expression.

Several other Bayesian approaches have been presented recently for microarray data analysis [Medvedovic, 2000; Keller *et al.*, 2000; Long *et al.*, 2001; Baldi and Long 2001; Newton *et al.*, 2001; Ibrahim *et al.*, 2002; Dror *et al.*, 2002; Parmigiani *et al.*, 2002]. However, our approach differs substantially from the others in several ways. For example, we consider model-based normalization of data, whereas normalization is usually considered as a separate procedure without accounting for its effect on subsequent classification. Also, given the unusual nature of the data considered here, we have primarily considered classification with respect to variance whereas most microarray data analysis concentrates on classification with respect to the central tendency of the (log) expression ratios for a gene.

2. DATA

The data contained median foreground and background intensities for 5776 spots from experimental and reference samples taken from 3 organs of 6 mice each applied with 2 dyes and 2 replicates. This resulted in more than 1.5 million data points. Of the 5776 spots on the arrays 5401 were mouse genes. The differences of foreground and background intensities in the experimental and reference probes were treated as observed intensities. On several occasions the resulting intensities turned out to be negative. In absence of further clarification for such measured intensities, these were treated as missing data. Modeling was carried out after taking the natural logarithm of the experimental and reference probe intensity ratio.

3. PRELIMINARY MODEL (MODEL A)

Given that both the experimental and the reference sample are to represent the normal expression for that gene, the observed variations in the log ratio of the corresponding measured intensities can then be attributed to different and possibly nested sources, such as, mouse or dye or replicate. Without further subdividing the possible sources of variation, we assume that such factors affect all the genes and accordingly adjust the observed

expression log ratios by an effect for each organ and each of the 24 arrays, denoted by μ_{oj}, $o = K$ (kidney), L (liver), and T (testis), $j=1,\ldots,24$.

The adjusted data were then inspected for possible variation still remaining (if any) exhibited by the genes. It is anticipated that the genes may naturally behave differently in different organs, e.g. by varying highly in one organ but not in another. Accordingly each gene was classified independently for each organ with respect to its corresponding residual variance. The adjusted log-ratios were classified for the unknown (possibly natural) variation of the genes for each organ into three different categories. We assume three latent variance classes with (unknown) ordered variances $\sigma_1^2 > \sigma_2^2 > \sigma_3^2$. For each gene i and for each organ o, let c_{io} indicate its variance-class membership for that organ. We assume that c_{io} takes a value in range $\{1,2,3\}$ with probabilities $\{\lambda_1, \lambda_2, \lambda_3\}$. The modeling was actually carried out using corresponding precision parameters (i.e. inverse of variance) $\tau_1 < \tau_2 < \tau_3$. It may be mentioned here that it is also possible to consider large or even an infinite number of variance classes through Bayesian infinite mixture, but for practical implementation and better interpretability we restrict our models to a finite mixture of latent classes (see Section 7 for a discussion).

Following the above notation, for the i^{th} gene, o^{th} organ and j^{th} array the conditional distribution of the log-ratio of intensity I_{ioj} given the corresponding membership indicator c_{io} is assumed to be given by

$$I_{ioj} = \mu_{oj} + E_{ioj}, \text{ where } E_{ioj} \sim N(0, 1/\tau(c_{io})) \quad (1)$$

where the arrays are ordered as in the original data. That is, we start with four arrays from mouse 1 with the first two being the two replicates with green dye applied to the experimental sample, and so on.

In the following we adopt the notation that a vector or a collection of certain parameters/variables is represented by suppressing the subscripts. The posterior density $P(\mu,\tau,C,\lambda|I)$ is proportional to the joint density $P(\mu,\tau,C,\lambda,I)$ and by assuming suitable conditional independence properties between the parameters to hold, it can be presented in the product form $P(I \mid \mu,\tau,C)\ P(C \mid \lambda)\ P(\tau)\ P(\mu)\ P(\lambda)$. We assume vague priors for all model parameters. The array effects (μ) were assigned Normal priors, N(0,10). The precision parameters (τ) were assumed to have Gamma distributions *a priori*, $\tau_1 \sim \Gamma(1,1)$, $\tau_{j+1} = \tau_j + \eta_j$ where $\eta_j \sim \Gamma(1,1)$ and $j = 1, 2$. The latent class-indicators (C) were assigned Multinomial distributions with corresponding probabilities (λ) drawn from a Dirichlet distribution, $D(\mathbf{1})$. In order to preserve compatibility, the estimation of the model parameters for all three organs was carried out simultaneously.

