

Analysis of gene expression data of the NCI 60 cancer cell lines using Bayesian hierarchical effects model

Jae K. Lee^{a*}, Uwe Scherf^b, Lawrence H. Smith^b, Lorraine Tanabe^b, and John N. Weinstein^b

^aUniversity of Virginia, P.O.Box 800717, Charlottesville, VA 22908-0717

^bLab of Molecular Pharmacology, National Cancer Institute, Bethesda, MD 20892

Abstract

From the end of the last decade, NCI has been performing large screening of anticancer drug compounds and molecular targets on a pool of 60 cell lines of various types of cancer. In particular, a complete set of cDNA expression array data on the 60 cell lines are now available (Scherf et al., 2000; Ross et al., 2000). To discover differentially-expressed genes in each type of cancer cell lines, we need to estimate a large number of genetic parameters, especially interaction effects for all combinations of cancer types and genes, by decomposing the total variance into biological and array instrumental components. This error decomposition is important to identify subtle genes with low biological variability. An innovative statistical method is required for simultaneously estimating more than 100,000 parameters of interaction effects and error components. We propose a Bayesian statistical approach based on the construction of a hierarchical model adopting parametrization of a linear effects model. The estimation of the model parameters is performed by Markov Chain Monte Carlo, a recent computer-intensive statistical resampling technique. We have identified novel genes whose effects have not been revealed by the previous clustering approaches to the gene expression data.

Key words: Cancer; Gene expression array; Hierarchical effects model; Gibbs sampling; NCI 60 cell line screening

1 Introduction

Since 1990, the Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI) has been screening potential anticancer agents for their activity in culture against a battery of 60 human cancer cell lines representing 9 different organs of origin. Included are leukemias, melanomas, and cancer cells of colorectal, renal, ovarian, breast, prostate, central nervous system, and lung origin. Using a two-day growth inhibition assay, drug potency values have been measured for >70,000 of the NCI's inventory of >500,000 chemically defined compound and also for >100,000 natural product extracts. The resulting patterns of activity across the 60 cell types have been found to provide incisive information on mechanisms of drug action, resistance, and modulation (Paull *et al.*, 1989; Weinstein *et al.*, 1992; Weinstein *et al.*, 1997). In that sense, the drug screen is also a drug profiling system. In order to understand better the meaning of the drug activity profiles, we and others have been studying the 60 cell lines with respect to a variety of molecular characteristics at the DNA, mRNA, protein, and functional levels. For example, we and our collaborators have (i) generated a database of expression patterns for 1,014 proteins in the cells using 2-D polyacrylamide gel electrophoresis (Myers *et al.*, 1997); and (ii) assessed mRNA expression using 10,000-gene Synteni/Incyte cDNA microarrays (Scherf et al., 2000; Ross et al., 2000).

*To whom correspondence should be addressed. Tel: (804) 924-8430, Fax: (804) 924-8437, Email: jaeklee@virginia.edu, Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences

Given these rich databases on drug activity and molecular characteristics of the 60 cell lines, a number of different mathematical methods to analyze them have been introduced—in part to illuminate the basic biology and in part to aid in the drug discovery process. The first was the COMPARE program, which uses the Pearson correlation coefficient as a measure of the similarity of drug to drug or drug to target in terms of pattern across the 60 cell lines (Paull *et al.*, 1989). A variety of other statistical and artificial intelligence methods have been introduced to address various questions about these databases. Included have been back-propagation neural networks (Weinstein *et al.*, 1992), principal component analysis (Koutsoukos *et al.*, 1994), hierarchical cluster analysis (Weinstein *et al.*, 1997; Shi *et al.*, 2000; Myers *et al.*, 1997), and color-coded clustered image maps (CIMs) for data visualization (Weinstein *et al.*, 1994, 1997; Myers *et al.*, 1997).

These methods have successfully addressed a number of important issues with respect to the data in the spirit of exploratory data analysis. They have not, however, provided us with statistically well-defined, rigorous models for statistical inference and parameter estimation with respect, for example, to the multifactorial relationships between cell type and drug activity or between cell type and gene expression. This is the type of problem often addressed statistically by generalized linear models (GLM), but straightforward GLM approaches to these data would be unsatisfactory for two fundamental reasons: (i) GLM approaches cannot flexibly decompose different layers or components of statistical error, such as those that can be identified with respect to the current large genomic data sets. Specialized GLM techniques such as split-plot designs, intended for such complex application, do not generally provide reliable estimates of the effects and variances of the multiple factors (Rao, 1973; Pukelsheim, 1981); (ii) GLM cannot reasonably handle the astronomical number of parameters to be estimated. For example, a two-way GLM model for 10,000-gene data across 9 different cell line types would require the estimation of more than 100,000 parameters for the main effects and binary interactions among parameters.

