Bayesian Analysis of Mutational Spectra

David B. Dunson* and Kenneth R. Tindall†

**Biostatistics Branch and* †*Laboratory of Environmental Carcinogenesis and Mutagenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709*

Manuscript received April 24, 2000
Accepted for publication July 5, 2000

ABSTRACT

Studies that examine both the frequency of gene mutation and the pattern or spectrum of mutational changes can be used to identify chemical mutagens and to explore the molecular mechanisms of mutagenesis. In this article, we propose a Bayesian hierarchical modeling approach for the analysis of mutational spectra. We assume that the total number of independent mutations and the numbers of mutations falling into different response categories, defined by location within a gene and/or type of alteration, follow binomial and multinomial sampling distributions, respectively. We use prior distributions to summarize past information about the overall mutation frequency and the probabilities corresponding to the different mutational categories. These priors can be chosen on the basis of data from previous studies using an approach that accounts for heterogeneity among studies. Inferences about the overall mutation frequency, the proportions of mutations in each response category, and the category-specific mutation frequencies can be based on posterior distributions, which incorporate past and current data on the mutant frequency and on DNA sequence alterations. Methods are described for comparing groups and for assessing dose-related trends. We illustrate our approach using data from the literature.

STUDIES of the frequencies at which DNA alterations of different types occur within a gene have improved our understanding of both spontaneous and induced mutagenesis. Current approaches for the analysis of mutational spectra test for differences between groups in the mutant frequency (Carr and Gorelick 1994, 1995; Fung *et al.* 1994, 1998) or in the proportions of mutations falling into different response categories (Adams and Skopek 1987; Roff and Bentzen 1989; Piegorsch and Bailer 1994). New analytic methods are needed for (1) better characterizing changes in mutational spectra; (2) assessing differences in the frequencies at which mutations of various types occur within a gene; (3) identifying dose-related trends in spectra; and (4) accounting for heterogeneity among studies when incorporating data from previous studies. As we describe in this article, each of these analytic goals can be addressed by using a Bayesian hierarchical modeling approach.

In a standard mutational spectra study, a subset of the mutants are genotyped, usually by DNA sequence analysis, and these mutants are then assigned to specific response categories. Categories can be defined by the type of DNA alteration and/or the position of the mutated base pair. Since a single "jackpot" mutation that occurs early in the replication of a population can result in a large pool of mutants carrying the same mutation (Nishino *et al.* 1996), mutation frequencies are most accurately determined by removing identical mutations that were recovered from the same tissue of the same animal. While there is a small probability that these mutations were of independent origin, it is much more likely that these identical mutations represent the clonal expansion of a single mutant. This conservative approach to scoring mutations guarantees that all mutations reported were of independent origin. A limitation to this approach is that the site-specific mutational frequency may be slightly underestimated at mutational hotspots. However, hotspots can still be identified using this approach, and future studies can then be designed to assess the mechanistic origin of a mutational hotspot.

We assume that the total number of independent mutants and the numbers of mutations falling into the different response categories follow binomial and multinomial sampling distributions, respectively. Nishino *et al.* (1996) demonstrate that it is often reasonable to assume a Poisson distribution for the number of independent mutants. Since the mutation frequency is extremely small, the Poisson is an excellent approximation of the binomial distribution. We use the binomial, since it results in simplified implementation and interpretation of our Bayesian model. The assumption of a multinomial sampling distribution for the counts in the different mutational categories is standard in mutational spectra analysis (Piegorsch and Bailer 1994) and is a requirement of the widely used Monte Carlo hypergeometric test (Agresti *et al.* 1979; Adams and Skopek 1987; Roff and Bentzen 1989).

To complete a Bayesian specification of our model, we choose Beta and Dirichlet prior distributions for

Corresponding author: David B. Dunson, Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709.
E-mail: dunson1@niehs.nih.gov

the overall mutation frequency and the proportions of mutations in each response category, respectively. As is well known in the Bayesian literature, the Beta and Dirichlet distributions have advantageous computational properties (*e.g.*, conjugacy) and the parameters have appealing interpretations as prior sample sizes (Gelman *et al.* 1996). The prior parameters can be elicited on the basis of data from previous studies, using an approach we propose that adjusts for heterogeneity among studies, or a noninformative prior can be chosen. Our prior elicitation procedure advances the statistical literature on methods for incorporating historical data into the analysis of a current study (*e.g.*, Tarone 1982; Prentice *et al.* 1992; Ibrahim *et al.* 1998).