3.1 Model Implementation and Sensitivity Analysis

We implemented the model and performed parameter estimation using WinBUGS [Gilks *et al.*, 1994]. Missing data points were treated in the same way as parameters in our model and were completed during estimation using Bayesian data augmentation [Gelman *et al.*, 1995]. Ten thousand Markov chain Monte Carlo (MCMC) samples were drawn based on multiple (parallel) chains with additional burn-in rounds. The convergence of the chain was monitored by CODA [Best *et al.*, 1995] and by inspecting the sample paths of several of the model parameter estimates. Additional sensitivity analysis was carried out by comparing the posterior distributions arising using different choices of the hyper-parameters in the priors. Based on results of these analyses, we concluded that, due to the large sample size and well-structured data, the posterior distributions of population parameters were almost identical independently of the choice of the prior.

3.2 Model A: Results

From the posterior estimates of μ's (Figure 1), it was observed that, apart from array specific variations in the estimates, the estimates clearly depict an effect of dye on the observed log-ratio of intensities, especially in samples from kidney and testis. For these, all arrays in which the experimental sample had been treated with green dye had estimated means higher than the corresponding means resulting from treatment with red dye, which is a phenomenon commonly observed in many microarray experiments. Such an effect may be present in all genes but the magnitudes might differ from gene to gene. Arrays based on the samples from testis showed not only a dye effect in the array means but also indicated a possible mouse effect.

However, to what extent the observed mouse effect is caused by biological factors and to what extent it is affected by experimental artifacts like RNA preparation, is not identifiable from the present data.

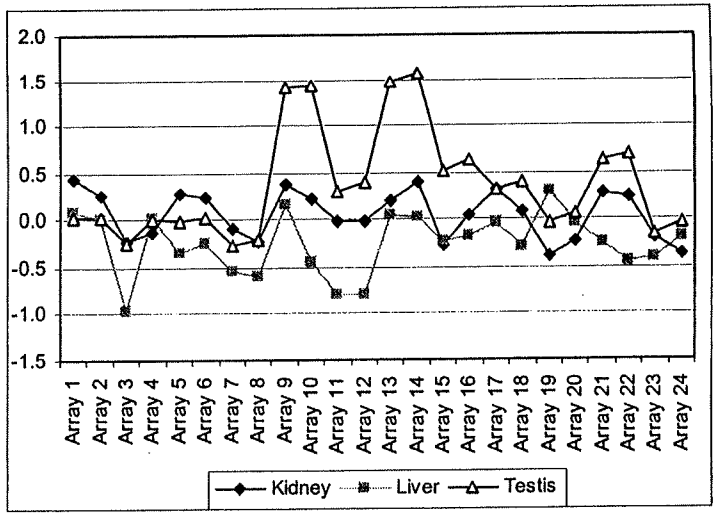

Figure 1. Plots of estimated posterior means for 24 arrays each from the three organs (kidney, liver, and testis) with Model A. Different line types are given in the bottom. Arrays are ordered according to 6 mice (M), 2 dyes (D) and 2 replicates (R) as (M1,D1,R1), (M1,D1,R2), (M1,D2,R1), (M1,D2,R2), (M2,D1,R1), ...

The posterior distributions of the three precision parameters (τ's) were quite separated from each other, depicting three distinct variance categories for the genes. Also, the estimated distributions were highly concentrated around the posterior mean, indicating un-ambiguity about the corresponding variance class.

Table 1. Posterior estimates of (common) precision parameters and gene proportions in the three precision groups (1,2,3) under Model A in the three organs.

Parameter	Group	Kidney	Liver	Testis
Precision	1	0.3	0.3	0.3
	2	2.9	2.9	2.9
	3	14.2	14.2	14.2
Percentages of genes	1	13.0	12.2	7.6
	2	40.1	40.1	41.7
	3	46.9	47.7	50.7

The estimated variance class for the residual variances for each gene (I) in each organ (O) was obtained from posterior distributions of the corresponding C_{IO}'s. It was characteristic of the results that, *a posteriori*, the genes were assigned to a variance class (within each organ) quite distinctly

and for many genes the estimated membership probabilities were either near zero or near one, although the model puts no such constraints and priors were chosen to be non-informative. Also possibilities of bad mixing / non-convergence were effectively ruled out by validating the posterior estimates from several parallel chains with distinctly different initial values.

From Table 1 we observe that the smallest number of highly varying genes was observed in testis (7.6%) and the largest in kidney (13.0%).