To deal with the first problem, hierarchical error structure with different components of error is introduced sequentially, one error component built upon the last. Toward that end, we introduce here a hierarchical effects model for the variability of chronological experimental stages and/or different biological factors within the 60 cell line data. Hierarchical effects modeling is a rapidly emerging area in the statistical literature (see, for example, Searle *et al.*, 1992; Daniels and Gatsonis, 1999), but it has not, to our knowledge, been applied to problems such as these in genomics or pharmacology. The second problem, the large number of variables, must also be solved, since exhaustive analytic calculations for the parameters or their approximations would be intractable in GLM. Hence, we adopt a computationally intensive but practical approach to calculation based on Gibbs sampling (GS), a recent computational resampling technique for complex models with large numbers of parameters and missing or censored data. GS is often used, for example, in the context of hidden Markov chain Monte Carlo techniques (Smith and Roberts, 1993; Besag *et al.*, 1995). In this study we apply this approach to gene expression profiling data on cDNA microarrays. In the next sections we will develop the mathematical basis for application of our approach to those data.

2 The Model

Suppose there are G genes on each microarray chip against a particular cell type and C categories or subgroups of cancer cell lines for analysis, e.g. nine cancer types in 60 cell lines—leukemia, melanoma, carcinoma of colon, ovarian, lung, renal, prostate, CNS, or breast origin. We then wish to estimate the parameters for the interaction effects — the effects of each of $G \times C$ combinations “adjusted” by general gene and group effects since we are interested in differential gene expression patterns across different cell types. In this estimation procedure we especially want to decompose the total variability into several biological and experimental components, so that we can evaluate the significance of such biological interaction effects purely based on biological variability, removing that of experimental variability. These will provide more biologically relevant interpretation on the interaction effects than the classical GLM estimators since the magnitude of biological effects is directly compared with its corresponding variability. To do so, the relationship between observed data and parameters is constructed by two hierarchical stages of biological and experimental error components, using consecutive conditional normal distributions. We note that error components for several array instrumental factors such as labeling, chip, and sample prep, are lumped together because these error components are confounded in the current array experiment of the 60 cell lines. However, as some independent replicates of these factors are available, variability of these error components can be separately estimated by a minor modification of the current model.

We can mathematically formulate the hierarchical model as follows. The first step is for the experimental error variability. It is well known that the variability of ratio statistics in array data is proportional to the mean expression value of each gene (Chen et al., 1997). Therefore, in this study we assume that an appropriate transformation, such as log-transformation of ratio statistics has been performed to stabilize the heterogeneous variances over different ranges of gene expression intensity. In our preliminary analysis (data not shown) and some other studies on gene expression data (e.g., Scherf et al., 2000), a log-transformation has generally been shown to result in a reasonably homogeneous variance of gene expression data. Therefore, at the first stage of our modeling we assume a constant variance of the instrumental error component of gene expression values, σ_ϵ^2 , given a fixed biological mean response of individual k in group C_j for gene G_i , for $r_{i,j,k}$, the l -th (repeatedly) observed gene expression value of the i -th gene for a particular k -th cell line in the j -th group is considered as

$$y_{i,j,k,l} \mid \{\text{fixed } r_{i,j,k}\} = r_{i,j,k} + \epsilon_{i,j,k,l}, \quad \text{with } \epsilon_{i,j,k,l} \sim \text{Normal}(0, \sigma_\epsilon^2), \quad (1)$$

where $i = 1, \dots, G; j = 1, \dots, C; k = 1, \dots, m_{i,j}; l = 1, \dots, n_{i,j,k}$. However, in the second layer of our hierarchical model, $r_{i,j,k}$ is also considered as a random component with a normal distribution:

$$r_{i,j,k} \mid \{\text{fixed } \mu, g_i, c_j, \delta_{i,j}, \sigma_{r_{ij}}^2\} = \mu + g_i + c_j + \delta_{i,j} + \nu_{i,j,k}, \quad \text{with } \nu_{i,j,k} \sim \text{Normal}(0, \sigma_{r_{ij}}^2), \quad (2)$$