Inferences about the overall mutation frequency, the proportions of mutations in each response category, and the category-specific mutation frequencies can be based on Bayesian posterior distributions, which synthesize information in the prior and in the likelihood. Our Bayesian approach has several important advantages over current standard methods (*e.g.*, Adams and Skopek 1987; Carr and Gorelick 1994). First, in addition to testing for significant differences between groups, we can easily obtain point and interval estimates for any function of the mutation parameters. For example, in studies with multiple dose groups, we can estimate slope parameters that characterize the category-specific changes in the mutation frequency with increasing dose. Such estimates can be extremely useful in interpreting study results. Second, we can incorporate DNA sequence information into tests for overall differences (or trends) in the mutation frequency. Such information can potentially improve power to detect an effect relative to tests based on the mutant fraction. With the exception of the approach of Carr and Gorelick (1996), procedures that incorporate information on DNA sequence alterations have based inference on the proportions of mutations within different response categories (*i.e.*, the category probabilities). In most cases, differences in the category-specific mutation frequencies are more interpretable and biologically relevant than differences in the category probabilities. Third, our procedure can be used to assess dose-related trends, while the Monte Carlo hypergeometric test applied by Adams and Skopek (1987) and others is not designed to be sensitive to trends. Fourth, our approach allows for the natural incorporation of data from previous studies through elicited prior distributions. For commonly studied genes, mutational spectra databases containing thousands of mutations have been established and can be accessed through the internet (Cariello *et al.* 1997; Hutchison and Donnelan 1997). Such information can potentially enhance the sensitivity of statistical analyses.

In what follows, we describe the Bayesian hierarchical model, we outline tests for differences in the category probabilities and in the category-specific mutation frequencies, and we propose methods for incorporating historical data into the analysis. We illustrate our approach through application to data from a study of mutation induction in *lacI* transgenic mice after exposure to the flame retardant tris(2,3-dibromopropyl) phosphate (TDBP; De Boer *et al.* 1996).

THE STATISTICAL MODEL

Modeling the mutation frequency: Consider an experiment involving s mutational classes ($i = 1, \ldots, s$) and t treatments or groups under study ($j = 1, \ldots, t$). For group j, let v_j be the number of tissues or cell cultures that are examined for mutations, let m_{jk} be the number of cells (or plaques) in tissue k that have detectable mutations, and let c_{jk} be the number of cells in tissue k that are screened for mutations ($j = 1, \ldots, t$; $k = 1, \ldots, v_j$). The standard estimate of the mutant frequency in group j is the mutant fraction, $\sum_{k=1}^{v_j} m_{jk} / \sum_{k=1}^{v_j} c_{jk}$.

For group j and tissue k, let z_{jk} denote the number of mutants that are sequenced out of m_{jk}, and let n_{jk} denote the number of mutants out of z_{jk} that remain after removing all recurrent mutations. An estimate of the mutation frequency in group j, which was originally proposed by Carr and Gorelick (1996), is the number of independent mutations ($n_j = \sum_{k=1}^{v_j} n_{jk}$) divided by the effective number of cells at risk ($r_j = \sum_{k=1}^{v_j} c_{jk} z_{jk} / m_{jk}$).

We assume that the number of independent mutants follows a binomial distribution,

$$\Pr(N_j = n_j \mid r_j, \phi_j) = \binom{r_j}{n_j} \phi_j^{n_j} (1 - \phi_j)^{r_j - n_j}, \quad \text{for } n_j = 0, 1, 2, \ldots, r_j,$$

where ϕ_j is the mutation frequency in group j. For reasons that will become clear, r_j does not need to be an integer. To represent the uncertainty in ϕ_j before conducting the current study, we assign ϕ_j a Beta prior distribution with parameters γ_j and β_j (Gelman *et al.* 1996). The resulting posterior distribution for ϕ_j is equivalent to the posterior that would have been obtained had we chosen a noninformative Beta(0, 0) prior for ϕ_j and then added an additional γ_j independent mutants and β_j normal cells to the group j data; that is, the Beta(γ_j, β_j) prior contains equivalent information to γ_j independent mutants out of $\gamma_j + \beta_j$ cells. Therefore, $\gamma_j + \beta_j$ can be considered the prior sample size.

The prior parameters can be chosen on the basis of data from previous studies, as we illustrate later in the article. Alternatively, a subjective prior can be chosen by setting the prior mean $\gamma_j / (\gamma_j + \beta_j)$ equal to the investigator's best guess for ϕ_j and choosing the prior variance (or sample size) to reflect the uncertainty in this choice. If relevant historical data or substantive information are not available, then a noninformative prior can be specified by setting γ_j and β_j equal to very small positive numbers. On the basis of exploratory analyses, we recommend using $\gamma_j = \beta_j = 0.001$, though setting γ_j and β_j to slightly lower or higher values should have no noticeable effect on analyses. Traditional noninfor-

mative priors, such as the Bayes-Laplace uniform prior ($\gamma = \beta = 1$) or Jeffrey's prior ($\gamma = \beta = 0.5$; Gelman *et al.* 1996), can result in noticeable bias in estimates of the mutation frequency when the number of independent mutants is small.