From the estimated variance-classes, it was observed that some of the genes behave differently across organs (Table 2). Although a large proportion of the genes was identified to be less varying in all three organs, there were genes which dramatically changed the assignment of variance class from one organ to another. About 79% of genes were estimated to belong to moderate or low variance classes in all three organs. Only 1.7% of genes were estimated to have high variation in all three tissues.

Table 2. Cross tabulation of genes according to their estimated variation groups (1: High, 2: Moderate, 3: Low) in different organs (K: kidney, L: liver, T: testis) under Model A (in %).

% of genes		(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	Total
(K,1)	(L,1)	1.7	6.0	0.4							8.1
	(L,2)	0.7	2.7	0.9							4.3
	(L,3)	0.2	0.2	0.2							0.6
(K,2)	(L,1)				0.7	1.5	0.5				2.8
	(L,2)				2.6	11.4	7.7				21.8
	(L,3)				0.6	8.0	7.0				15.6
(K,3)	(L,1)							0.6	0.4	0.3	1.3
	(L,2)							0.6	6.0	7.5	14.1
	(L,3)							0.0	5.3	26.2	31.5
Total		2.6	9.0	1.4	4.0	20.9	15.2	1.1	11.8	34.0	100.0

4. EXTENDED MODEL (MODEL B)

It is expected that some genes may be expressed differently in one organ compared to their average expression in all three organs. This indicates that for these genes, for a particular organ, the observed log-ratio-of-intensities can be away from the expected zero value. That is, they can have a higher or lower expression in the experimental sample in one particular organ when compared to the expression in the reference sample. Since the reference samples were taken from all three tissues, it can then be expected that for the same genes in one or both of the remaining organs the log-ratio-of-intensities would behave in the opposite way than in the first organ.

In order to account for such between-organ differences in gene expression we modeled the log-ratio of intensities for the genes in the

following manner. The extended model continued to have array mean effects across all genes (i.e. μ's) as in Model A. Additionally each gene (I) was classified independently in each organ (O) as belonging to one of three possible expression groups (D_{IO}). Accordingly for each gene, an expression level ($\theta(D_{IO})$) was added to the corresponding array mean effect (μ_{OJ}) to explain the expected log-ratio of intensities of that gene.

As in Model A, each gene was classified independently for each organ with respect to its residual variance (C_{IO}).

In summary, the conditional distribution of the log-ratio-of-intensity $I_{IOJ,}$, given C_{IO} and D_{IO}, was assumed to be given by (with I, O, J as before),

$$I_{IOJ} = \mu_{OJ} + \theta(D_{IO}) + E_{IOJ}\ , \text{ where } E_{IOJ} \sim N(0,\ 1/\tau(C_{IO})) \qquad (2)$$

The prior assumptions on the original parameters were kept unchanged. In addition, the θ's were assigned Normal priors (on appropriate ranges) with $\theta_1 < \theta_2 = 0 < \theta_3$ and, similar to C_{IO}'s, the latent variables D_{IO} were assumed to be *a priori* Multinomial with corresponding probabilities (λ_D) drawn from a Dirichlet distribution. The posterior density $P(\mu,\tau, C, \lambda_C, D, \lambda_D | I)$ is then defined in a manner similar to the previous model.

4.1 Model B: Results

According to the extended model a gene in organ o and array J can now be assigned to one of three different means parameterized by $\mu_{OJ} + \theta(D_{IO})$. The estimated values of $\mu_{OJ} + \theta(D_{IO})$ with $D_{IO} = 2$ were comparable to the average array effects μ_{OJ} obtained under the previous model. On the other hand, $\mu_{OJ} + \theta(D_{IO})$ with $D_{IO} = 1$ correspond to a lower expression class, and those with $D_{IO} = 3$ to a higher expression category. Plots of posterior estimates of $\mu_{OJ} + \theta(D_{IO})$ clearly depict the distinct nature of the three expression levels. Those for kidney samples have been presented in Figure 2. Similar observations were made for liver and testis samples.

The estimated variance-classes improved in the sense that each of the estimated precision parameters (τ) under the new model is higher than that under the preliminary model. In other words, allowing for gene specific adjustment of expression levels in the extended model explained more variation in the observed expressions thereby leading to reduced residual variances (Table 3).