where μ is the parameter for the grand mean, g_i and c_j are the parameters for the general gene and cell-type mean effects, $\delta_{i,j}$ is the parameter for the interaction effect, and $\sigma_{r_{ij}}^2$ is the parameter for the biological variability in the ij -th combination. Note that $m_{i,j}$ and $n_{i,j,k}$ may vary in each combination, especially due to missing data. Note also that this would be considered as over-parametrization as often done in GLMs, in which some constraints are required to estimate these parameters. In Bayesian paradigm over-parametrization is dealt somewhat differently by examining the posterior variance reduction for the informative parameters from the observed data. In this regard non-identifiability of the model parameters is not directly relevant in this case. From the two stages of our modeling, the joint probability of the observed and unobserved variables (complete likelihood) is:

$$\mathbb{P}r(\mathbf{y}, \mathbf{r}; \theta = (\mu, \mathbf{d}, \mathbf{c}, \delta, \sigma_{\mathbf{r}}^2, \sigma_\epsilon^2)) = \prod_{i,j,k,l} \phi\left(\frac{y_{i,j,k,l} - r_{i,j,k}}{\sigma_\epsilon}\right) \times \prod_{i,j,k} \phi\left(\frac{r_{i,j,k} - \mu - d_i - c_j - \delta_{i,j}}{\sigma_{r_{ij}}}\right), \quad (3)$$

where ϕ is the density function of the standard normal distribution.

The prior distributions are a uniform prior on μ and normal priors on g_i, c_j , and $\delta_{i,j}$ with mean zero and variance σ_g^2, σ_c^2 , and σ_δ^2 , respectively. For variance parameters $\sigma_{r_{ij}}^{-2}$ and σ_ϵ^{-2} , we use gamma priors with parameters $(\alpha_\gamma, \beta_\gamma)$ and $(\alpha_\epsilon, \beta_\epsilon)$, respectively, for mathematical convenience. Our inference on the model is then made by the *posterior distribution* $\pi(\mathbf{r}, \theta \mid \mathbf{y})$, the conditional distribution of the unobserved data \mathbf{r} and the parameters $\theta = (\mu, \mathbf{g}, \mathbf{c}, \delta, \sigma_{\mathbf{r}}^2, \sigma_\epsilon^2)$, given the observed data \mathbf{y} , which is proportional to

$$\mathbb{P}r(\mathbf{y}, \mathbf{r}; \theta) \times \prod_i \phi\left(\frac{g_i}{\sigma_g}\right) \times \prod_j \phi\left(\frac{c_j}{\sigma_c}\right) \times \prod_{i,j} \phi\left(\frac{\delta_{i,j}}{\sigma_\delta}\right) \times \prod_{i,j} \Gamma(\sigma_{r_{ij}}^2; \alpha_\gamma, \beta_\gamma) \times \Gamma(\sigma_\epsilon^2; \alpha_\epsilon, \beta_\epsilon), \quad (4)$$

where $\mathbb{P}r(\mathbf{y}, \mathbf{r}; \theta)$ is the joint probability in (3) and $\Gamma(*; \alpha, \beta)$ is the density function of a Gamma distribution with parameters α and β .

3 Implementation: Gibbs Sampling

Direct inference on the model is difficult due to various reasons, most notably, a large number of model parameters and missing data and complexity of the model. A recent computational, stochastic resampling approach is thus adopted (Gibbs sampling, Besag *et al.*, 1995). This technique enables us to sample both the parameters and the missing data directly from a complex (high-dimensional) model probability function known only up to constant. Our mathematical derivation for our Gibbs sampling algorithm is straightforward

because, even though the whole probability function of the model is complex, each conditional distribution of the parameters and unobserved data can be easily derived. Our sampling algorithm consists of consecutive updates of missing data \mathbf{r} and parameters θ from each of their full conditional distributions given the observed data. These (posterior) conditional distributions are:

$$\begin{aligned} \pi(\mu|rest) &= \text{Normal}\left(\sum_{i,j} \frac{\sum_k \frac{(r_{i,j,k} - g_i - c_j - \delta_{i,j})}{\sigma_{r_{i,j}}^2}}{\frac{m_{1,1}}{\sigma_{r_{1,1}}^2} + \dots + \frac{m_{G,C}}{\sigma_{r_{G,C}}^2}}, \left(\frac{m_{1,1}}{\sigma_{r_{1,1}}^2} + \dots + \frac{m_{G,C}}{\sigma_{r_{G,C}}^2}\right)^{-1}\right) \\ \pi(g_i|rest) &= \text{Normal}\left(\sum_j \frac{\sum_k \frac{(r_{i,j,k} - \mu - c_j - \delta_{i,j})/\sigma_{r_{i,j}}^2}{\frac{m_{1,1}}{\sigma_{r_{1,1}}^2} + \dots + \frac{m_{i,C}}{\sigma_{r_{i,C}}^2} + \frac{1}{\sigma_g^2}}}{\frac{m_{1,1}}{\sigma_{r_{1,1}}^2} + \dots + \frac{m_{i,C}}{\sigma_{r_{i,C}}^2} + \frac{1}{\sigma_g^2}}, \left(\frac{m_{1,1}}{\sigma_{r_{1,1}}^2} + \dots + \frac{m_{i,C}}{\sigma_{r_{i,C}}^2} + \frac{1}{\sigma_g^2}\right)^{-1}\right), \\ \pi(c_j|rest) &= \text{Normal}\left(\sum_i \frac{\sum_k \frac{(r_{i,j,k} - \mu - g_i - \delta_{i,j})/\sigma_{r_{i,j}}^2}{\frac{m_{1,j}}{\sigma_{r_{1,j}}^2} + \dots + \frac{m_{G,j}}{\sigma_{r_{G,j}}^2} + \frac{1}{\sigma_c^2}}}{\frac{m_{1,j}}{\sigma_{r_{1,j}}^2} + \dots + \frac{m_{G,j}}{\sigma_{r_{G,j}}^2} + \frac{1}{\sigma_c^2}}, \left(\frac{m_{1,j}}{\sigma_{r_{1,j}}^2} + \dots + \frac{m_{G,j}}{\sigma_{r_{G,j}}^2} + \frac{1}{\sigma_c^2}\right)^{-1}\right), \\ \pi(\delta_{i,j}|rest) &= \text{Normal}\left(\frac{\sigma_\delta^2}{\sigma_\delta^2 + \sigma_{r_{i,j}}^2/m_{i,j}} \sum_k \frac{r_{i,j,k} - \mu - g_i - c_j}{m_{i,j}}, \left(\frac{1}{\sigma_\delta^2} + \frac{1}{\sigma_{r_{i,j}}^2/m_{i,j}}\right)^{-1}\right), \\ \pi(r_{i,j,k}|rest) &= \begin{cases} \text{Normal}(\mu + g_i + c_j + \delta_{i,j}, \sigma_{r_{i,j}}^2), & \text{if all } y_{i,j,k,l} \text{'s missing,} \\ \text{Normal}\left(\frac{\sigma_{r_{i,j}}^2}{\sigma_{r_{i,j}}^2 + \frac{\sigma_\epsilon^2}{n_{i,j,k}}} \sum_l \frac{y_{i,j,k,l}}{n_{i,j,k}} + \frac{\frac{\sigma_\epsilon^2}{n_{i,j,k}}}{\sigma_{r_{i,j}}^2 + \frac{\sigma_\epsilon^2}{n_{i,j,k}}}(\mu + g_i + c_j + \delta_{i,j}), \right. \\ \left. \left(\frac{1}{\sigma_{r_{i,j}}^2} + \frac{n_{i,j,k}}{\sigma_\epsilon^2}\right)^{-1}\right), & \text{otherwise.} \end{cases} \\ \pi(\sigma_{r_{i,j}}^{-2}|rest) &= \text{Gamma}\left(\frac{m_{i,j}}{2} + \alpha_r, \sum_k \frac{(r_{i,j,k} - \mu - g_i - c_j - \delta_{i,j})^2}{2} + \beta_r\right), \\ \pi(\sigma_\epsilon^{-2}|rest) &= \text{Gamma}\left(\frac{N}{2} + \alpha_\epsilon, \sum_{i,j,k,l} \frac{(y_{i,j,k,l} - r_{i,j,k})^2}{2} + \beta_\epsilon\right). \end{aligned}$$

These steps are repeated, and the sample of size L , $\xi_1 = (\mathbf{r}_1, \theta_1), \dots, \xi_L = (\mathbf{r}_L, \theta_L)$, from the consecutive iterations of sampling are collected after discarding some portion of the initial burn-in time for mathematical convergence. In our current study, since all conditional distributions can be implemented by the Gibbs sampling, the convergence of our MCMC run has been achieved fairly quickly (within the first 100 iterations), and consecutive samples have relatively low correlations (data not shown). In the applications to 60 cell line data here, we use 300 and 2,000 iterations for our burn-in time and for the total length of the sampling, respectively. These choices, however, are flexible and may be different in other cases.