Conditional on the prior and on the data from the current study, the posterior distribution of the mutation frequency ϕ_j is Beta with parameters $\gamma_j + n_j$ and $\beta_j + r_j - n_j$:

$$f(\phi_j | n_j, r_j, \gamma_j, \beta_j) = \frac{\Gamma(\gamma_j + \beta_j + r_j)}{\Gamma(\gamma_j + n_j)\Gamma(\beta_j + r_j - n_j)} \times \phi_j^{\gamma_j + n_j - 1}(1 - \phi_j)^{\beta_j + r_j - n_j - 1}. \quad (1)$$

This posterior distribution quantifies the current information about the mutation frequency in group j. Point and interval estimates can easily be calculated to summarize this posterior. In the case where a completely noninformative prior is chosen, the posterior mean $\hat{\phi}_j = (\gamma_j + n_j)/(\gamma_j + \beta_j + r_j)$ will equal the maximum-likelihood estimate n_j/r_j. Otherwise, the posterior mean will equal a weighted average of the prior mean $\gamma_j/(\gamma_j + \beta_j)$ and the maximum-likelihood estimate. Tests can be formulated on the basis of the posterior distributions for the mutation frequencies within the different groups, as we illustrate in this article.

Modeling the category probabilities: The mutants that are sequenced can be classified according to position within the gene and/or type of genetic damage. The counts of the number of independent mutants falling into each category within each group form an $s \times t$ contingency table, with the rows representing mutation categories $i = 1, \ldots, s$ and the columns representing groups $j = 1, \ldots, t$. We let y_{ij} denote the number of mutants that are in category i out of the n_j independent mutants that are sequenced in group j ($i = 1, \ldots, s$; $j = 1, \ldots, t$). We make the standard assumption that the jth column of the contingency table $\mathbf{y}_j = (y_{1j}, \ldots, y_{sj})$ has a multinomial sampling distribution with parameters n_j and $\boldsymbol{\pi}_j = (\pi_{1j}, \ldots, \pi_{sj})$. The probability that a mutation in group j falls into category i is π_{ij}.

Since the overall mutation frequency is extremely small, the sampling distribution of the total number of independent mutants (n_j) is approximately Poisson with mean $r_j\phi_j$ under expression (1). Under the multinomial conditional distribution for the numbers of mutations in each category ($\mathbf{y}_j$), the unconditional distribution for the number of mutations in category i (y_{ij}) is approximately Poisson with mean $r_j\lambda_{ij}$, where $\lambda_{ij} = \phi_j\pi_{ij}$ is the mutation frequency corresponding to category i ($i = 1, \ldots, s$). Thus, our hierarchical model provides a unified framework for incorporating mutant fraction and DNA sequence information into analyses of the overall mutation frequency (ϕ_j), the category probabilities ($\boldsymbol{\pi}_j$), and the category-specific mutation frequencies ($\lambda_{1j}, \ldots, \lambda_{sj}$).

To represent the uncertainty in the category probabilities before conducting the current study, we assign $\boldsymbol{\pi}_j$ a Dirichlet prior distribution with parameters $\mu_{1j}, \ldots, \mu_{sj}$ (Gelman *et al.* 1996). The resulting posterior distribution for $\boldsymbol{\pi}_j$ is equivalent to the posterior that would have been obtained had we chosen a noninformative Dirichlet $(0, \ldots, 0)$ prior for $\boldsymbol{\pi}_j$ and then added an additional $\mu_{1j}, \ldots, \mu_{sj}$ mutations to categories $i = 1, \ldots, s$ of the group j data. Therefore, the prior contains equivalent information to $\mu_j = \sum_{i=1}^{s}\mu_{ij}$ mutations in group j with μ_{ij} of type i ($i = 1, \ldots, s$).

The prior parameters can be chosen on the basis of data from previous studies, as we illustrate in this article. Alternatively, a subjective prior can be chosen by setting each μ_{ij}/μ_j equal to the investigator's best guess at the proportion of mutations falling into category i in group j. The prior sample size μ_j can then be chosen to reflect uncertainty in this choice. In the absence of historical or substantive information, a noninformative prior can be chosen by setting the prior parameters equal to small positive numbers. On the basis of exploratory analyses, we recommend using $\mu_{1j} = \ldots = \mu_{sj} = 0.01$. However, using slightly higher or lower values should have no noticeable effect on the analytic results, and the sensitivity to the specified values drops off rapidly as the number of sequenced mutants increases.

Conditional on the prior and on the data from the current study, the posterior distribution of the category probabilities in group j is Dirichlet with parameters $\mu_{1j} + y_{1j}, \mu_{2j} + y_{2j}, \ldots, \mu_{sj} + y_{sj}$:

$$f(\pi_{1j}, \ldots, \pi_{sj} \mid \mathbf{y}_j, n_j, \mu_{1j}, \ldots, \mu_{sj}) = \frac{\Gamma(\mu_j + n_j)}{\prod_{i=1}^{s}\Gamma(\mu_{ij} + y_{ij})}\prod_{i=1}^{s}\pi_{ij}^{\mu_{ij} + y_{ij} - 1}. \quad (2)$$

This posterior distribution quantifies the current information about the category probabilities in group j. Point and interval estimates can easily be calculated to summarize this posterior. In the case where a completely noninformative prior is chosen, the posterior mean $\hat{\pi}_{ij} = (\mu_{ij} + y_{ij})/(\mu_j + n_j)$ equals the maximum-likelihood estimate y_{ij}/n_j. Otherwise, the posterior means for the category probabilities will equal a weighted average of the prior means and the maximum-likelihood estimates. The posterior distributions described in expressions (1) and (2) can be used for statistical inference about the category probabilities and the category-specific mutation frequencies, as we illustrate in the next section. They can also be used to obtain point and interval estimates of any function of the mutational parameters. Such estimates do not rely on large sample approximations and can be extremely useful in characterizing differences between spectra.