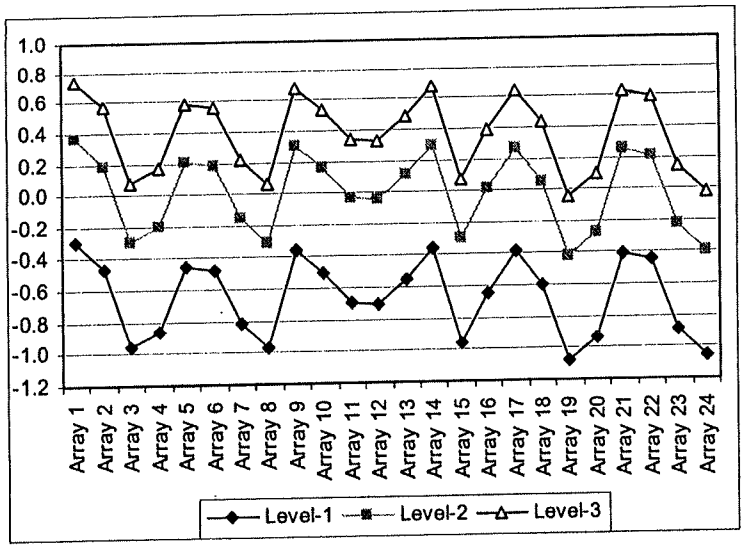

Figure 2. Kidney data: Plots of posterior estimates of $\mu_{OJ} + \theta(D_{IO})$ for O = K, J = 1, ..., 24 and three different expression levels corresponding to D_{IO} = 1 (lower expression), 2 (average expression) and 3 (higher expression) with array ordering as in Figure 1.

Moreover, even though the new variance classes gave rise to narrower distributions for the expressions, the estimated proportions of genes in the lowest variance-class still increased from Model A to Model B. Also the number of genes in the highest variance class was reduced compared to Model A (Table 3).

Table 3. Posterior estimates of (common) precision parameters and gene proportions in the three precision groups (1,2,3) under Model B in the three organs.

Parameter	Group	Kidney	Liver	Testis
Precision	1	0.5	0.5	0.5
	2	4.7	4.7	4.7
	3	19.0	19.0	19.0
Percentages of genes	1	11.1	8.7	5.0
	2	34.3	42.7	36.7
	3	54.7	48.6	58.3

Under Model B, more genes were estimated to have moderate or low variation in all three organs, compared to A (Table 4). For some genes, the estimated variance classes still varied across the three organs, although the number of such genes was smaller than in the previous model.

Table 4. Cross tabulation of genes according to their estimated variation groups (1: High, 2: Moderate, 3: Low) in different organs (K: kidney, L: liver, T: testis) under Model B (in %).

% of genes		(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	Total
(K,1)	(L,1)	1.4	2.7	1.2							5.2
	(L,2)	0.8	3.4	0.7							4.9
	(L,3)	0.1	0.6	0.2							0.9
(K,2)	(L,1)				0.5	1.1	0.4				2.0
	(L,2)				0.8	12.7	7.7				21.2
	(L,3)				0.3	5.2	5.7				11.1
(K,3)	(L,1)							0.4	0.4	0.6	1.4
	(L,2)							0.6	4.8	11.2	16.6
	(L,3)							0.1	5.9	30.6	36.6
Total		2.3	6.6	2.1	1.6	19.0	13.7	1.1	11.1	42.5	100.0

From the estimated θ values and the estimated posterior distributions of D_{IO}'s, we observed that several genes had a higher expression level in one organ and a lower expression in another, supporting our motivation for using the extended model (Model B) to incorporate organ specific differential expression. For example, in Table 5 the relatively large number of genes in the furthest off-diagonal positions indicate opposite (lower/higher) expression levels for those genes in different organs.

Table 5. Cross tabulation of genes according to their estimated expression groups (1-lower, 2-average, 3-higher) in the three organs (viz. K: kidney, L: liver and T: testis) under Model B.

% of genes		(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	Total
(K,1)	(L,1)	0	8	691							699
	(L,2)	6	87	208							301
	(L,3)	93	50	16							159
(K,2)	(L,1)				0	25	124				149
	(L,2)				7	914	812				1733
	(L,3)				103	635	87				825
(K,3)	(L,1)							97	48	20	165
	(L,2)							119	443	111	673
	(L,3)							346	327	24	697
Total		99	145	915	110	1574	1023	562	818	155	5401

In all three organs, 90% or more of the genes with a higher than average expression also had moderate or low variance (Table 6). As expected, genes with a lower expression showed greater variation and about 70% of the genes with lower expression had low or moderate variance.