IMAGE clone ID: gene description	Nscore	mean	SD
488119: Homo sapiens clone 24589 mRNA sequence Chr.11	4.42	2.39	0.54
509820: ETV4 Ets variant gene 4 (E1A enhancer-binding protein)	4.28	2.51	0.58
299290: ESTs [5':W05338, 3':N75545]	4.07	2.55	0.62
49494: ESTs, Weakly similar to HYPOTHETICAL 81.5 KD PROTEIN	4.04	2.52	0.62
75340: [5':T57560, 3':T57514]:UA	4.00	1.94	0.48
486215: Urokinase-type plasminogen activator	3.87	1.91	0.49
344848: ESTs [5':W76206, 3':W72969]:UA	3.84	2.27	0.59
484681: Homo sapiens ES/130 mRNA, complete cds	3.82	2.43	0.63
360768: Human 19.8 kDa protein mRNA, complete cds Chr.8	3.82	2.18	0.57
487502: ESTs Chr.1	3.81	2.23	0.58
510115: NRAMP2 Natural resistance-associated macrophage protein	3.77	2.68	0.71
287239: ESTs [5':, 3':N66980]:A	3.76	2.15	0.57
510482: ESTs [5':AA055725, 3':AA055668]:UA	3.75	2.74	0.73
489060: H.sapiens mRNA for TRAMP protein Chr.8	3.73	2.12	0.56
358754: Human mRNA for cysteine protease, complete cds	3.65	2.36	0.64
343063: Homo sapiens T245 protein (T245) mRNA, complete cds Chr.X	3.65	2.29	0.62
297055: ESTs Chr.X [297055, (EW), 5':W03870, 3':N73758]:EWA	3.64	1.78	0.49
293891: ESTs [5':, 3':N66023]:A	3.59	2.45	0.68
377004: ESTs [5':AA047777, 3':AA057780]:A	3.59	1.71	0.47
510189: Homo sapiens CAG-isl 7 mRNA, complete cds	3.58	2.53	0.70

Table 1: Genes with highest normal scores for colon cell lines, together with their group means and standard deviations

4 Results

Here, we provide preliminary results for 10K microarray data of the 60 cell lines, whose subset of 1,376 genes has been analyzed by a hierarchical clustering approach in Scherf et al. (2000). For this analysis, the 39 cancer cell lines whose cell line clusters have been found reliable in Scherf et al. (2000) were divided into 6 groups based on organ of origin. We, however, note that this is one of the many possible ways in which the cells could be categorized. Classification by p53 genotype, for example, would address questions related to apoptosis, G1 arrest, and DNA repair. Sub-classification of the leukemias as B-cell or T-cell in origin would address genotypic and phenotypic differences between the two. Although we will not consider such calculations here, our algorithm could just as easily be used to score the cell lines on the basis of their relationships to pre-defined subclasses of genes or targets.

From our MCMC results, we first calculated estimates of the interaction effects ($\delta_{i,j}$) and variance component ($\sigma_{r_{ij}}^2$), as in (2), for each combination of a gene and a cell line group. Then, a *relative normal score* ($\delta_{ij}/\sigma_{r_{ij}}$) was calculated based on the interaction estimates and sample standard error. This normalized score (Nscore) can be interpreted as the ratio of each interaction effect and its corresponding sample standard error, which will directly represent the statistical significance of the effects of each combination of the biological factors. Using this normalized score, the magnitudes of the interaction effects across all combinations of genes and cell types were compared, and the statistical significance of each interaction effect was assessed.