STATISTICAL TESTS

Tests of homogeneity in the category probabilities: Tests for differences in mutational spectra between groups can be based on either the category probabilities $\pi_{1j}, \ldots, \pi_{sj}$ or the category-specific mutation frequencies

$\lambda_{1j}, \ldots, \lambda_{sj}$. The hypothesis of homogeneity in the category probabilities can be expressed as

$$H_{01}: \pi_{i1} = \pi_{i2} = \ldots = \pi_{it} \quad \text{(for all } i\text{)}.$$

A natural measure of deviation from H_{01} is the Pearson chi-square goodness-of-fit statistic,

$$X^2 = d(\mathbf{y}) = \sum_{i=1}^{s}\sum_{j=1}^{t} \frac{(y_{ij} - n_j\hat{\pi}_i)^2}{n_j\hat{\pi}_i}, \tag{3}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ denotes the observed mutation counts and

$$\hat{\pi}_i = \sum_{j=1}^{t}(\mu_{ij} + y_{ij}) / \sum_{j=1}^{t}(\mu_j + n_j)$$

denotes the estimated probability that a randomly selected mutant falls in category i under H_{01}. Classical tests compare X^2 to a χ^2 reference distribution, which approximates the posterior of X^2 under the null hypothesis in large samples. The large sample approximation can perform poorly with mutational spectrum data sets, since the expected number of mutations is often low within some of the categories (Adams and Skopek 1987).

Alternatively, exact P values can be calculated using either a conditional approach (Agresti 1992) or an unconditional approach (Bayarri and Berger 1999). The most common example of the conditional approach is Fisher's exact test, which conditions on the row and column totals. We use the following unconditional P value here,

$$P = \Pr\{d(\mathbf{Y}) \geq d(\mathbf{y})\},$$

where $\mathbf{Y}$ denotes the mutation counts that would have been observed had H_{01} been true and had the experiment been replicated under the same conditions. A simple Monte Carlo procedure can be used to estimate P:

1. Sample $\mathbf{Y}_j$ from a multinomial distribution with parameters n_j and $\hat{\pi}_1, \ldots, \hat{\pi}_s$ for groups $j = 1, \ldots, t$ and calculate $d(\mathbf{Y})$.
2. Repeat 1 for a large number of iterations, and let P equal the proportion of samples where $d(\mathbf{Y}) \geq d(\mathbf{y})$.

When a noninformative prior is specified for the category probabilities, this procedure is similar to the Monte Carlo hypergeometric test with one distinction: we do not condition upon the number of mutants per category (row totals). Since the number of mutants per category is not fixed by design, allowing the row totals to vary better represents the true sampling distribution of the data and can potentially result in an increase in power (Agresti 1990).

Tests that use measures of deviation that are not designed to be sensitive to dose-related trends, including the Adams and Skopek (1987) Monte Carlo hypergeometric test and the unconditional test we just described, may fail to detect important effects. Suppose that the category-specific mutation frequencies increase with dose and that the rate of increase is category dependent. This scenario will result in an increasing trend in the proportion of mutations in categories with a relatively high rate of increase and a decreasing trend in categories with a relatively low rate of increase. We expect that this scenario is quite common, since the mutability of DNA can vary substantially across sites in a genome (Foster *et al.* 1982). The following measure of deviation from H_{01} is sensitive to trends in the category probabilities,

$$d(\mathbf{y}) = \sum_{i=1}^{s}\left[\frac{\{\Sigma_{j=1}^{t}(x_j - \bar{x})(y_{ij} - n_j\hat{\pi}_i)\}^2}{\hat{\pi}_i(1 - \hat{\pi}_i)\Sigma_{j=1}^{t} n_j(x_j - \bar{x})^2}\right], \tag{4}$$

where $x_1, \ldots, x_t$ are the dose levels for treatment groups $j = 1, \ldots, t$ and $\bar{x} = \Sigma_j x_j n_j / \Sigma_j n_j$. This measure of deviation is the sum across mutational categories of the category-specific Cochran-Armitage score test statistics (Cochran 1954; Armitage 1955). By using deviation measure (4) instead of measure (3) when implementing steps 1 and 2 of the Monte Carlo procedure described above, a P value can be estimated for testing H_{01} against the alternative hypothesis of a trend in the category probabilities with dose.