The expression measures from testis samples showed the larger share of genes within the high expression group (38.8%), compared to kidney (28.4%) and liver (31.1%). This may be due to over-representation of testis genes in the reference. Testis is known to have a greater percentage of

messenger RNA than liver and kidney. When equal amounts of total RNA are used to make the reference, more of the reference ends up being from the testis since only messenger RNA is made into array probes.

Also although 5% of all genes in the testis is estimated to have high variation, only 0.3% of the 38.8% highly expressed genes were estimated to be highly varying.

Table 6. Organ-wise cross tabulation of genes (in %) according to their estimated variance and expression levels.

		Kidney data				Liver data				Testis data			
		Variance group				Variance group				Variance group			
		1	2	3	All	1	2	3	All	1	2	3	All
Exp level	1	6.7	9.1	5.7	21.5	5.1	9.8	3.8	18.8	3.5	6.8	3.9	14.3
	2	2.7	15.7	31.8	50.1	0.9	20.4	28.8	50.1	1.3	20.7	25.0	47.0
	3	1.7	9.5	17.2	28.4	2.6	12.5	16.0	31.1	0.1	9.2	29.4	38.8
	All	11.1	34.3	54.7	100	8.7	42.7	48.6	100	5.0	36.7	58.3	100

5. MOUSE MODEL (MODEL C)

Pritchard *et al.* [2001] observed that for some genes the expression varied significantly across mice. Note that both in Model A and B, by introducing an array-level effect, namely μ, a mouse-level effect has already been nested into the model. We can further extend the proposed models A and B to incorporate an additional mouse-effect. We have already established that the genes have varied expression levels across organs. Moreover, it is possible that even if a certain gene has an above average expression in a certain organ, its magnitude may vary across mice. Similarly, since the reference samples were taken from all three tissues, it is possible that the same genes might exhibit average or low expression in one or both of the remaining organs, but again its magnitude might vary across mice.

In order to account for such between-organ and between mice differences in gene expression we modeled the log-ratio of intensities of the genes in the following manner. The extended model continued to have array mean effects across all genes (i.e. μ's) as in Models A and B. Additionally each gene (I) was classified independently in each organ (O) as belonging to one of three possible expression groups (D_{IO}). Accordingly for each gene and for each mouse, an expression-level ($\theta(D_{IO}, M_J)$) with M_J indicating the mouse number for the J-th array, was added to the corresponding array mean effect (μ_{OJ}) to explain the expected log-ratio of intensities of that gene.

As in Model A, each gene was classified independently for each organ with respect to its residual variance (C_{IO}).

In summary, the conditional distribution of the log-ratio-of-intensity I_{IOJ}, given C_{IO} and D_{IO}, was assumed to be given by (with I, O, J as before),

$$I_{IOJ} = \mu_{OJ} + \theta(D_{IO}, M_J) + E_{IOJ}, \text{ where } E_{IOJ} \sim N(0, 1/\tau(C_{IO})) \quad (3)$$

The prior assumptions on the original parameters were kept unchanged, with the θ's for the L-th mouse being assigned Normal priors (on appropriate ranges as before) with $\theta_{1L} < \theta_{2L} = 0 < \theta_{3L}$. The latent variables C_{IO}, and D_{IO} were assumed to be *a priori* Multinomial with corresponding probabilities (λ_C and λ_D) drawn from Dirichlet distributions. The posterior density $P(\mu, \tau, C, \lambda_C, D, \lambda_D | I)$ was then defined in a manner similar to the previous model.

5.1 Model C: Results

The estimated variance classes were comparable with the estimates obtained under Model B. The model continued to improve over the previous models, in the sense of explaining the observed variation better, by reducing the proportion of highly varying genes and increasing that of less varying genes (Table 7).

From the estimated mouse-specific expression levels it was observed that there were indeed differences in the expression levels across mice, although the differences were not significantly large.

Table 7. Cross tabulation of genes according to their estimated variance groups (1: High, 2: Moderate, 3: Low) in different organs (K: kidney, L: liver, T: testis) under Model C (in %).

% of genes		(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	(T,1)	(T,2)	(T,3)	Total
(K,1)	(L,1)	1.2	2.7	1.3							5.2
	(L,2)	0.7	3.1	0.7							4.5
	(L,3)	0.1	0.6	0.2							0.8
(K,2)	(L,1)				0.5	0.9	0.4				1.9
	(L,2)				0.8	12.1	7.0				19.9
	(L,3)				0.3	5.3	5.3				11.0
(K,3)	(L,1)							0.4	0.4	0.6	1.4
	(L,2)							0.6	5.1	11.9	17.6
	(L,3)							0.1	6.4	31.2	37.7
Total		2.1	6.4	2.1	1.6	18.4	12.7	1.1	11.9	43.7	100.0

Similar computations as presented in Table 5 were carried out with the result that a large number of genes continued to have different expression levels across the three organs as before.