However, owing to the design of the 60 cell line data, the cancer relevant and general cell origin gene factors are confounded for the genes with high interaction effects in this application. In fact, there may be more genes relevant to specific origin of cell lines rather than cancer relevant ones, so that the genes identified

IMAGE clone ID: gene description	Nscore	mean	SD
323662: Homo sapiens mRNA for KIAA0607 protein, partial cds	3.02	1.64	0.54
307717: Homo sapiens KIAA0430 mRNA, complete cds	2.93	1.50	0.51
377611: Homo sapiens cysteine and glycine-rich protein 2	2.84	1.43	0.50
230124: Radixin Chr. [230124, (IE), 5':H80175, 3':H78800]	2.78	1.46	0.52
265494: ESTs, Weakly similar to coagulation factor v precursor	2.70	1.65	0.61
347761: ESTs Chr.7 [347761, (DIW), 5':W81508, 3':W81605]	2.69	1.33	0.49
484776: Human dystroglycan (DAG1) mRNA, complete cds	2.67	1.54	0.57
262691: GRL Glucocorticoid receptor Chr.5	2.60	1.15	0.44
469414: ESTs Chr.1 [469414, (I), 5':, 3':AA026919]	2.55	1.65	0.64
510489: Transferrin receptor (p90, CD71)	2.49	1.22	0.49
345527: Human transcription factor, forkhead related activator	2.46	1.51	0.61
488362: ESTs [5':AA046764, 3':AA046492]:A	2.38	1.57	0.65
291620: Restin (Reed-Steinberg cell-expressed filament protein)	2.38	1.61	0.67
298128: Homo sapiens KIAA0400 mRNA, complete cds Chr.2	2.35	1.35	0.57
512258: Human tazarotene-induced gene 2 (TIG2) mRNA	2.35	1.62	0.69
362926: PRKACB Protein kinase, cAMP-dependent, catalytic	2.34	1.31	0.55
309395: ESTs, Weakly similar to W09D10.2 [C.elegans]	2.30	1.24	0.54
30902: *EST AA359563 SID 30902, ESTs [5':R17787, 3':R41930]	2.26	1.27	0.56
343291: ESTs, Highly similar to DIAMINE ACETYLTRANSFERASE	2.24	1.46	0.65
470746: Human microfibril-associated glycoprotein-2 MAGP-2	2.24	1.08	0.48

Table 2: Genes with highest normal scores for melanoma cell lines, together with their group means and standard deviations

with high interaction effects may not directly lead us to discovery of genes responsible for cancer; if the data consisted of a direct comparison between normal and tumor cell lines, the genes with high interaction effects would be considered as gene markers directly associated with tumor progression. We, however, value our current development in two perspectives: (1) the general method will be useful when a chip experiment is performed for revealing differentially-expressed genes of specific biological factors and (2) the genes identified from the 60 cell line chip data will provide good reference lists for various gene expression studies of origin specific cancer. In this regard, we tabulate the results from two types of cancer—Colon and Melanoma—with the highest scores of gene/cell interactions (Tables 1 & 2). For example, in table 1 IMAGE clones 509820, 299290, and 486215 are found to be highly expressed in colon tissue. In table 2 IMAGE clones 323662 and 377611 are the genes expressed in skin and 307717 and 230124 are found in muscle and eye cells. A complete list of these interaction estimates can be obtained directly from the authors in this paper.

5 Discussion

Our hierarchical modeling provides a statistically appropriate way to analyze large scale, high-dimensional databases from large genomic data. We have shown here how our algorithm can be applied to such data sets, using as an example the databases on gene expression data in the NCI 60 cell line discovery program. In particular, we have used this to address the following questions for each of the 6 organs of origin represented among the 60 cell lines: “Which genes out of a database of 10K are most highly expressed for cancer cell lines derived from that particular organ?” Although we have not fully investigated other biological characterizations here, our algorithm could just as easily be used to ask which cell lines were most sensitive to pre-defined subclasses of gene expression.

Hierarchical modeling has several advantages for addressing these questions: (i) Although computationally intensive, the Gibbs sampling algorithm is feasible for large, highly multivariate genomic data sets; (ii) the error is decomposed into components with different biological and/or instrumental sources; (iii) the error

structure is modeled as hierarchical, which would be suitable for understanding sequential sources of error in biological experiments; (iv) our Gibbs sampling algorithm handles missing and censored data in a natural and appropriate way; (v) it has a statistically sound basis that permits inference on parameter estimates and tests of hypothesis.