Tests of homogeneity in the category-specific mutation frequencies: As an alternative to hypothesis H_{01}, we could test the following hypothesis of homogeneity in the category-specific mutation frequencies:

$$H_{02}: \lambda_{i1} = \lambda_{i2} = \ldots = \lambda_{it} \quad \text{(for all } i\text{)}.$$

The Pearson goodness-of-fit statistic can be used as a measure of deviation from H_{02},

$$d(\mathbf{y}) = \sum_{i=1}^{s}\sum_{j=1}^{t} \frac{(y_{ij} - r_j\hat{\lambda}_i)^2}{r_j\hat{\lambda}_i}, \tag{5}$$

where $\hat{\lambda}_i = \hat{\pi}_i \Sigma_{j=1}^{t} (\gamma_j + n_j) / \Sigma_{j=1}^{t} (\gamma_j + \beta_j + r_j)$ for $i = 1, \ldots, s$. This measure is not sensitive to trends in the mutation frequencies. An alternative measure of deviation from H_{02} that is sensitive to increasing dose-related trends is

$$d(\mathbf{y}) = \sum_{i=1}^{s}\left[\frac{\Sigma_{j=1}^{t} x_j(y_{ij} - r_j\hat{\lambda}_i)}{\{\hat{\lambda}_i\Sigma_{j=1}^{t} r_j(x_j - \bar{x})^2\}^{1/2}}\right]. \tag{6}$$

This measure of deviation is the sum across mutational categories of the category-specific Armitage (1955) score test statistics. The P value for testing H_{02} can be estimated as follows:

1. Sample Y_{ij} from a Poisson distribution with mean $r_j\hat{\lambda}_i$ for categories $i = 1, \ldots, s$ and groups $j = 1, \ldots, t$, and calculate $d(\mathbf{Y})$.
2. Repeat 1 for a large number of iterations. The estimated P value is the proportion of samples where $d(\mathbf{Y}) \geq d(\mathbf{y})$.

When using measure of deviation (5), this procedure estimates a P value for testing the null hypothesis of homogeneity in the category-specific mutation frequencies (H_{02}) against the unordered alternative hypothesis of any difference between groups. When using measure of deviation (6), this procedure estimates a P value for testing H_{02} against the alternative hypothesis of an overall increase in the mutation frequencies with dose.

EXAMPLE

We illustrate the proposed approach through application to data from a study of the flame-retardant TDBP (De Boer *et al.* 1996). In this study, *lacI* transgenic male B6C3F1 mice (Big Blue) were used to examine mutation induction in the kidney, liver, and stomach after exposure to 0 mg/kg, 150 mg/kg (2 days), 300 mg/kg (4 days), or 600 mg/kg (4 days) of TDBP. There were six mice in the control group and five mice in each of the exposure groups. Animals were sacrificed 14 days after the last dose of TDBP. Tissues were removed from the animals and were later examined for mutations. The authors concluded that exposure to TDBP induced tissue-specific mutations in the kidney that were distinct from spontaneous mutations.

These data were later reanalyzed by Brackley *et al.* (1999) to explore the use of log-linear models for analyzing mutational spectra data. On the basis of a Cochran-Mantel-Haenszel test (CMH; Agresti 1990) they concluded that there was an ordinal effect of TDBP dose on the mutational spectra ($P = 0.021$). The CMH test has similar drawbacks to the Pearson goodness-of-fit test in that it can perform poorly when data are sparse. In the TDBP study, a relatively large number of mutants were sequenced in each group. However, many of the categories had fewer than five mutations, a commonly used cutoff for chi-square tests (Agresti 1990), raising concern about the validity of the CMH test.

We reanalyze the kidney data here. The estimated mutation frequencies for the control, low, medium, and high dose groups were, respectively,

$$2.8 \times 10^{-5},\ 3.4 \times 10^{-5},\ 5.5 \times 10^{-5},\ 4.9 \times 10^{-5},$$

after correction for potential clonal expansion. Table 1 lists the DNA alterations by class. We compared the category probabilities in each dose group with the control group using both a Monte Carlo hypergeometric test and our proposed test with a noninformative prior. The P values from the Monte Carlo hypergeometric test were 0.526, 0.671, and 0.135 for the 150, 300, and 600 mg/kg dose groups, respectively. The comparable P values based on our procedure were 0.461, 0.676, and 0.129, respectively. We also tested for a dose response trend in the category probabilities using our proposed trend test. The estimated P value based on 5000 Monte Carlo samples was $P = 0.011$ (99% confidence interval, $0.007 < P < 0.015$), suggesting a highly significant dose-related trend in the category probabilities. This effect was not apparent based on the conventional Monte Carlo hypergeometric test.