6. MODEL COMPARISON

6.1 Comparison between Models A, B and C

In most microarray data analysis normalization is done as a separate procedure without considering its effect on subsequent classification. In the three models proposed here, the data have been normalized/adjusted using increasingly available information on biological and experimental factors possibly affecting the observed data. As a result we observed gradual improvement and refinement in classification of the genes.

We have already noted that over the three models for each of the three organs the number of genes identified as having a high variance decreased, and also the share of genes with low variance increased, when moving from Model A to B and then to C.

When the behavior of any particular gene in all three organs is considered simultaneously, then collectively for all genes from Tables 2, 4 and 7 we note that the proportions of genes with high variance in all three organs was reduced across the models, with the percentages being 1.7, 1.4 and 1.2 respectively under Model A, B and C. On the other hand the percentages of genes with average or low variance in all three organs increased (A: 79.1%, B: 83.7% and C: 84.3%). This implies that employing biological and experimental hypotheses in model building better explains the observed variations in the data.

In the following we give a brief example (Figure 3) of how the models work. Of the 5401 genes, a few were selected and log-ratio-of-intensities from the kidney sample were plotted. The plot of the original data shows wide variation across arrays.

However, *a posteriori* adjusted log-ratio-of-intensities under Model A show smoothing over arrays. Recall that the model did not by itself introduce any such assumptions. Hence such smoothing implies that possibly some common factors, experimental or biological, had affected the observed log-ratio of intensities of most of the genes along with those of the selected ones. These factors, when taken care of by the array effect components in the model, smooth out the observed log-ratios.

For the selected set of genes, although the Model-A-adjusted ratios appeared much smoother than the original data, these plots were still distinctly away from zero. Clearly some of these genes have above average expression in the kidney sample, whereas others had below average. This is supported by the plots of *a posteriori* adjusted log-ratio-of-intensities under Model B. This plot shows array-wise movement towards the origin resulting in shrinkage of the plotted region.

By applying Model C some further smoothing was observed for mice 5 and 6, although very little in magnitude.

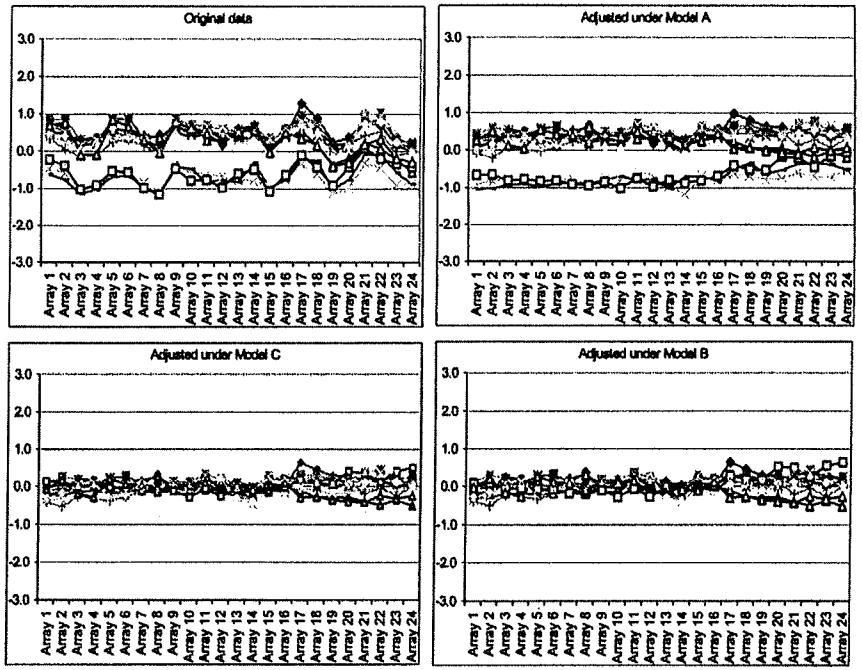

Figure 3. Plots of original and adjusted log-ratio of intensities for some genes as adjusted under Models A, B and C (Clockwise).