Hierarchical modeling is becoming increasingly popular but, to the best of our knowledge, has not been used in the studies of gene expression data. The Gibbs Sampling algorithm is often utilized in the hidden Markov chain approaches being used successfully for nucleic acid and protein sequence analysis. One of the restrictions of our current algorithm is that it assumes the variates to be normally distributed. Since this may not be generally the case for real data such as those from genomic and pharmacological studies, the confidence limits on the parameter estimates and critical values for hypothesis testing in our current study are considered somewhat provisional. Therefore, it is necessary for us to investigate other distribution (both parametric and non-parametric) structures. Such extension is also relatively straightforward in our current strategy—hierarchical modeling and inference by stochastic resampling techniques. A further study with non-normal assumptions is also in progress. We have developed an interactive web-based tool of our current Gibbs sampling algorithm at the web site of the NCI Laboratory of Molecular Pharmacology (<http://discover.nci.nih.gov/>).

It may be difficult here for us to examine many possible associations that are highlighted by our analysis because it requires considerably enough outside information about most of these relationships to identify whether they are interesting cases. We thus illustrate only a case of origin of cancer as a suggestive direction for our further bioinformatic investigation. In the current study, just two layers of hierarchical structure were considered for components of the variance. This two-stage model is our first attempt to capture multiple layers of error variability in gene expression data, lumping the whole biological variability as one factor and all experimental variability as another. We, however, note that more refined layers can be modeled for each of the two error factors. For example, the biological variability can be decomposed into individual and cell-type specific errors and experimental variability into labeling, hybridization, and scanning variability, provided that our data support for such replicated observations. Thus, the algorithm can be generalized to any number of layers, insofar as the data and/or a priori information about the true error structure warrant. This plasticity may be especially useful when time series data sets are to be analyzed.

References

- [1] Besag, J., Green, P. J., Higdon, D., and Mengersen, K. (1995) Bayesian computation and stochastic systems. (with comments) *Statistical Science*, **10**, 3-66.
- [2] Boyd, M.R. (1997) Status of the NCI preclinical antitumor drug discover screen. *Cancer: Principles and Practice of Oncology Update*, **3**, 23-42.
- [3] Chen Y, Dougherty ER, and Bittner ML (1997). Ratio-bases dcisions and the quantitative analysis of cDNA microarray images, *Biomedical Optics*, **2**, 364-374.
- [4] Daniels, M.J. and Gatsonis, C. (1999) Hierarchical generalized linear models in the analysis of variations in health care Utilization, *J. of Amer. Stat. Assoc.*, **94**, 29-42.
- [5] Harville, D.A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. , *J. Amer. Stat. Assoc.*, **72**, 320-340.
- [6] Koutsoukos, A.D. et al. (1994) Discrimination techniques applied to the NCI in vitro anti-tumor drug screen: predicting biochemical mechnism of action. *Stat. Med.*, **13**, 719-730.
- [7] Myers, T.G., Waltham, M., Unsworth, E., Treston, A., Mushine, J., Anderson N.L., Kohn, K.W., Weinstein, J.N. (1997). A protein expression database for the molecular pharmacology of cancer. *Electrophoresis*, **18**, 391-402.
- [8] Paull et al. (1989). Display and analysis of patterns of differential activity of drugs against human tumor cell lines: Development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.*, **81**, 1088-1092.

- [9] Pukelsheim, F. (1981). On the existence of unbiased nonnegative estimates of variance components. *Ann. Stat.*, 9, 293-299.
- [10] Rao, C.R. (1973). Linear statistical inference and its applications, 2nd edn., *John Wiley & Sons*, New York.
- [11] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, Mar;24(3):227-35
- [12] Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, and Weinstein JN (2000). A cDNA microarray gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 24 (3), 236-244.
- [13] Searle, S.R., Casella, G., and McCulloch, C.E. (1992) Variance components, *John Wiley & Sons*, New York.
- [14] Shi LM, Fan Y, Lee JK, Myers T, Waltham M, Andrews A, Scherf U, Paull KD, Weinstein JN. (2000). Mining and Visualizing Large Anticancer Drug Databases. *J. of Chem. Inf. & Com. Sci.*, 40 (2), 367-379.
- [15] Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods. *J. Royal Stat. Soc., B*, 55, 3-23.
- [16] Weinstein, J.N. et al. (1992) Neural computing in cancer drug development: predicting mechanism of action, *Science*, 258, 447-451.
- [17] Weinstein, J.N. et al. (1997) An information-intensive approach to the molecular pharmacology of cancer, *Science*, 275, 343-349.