TABLE 1

Mutational spectra of independently recovered *lacI* mutations in the kidney of Big Blue mice exposed to TDBP

Class	Control 60[a] (81)[b]	2 × 150 79[a] (86)[b]	4 × 300 92[a] (100)[b]	4 × 600 89[a] (96)[b]
G:C → A:T	33	39	40	31
A:T → G:C	1	3	5	4
G:C → T:A	12	21	14	14
G:C → C:G	4	4	8	5
A:T → T:A	2	2	4	5
A:T → C:G	1	2	4	1
Frameshift −1	2	7	9	13
Frameshift +1	1	0	2	4
Others[c]	4	1	7	12

Data from De Boer *et al.* (1996).
[a] The number of independent mutants after correction for possible clonal expansion.
[b] The number of mutants sequenced.
[c] Includes deletions, insertions, and complex changes.

Differences in the category probabilities can be difficult to interpret due to the constraint that the probabilities must sum to one in each dose group. Therefore, we reanalyzed the TDBP data to assess trends in the category-specific mutation frequencies. The estimated P value from our proposed trend test was $P = 0.0004$ (99% confidence interval, $0.000 < P < 0.001$), suggesting that treatment with TDBP causes a highly significant increase in the frequency of one or more types of *lacI* mutations. To identify differences in the rate of increase between mutational classes, we estimated posterior summaries of the slope parameters characterizing the change in the class-specific mutation frequencies with dose (Table 2). These posterior summaries were estimated from repeated samples, which were obtained by first sampling from the posterior distribution of the mutation frequencies (λ_{i1}, λ_{i2}, λ_{i3}, λ_{i4}) and then calculating the slope. The estimated slopes are positive for each mutational class, and treatment with TDBP causes significant increases in the frequency of A:T → G:C transitions, A:T → T:A transitions, frameshifts (both +1 and −1), and mutations in the category including deletions, insertions, and complex changes.

EXTENSION: INCORPORATING HISTORICAL DATA

Choosing the prior for the mutation frequency: We have described statistical models that quantify prior uncertainty in the mutation frequency and in the category probabilities using probability distributions. For commonly studied genes, mutational spectrum databases

TABLE 2

Posterior summaries of the slope parameters characterizing the change in the class-specific mutation frequencies with dose of TDBP (mg/kg)

Class	Estimated slope[a]	SD	90% interval	*P* value[b]
G:C → A:T	4.424	6.598	(−6.431, 15.491)	0.250
A:T → G:C	2.970	2.056	(0.040, 6.675)	0.047
G:C → T:A	2.277	4.348	(−4.570, 9.596)	0.305
G:C → C:G	2.262	2.565	(−1.562, 6.739)	0.181
A:T → T:A	3.486	2.339	(0.047, 7.582)	0.047
A:T → C:G	0.284	1.19	(−1.456, 2.453)	0.419
Frameshift −1	10.484	3.605	(5.087, 16.764)	0.000
Frameshift +1	3.435	2.023	(0.618, 7.096)	0.021
Others[c]	9.443	3.498	(4.229, 15.465)	0.001

Data from De Boer *et al.* (1996).
[a] Expressed as 10^{-9} per plaque per unit increase in dose.
[b] Represents posterior probabilities of a negative slope.
[c] Includes deletions, insertions, and complex changes.

containing thousands of mutations are available (Carello *et al.* 1997; Hutchison and Donnelan 1997). These data can be used to choose prior distributions, and thus information from previous studies can be incorporated into analyses of data from a current study.

Suppose that data are available from h previous studies that involve similar experimental conditions to group 1 of the current study, where group 1 is a reference or control group. For study l, let n_{1l} be the number of independent mutations, and let r_{1l} be the effective number of cells at risk ($l = 1, \ldots, h$). If we could assume that the mutation frequency (*i.e.*, the mutant frequency corrected for clonal expansion) in each historical study is identical to ϕ_1, the mutation frequency in group 1 of the current experiment, then we could set $\gamma_1 = \Sigma_{l=1}^{h} n_{1l}$ and $\beta_1 = \Sigma_{l=1}^{h} (r_{1l} - n_{1l})$. This approach is equivalent to pooling the data from all the studies. Such an approach can produce misleading results in the presence of variability between studies.

We instead assume that the mutation frequency in each previous study is a random variable from a distribution centered on the mutation frequency in group 1 of the current study. Prentice *et al.* (1992) made a similar assumption in developing statistical methods for incorporating historical control data into trend tests for dichotomous data. This formulation enables borrowing of information across studies and accounts for heterogeneity among studies in the mutation frequency.