The proposed models gave consistent results in the sense of providing an improved explanation of observed variation with increasing complexity of the model. In principle, the models can be extended further by increasing the number of layers of the hierarchy. For practical purposes, in the absence of any specific hypothesis about essential sources of variation, amongst the proposed models, model B would be preferred. Well-established biological effects, such as tissue type, have already been accounted for in all three models that were proposed. In addition, if the biological replicates (here mice) are also suspected to cause additional variation, as suggested earlier [e.g., Pritchard et al., 2001], model C would be selected.

6.2 Comparison with Pritchard *et al.*

The genes identified by us as belonging to a large variance class were compared with those found by Pritchard *et al.* [2001]. Several such genes were identified in both studies. But we also noted that several of the genes identified by Pritchard *et al.* as significantly varying were not similarly classified by our models.

Incidentally, the genes plotted in Figure 3 are some of the genes identified as varying (in kidney) by Pritchard *et al.* [2001] using an ANOVA model whereas our analyses indicated these as having average or low variance.

This appears to be so because the emphasis in ANOVA, when performed separately on every gene, is to study the *relative* contributions of different sources of variation. In particular, Pritchard *et al.* focused on identifying genes with a relatively higher contribution of variation due to mouse, irrespective of the absolute magnitude of variation. By contrast, our main aim was to identify genes with a high degree of variation that remained unexplained by the model(s).

The genes common in the two lists generally have a high variance in the sense of magnitude, in which there is a significant contribution from between mice variation. The remaining genes from the list of Pritchard *et al.* list are probably genes which exhibit variation across mice and moderate or low residual variance. Hence future observations from these genes might vary from mouse to mouse but overall differences are expected to stay within a "tolerable" limit. The genes additionally identified here were estimated to have a high variance making future observations from them less predictable. The observed variation in these genes could not be explained any further by introducing a model for between-mouse effects.

Four liver genes and three kidney genes were analyzed by Pritchard *et al.* using real-time quantitative RT-PCR to validate expression differences between mice. Three of the 4 genes confirmed to be variable in the liver by RT-PCR were in common between the two lists. These were CisH (NM_009895), Gadd45 (NM_007836), and PRH/Hhex (NM_008245). The fourth gene confirmed in the liver, Bcl-6 (NM_09744), was not found in the large variance class of the current analysis. Instead, under both Models B and C, this gene was identified as having above average expression in liver with moderate variance. However, applying the same models to the RT-PCR data from liver assigns Bcl-6 to the large variance class.

Three kidney genes confirmed by RT-PCR as varying are CisH, Bcl-6 and complement factor D (NM_013459). Of these, CisH was identified as varying by the present analysis too. Similar to liver, our models confirmed Bcl-6 as varying in kidney tissues based on RT-PCR data but not based on

microarray data. Although complement factor D was estimated to have moderate variance in kidney, its expression classification significantly changed from Model B to mouse Model C, supporting the existence of a mouse effect on this gene.

Results of Bayesian analysis of microarray data were compared for these selected genes which were identified as highly varying by Pritchard *et al*, [2001] based on microarray as well as RT-PCR data. In case of a disagreement between the two analyses, RT-PCR data was additionally used. This comparison is summarized in Table 8.

Table 8. Bayesian characterizations of selected genes which were estimated by Prichard *et al.* [2001] as highly varying in both microarray and RT-PCR data.

Organ	Gene annotation	Bayesian analysis
Kidney		
	CisH (NM_009895)	High variation
	Bcl-6 (NM_09744)	Microarray data: Avg. variation RT-PCR data: High variation
	Comp. Factor D (NM_013459)	Microarray data: Avg. variation with expression group changing from high to average from Model B to Model C indicating possible mouse effect
Liver		
	CisH (NM_009895)	High variation
	Gadd 45 (NM_007836)	High variation
	PRH/Hhex (NM_008245)	High variation
	Bcl-6 (NM_09744)	Microarray data: Avg. variation with high expression RT-PCR data: High variation

Another gene, β-2 microglobulin, which was identified by Pritchard *et al.* as varying in both kidney and in testis, was also estimated by us to have high variance in both samples under Model A. In testis under all three models this gene continued to be identified as varying. However, its estimated variance class in kidney dramatically changed when moving from Model A to Model B. Under Model B in kidney this gene was estimated to have low variance class with above average expression.

7. CONCLUDING REMARKS

Our approach to statistical modeling and analysis has been integrated in the sense that normalization and classification are being carried out simultaneously.

The normalization factors as obtained by Pritchard *et al.* were partly compared with those obtained by us under different models and were found to be comparable. Recall that our method of normalization is statistical and not deterministic as is commonly done. Additional adjustments suggested in Models B and C to incorporate above or below average expression, respectively without and with mouse specific variation, have certainly helped us to explain the observed variation. This is evident from the decrease in the number of genes estimated to have a higher variance. We emphasize the point here that all these were achieved using solely experimental and biological information and without resorting to any ad hoc or artificial adjustments.

Model A takes into account normalization for experimental factors distorting measurements of all genes on an array. Model B extends Model A by incorporating some available biological information. As an example of possible further extension Model C was proposed. With the availability of further knowledge on relevant biological factors, extended models can be formed using such information.

Classification as is proposed here helps one to reduce the dimensionality problem significantly. For example, a standard clustering algorithm would have treated the observed data from each organ as a vector of 24 dimensions. Instead we reduce the problem to two dimensions only and then perform classification with respect to the adjusted mean and residual variance of what is assumed to be 24 exchangeable observations.

It may be mentioned here that it is also possible to treat the number of groups as an unknown parameter in the model and estimated simultaneously along with the others by using techniques like reversible jump MCMC [Green, 1995; Medvedovic and Sivaganesan, 2002]. One could also consider a hierarchical model in which the variances are distributed according to some continuous distribution. In a special case this leads to Student's t-distribution for errors [Geweke, 1993]. However, even if we would consider the variances as continuous variables, in order to address the basic problem of differentiating between genes, one would finally have to resort to some sort of discretization, e.g., by introducing cut-off points.

The proposed model is based on the idea of applying a simple discrete approximation that leads to an easy implementation and intuitive interpretation of the results. In order to be able to differentiate between genes according to their variance level, suitable numbers of classes can be

explored, starting from two, and depending on the required fineness/resolution of the partitions in the variance scale. Several such numbers of classes were explored here and three classes were chosen only for the purpose of illustration and easy interpretability.

Estimation of uncertainty was obtained along with classification and consequently it was unnecessary to carry out a large number of testing of hypotheses which also avoids complications of multiple comparisons.

The proposed method of analysis brought out a previously unexplored aspect of this data set. Namely the relative level of activeness of the genes as estimated by the expression groups together with their predictability, which is given by their respective variance (or precision) class estimator.

Additionally from these models, gene-level information on their respective expression levels across the organs could certainly provide useful insight into understanding the relevant regulatory networks in different organs.

REFERENCES

Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17: 509-519.

Best, N. G., Cowles, M. K. and Vines. S. K. (1995) CODA: Convergence Diagnosis and Output Analysis software for Gibbs Sampler output: Version 0.3. Cambridge: Medical Research Council Biostatistic Unit.

Dror RO, Murnick JG, Rinaldi NA (2002) A Bayesian approach to transcript estimation from gene array data: the BEAM technique. RECOMB 2002: Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (ACM PRESS).

Gelman A, Carlin JB, Stern HS, Rubin DB. (1995) Models for missing data. In: Bayesian data analysis. London: Chapman & Hall; pp. 439-66.

Geweke J. (1993) Bayesian treatment of the independent Student-t linear model. Journal of Applied Econometrics 8: S19-S40.

Gilks WR, Thomas A, Spiegelhalter DJ (1994) A language and program for complex Bayesian modeling. The Statistician 43: 169-178.

Green PJ. (1995) Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika. 82: 711-732.

Ibrahim JG, Chen M-H, Gray RJ (2002) Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc 97: 88-99.

Keller AD, Schummer M, Hood L, Ruzzo WL (2000) Bayesian classification of DNA array expression data, *Technical Report*, UW-CSE-2000-08-01, Dept. of Comp. Sc. & Engg., Univ. of Washington, Seattle.

Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. J Biol Chem 276: 19937-19944.

Medvedovic M (2000) Identifying significant patterns of expression via Bayesian infinite mixture models. *CAMDA'00 : Critical Assessment of Techniques for Microarray Data Analysis*, Duke University.

Medvedovic M and Sivaganesan S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 18: 1194-1206.

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comp Biol 8: 37-52.

Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E (2002) A statistical framework for expression-based molecular classification in cancer. J Roy Stat Soc B, 64: 717-736.

Pritchard CC, Hsu L, Delrow J, Nelson PS (2001) Project normal: Defining normal variance in mouse gene expression. Proc Natl Acad Sci USA 98: 13266-13271.

SECTION VI

INVESTIGATING CROSS HYBRIDIZATION ON OLIGONUCLEOTIDE MICROARRAYS