Let ϕ_{1l} denote the mutation frequency in study l, and let $\bar{\phi}_1 = \Sigma_{l=1}^{h} r_{1l}\phi_{1l} / \Sigma_{l=1}^{h} r_{1l}$ denote the pooled mutation frequency. We assign the study-specific mutation frequencies a Beta(a, b) prior density. As described earlier in the article, a noninformative prior can be chosen by setting a and b close to 0. The prior for ϕ_1, the mutation frequency in the current study, is chosen on the basis of the posterior densities for the study-specific mutation frequencies $\phi_{11}, \ldots, \phi_{1h}$. In the presence of variability between studies, data from a past experiment do not contain as much information about the current mutation frequency as data from the current experiment. In choosing the prior for ϕ_1, we weight experiment l according to the ratio of estimated mean square errors,

$$u_l = \frac{\bar{\phi}_1}{\phi_{1l} + r_{1l}(\phi_{1l} - \bar{\phi}_1)^2}, \tag{7}$$

which represents the information about ϕ_1 in experiment l relative to the information that would have been available had r_{1l} cells been added to the current study that would have been sequenced had they contained detectable mutations in the gene of interest. Our proposed prior for ϕ_1 is given by

$$\phi_1 \sim \text{Beta}\left(\gamma_1 = \sum_{l=1}^{h} u_l n_{1l},\ \beta_1 = \sum_{l=1}^{h} u_l (r_{1l} - n_{1l})\right). \tag{8}$$

This prior assigns each historical study a weight between 0 and 1, where 0 indicates that the data from a particular past study are completely noninformative about the current mutation frequency and 1 indicates that data from the past and current studies can be pooled. The overall weight assigned to the historical data is inversely proportional to the magnitude of variability between studies.

To incorporate prior (8) into the Monte Carlo analyses described earlier in the article, two alternative methods can be used: (1) a plug-in approach or (2) a fully hierarchical approach. To implement the plug-in approach, simply plug in

$$\hat{\phi}_{1l} = \frac{a + n_{1l}}{a + b + r_{1l}} \quad \text{and} \quad \hat{\phi}_1 = \frac{\Sigma_{l=1}^{h}(a + n_{1l})}{\Sigma_{l=1}^{h}(a + b + r_{1l})}$$

for ϕ_{1l} and $\bar{\phi}_1$, respectively, in expression (7) and use the resulting weights to estimate γ_1 and β_1. This plug-in approach is simple to implement but does not account for error in estimating the weights. To instead

use the fully hierarchical approach, add the following steps to the Monte Carlo sampling procedure (before step 1):

i. Sample ϕ_{1l} from Beta($a + n_{1l}$, $b + r_{1l} - n_{1l}$) for $l = 1, \ldots, h$, and then calculate $\bar{\phi}_1$.
ii. Calculate u_l, $l = 1, \ldots, h$, and then γ_1 and β_1 conditional on the sampled ϕ_{1l}'s.

This approach accounts for uncertainty in estimation of γ_1 and β_1.

Choosing the prior for the category probabilities: We follow a similar approach to choose the prior for the category probabilities. We let y_{i1l} denote the number of mutations of type i out of the n_{1l} independent mutants in study l ($l = 1, \ldots, h$). We assume that the category probabilities in study l are random variables from a distribution centered on the category probabilities in group 1 of the current study. The probability that a mutation in experiment l is of type i is π_{i1l}, and the pooled probability that a mutation is of type i is $\bar{\pi}_{i1} = \Sigma_{l=1}^{h} n_{1l}\pi_{i1l}/\Sigma_{l=1}^{h} n_{1l}$. We assign the study-specific category probabilities $\boldsymbol{\pi}_{1l} = (\pi_{11l}, \pi_{21l}, \ldots, \pi_{s1l})$ a Dirichlet $(c_1, \ldots, c_s)$ prior density, where $c_1, \ldots, c_s$ can be set close to 0 to specify a noninformative prior. The prior for the category probabilities in group 1 of the current study ($\boldsymbol{\pi}_1$) is chosen on the basis of the posterior densities for the study-specific category probabilities. In formulating this prior, we weight experiment l according to the ratio of estimated mean square errors,

$$w_l = \frac{\Sigma_{i=1}^{s}\bar{\pi}_{i1}(1 - \bar{\pi}_{i1})}{\Sigma_{i=1}^{s}\{\pi_{i1l}(1 - \pi_{i1l}) + n_{1l}(\pi_{i1l} - \bar{\pi}_{i1})^2\}}, \qquad (9)$$

which represents the information about $\boldsymbol{\pi}_1$ in experiment i relative to the information that would have been available had n_{1l} additional independent mutants been sequenced in the current study. Our proposed prior for $\boldsymbol{\pi}_1$ is given by

$$\boldsymbol{\pi}_1 \sim \text{Dirichlet}(\mu_{11} = \sum_{l=1}^{h} w_l y_{11l},\ \mu_{21} = \sum_{l=1}^{h} w_l y_{21l}, \ldots, \mu_{s1} = \sum_{l=1}^{h} w_l y_{s1l}). \qquad (10)$$

This prior assigns each historical study a weight between 0 and 1, where 0 indicates that the mutations from a particular past study provide no information about the current category probabilities and 1 indicates that the past mutations are as informative as mutations in the current study. The overall weight assigned to the historical data is inversely proportional to the magnitude of variability between historical studies in the category probabilities. To incorporate the historical data into the Monte Carlo analyses described earlier in the article, we can use a plug-in or fully hierarchical approach. To implement the plug-in approach, simply plug in

$$\hat{\pi}_{i1l} = \frac{c_i + y_{i1l}}{n_{1l} + \Sigma_{m=1}^{s} c_m} \quad \text{and} \quad \hat{\pi}_{i1} = \frac{\Sigma_{l=1}^{h}(c_i + y_{i1l})}{\Sigma_{l=1}^{h}(n_{1l} + \Sigma_{m=1}^{s} c_m)}$$

for π_{i1l} and $\bar{\pi}_{i1}$, respectively, in expression (9) and use the resulting weights in estimating the prior parameters $\mu_{11}, \ldots, \mu_{s1}$. To instead follow a fully hierarchical approach, which accounts for error in estimating the prior parameters, add the following steps to the Monte Carlo sampling procedure (before step 1):

i. Sample $\boldsymbol{\pi}_{1l}$ from Dirichlet $(c_1 + y_{11l}, \ldots, c_s + y_{s1l})$ for $l = 1, \ldots, h$.
ii. Calculate w_l, $l = 1, \ldots, h$, and then $\mu_{11}, \ldots, \mu_{s1}$ conditional on the sampled $\boldsymbol{\pi}_{1l}$'s.

DISCUSSION

We have proposed a new Bayesian framework for the analysis of data from mutational spectra experiments. Our approach allows for the incorporation of data from previous studies without requiring restrictive assumptions of homogeneity across studies. The inclusion of historical data can potentially result in substantial improvements in the sensitivity of statistical tests, particularly when the data are sparse (Tarone 1982; Haseman *et al.* 1984; Fung *et al.* 1996). As mutations are extremely rare, data from several previous studies may be needed to detect a difference between groups in the frequency of mutation at a particular site within a gene. For example, without information from past studies, an increase from 0 mutants of a given type to 1 or 2 mutants of a given type will typically be judged to be nonsignificant. However, if no mutants of this type have been observed in any of several previous studies, 1 or 2 mutants may represent a true (and possibly biologically important) increase. Though including historical data can enable the detection of small absolute differences in mutation frequency, it may also lead to an inflation of the type I error rate if proper correction is not made for the variability between studies. We have described such a correction in this article, and in future work we plan to fully evaluate the operating characteristics of this approach.

Within our modeling framework, we have described easy-to-implement Monte Carlo test procedures for assessing differences between groups in the frequencies of mutation within categories defined by type of DNA alteration and/or position of the mutated base pair and in the proportions of mutants that fall within each of these categories. These tests are an alternative to the widely used Adams and Skopek (1987) analysis. When historical data are not available and interest focuses on differences between two groups in the category probabilities, our method should have modestly increased power relative to the Adams and Skopek test, since we do not condition on the number of mutants per category. However, simulation studies are needed to assess the magnitude of the difference in power under a variety of scenarios. In addition to allowing for the incorporation of historical data, our method can be expected to have substantially increased power relative to the Adams

and Skopek method in two common situations. First, when mutational spectra data are collected for several dose groups, our procedure allows testing for a dose-related trend in the category probabilities. As we have illustrated, a trend test can be much more sensitive than the conventional approach of separately comparing each dose group to the control group. Second, when interest focuses on assessing differences in the mutation frequency between groups and there is variability between the frequency of mutations at specific sites within a target gene, our procedure for testing for overall differences in the category-specific mutation frequencies should have improved power relative to methods based on the overall mutant frequency (*e.g.*, Carr and Gorelick 1994) or on the category probabilities (*e.g.*, Adams and Skopek 1987). Assessing the magnitude of this difference under a variety of scenarios is an area for future research.

A distinguishing feature of our approach is that exact estimates can be obtained for any function of the mutational parameters. As we illustrate in the example, such estimates can be extremely useful in characterizing differences between spectra and between mutational categories. For simplicity in presentation of the modeling framework, this article has not considered the incorporation of covariates, such as sex, age, tissue type, and species, into models for the mutation frequency and the category probabilities. However, covariates can easily be incorporated using dichotomous and multinomial response models, such as the logistic and the probit (see, for example, Chib and Greenberg 1998; Dunson 2000). Extended models that accommodate covariates and extrabinomial (see Piegorsch *et al.* 1994, 1997) or multinomial variation can be fit using BUGS (Best *et al.* 1996), a freely available program for Bayesian inference.

We thank David Umbach, Norman Kaplan, Joseph Haseman, and three anonymous reviewers for their many helpful suggestions.

LITERATURE CITED

Adams, W. T., and T. R. Skopek, 1987 Statistical test for comparison of samples from mutational spectra. J. Mol. Biol. **194:** 391–396.

Agresti, A., 1990 *Categorical Data Analysis.* John Wiley & Sons, New York.

Agresti, A., 1992 A survey of exact inference for contingency tables. Stat. Sci. **7:** 131–177.

Agresti, A., D. Wackerly and J. Boyett, 1979 Exact conditional tests for cross-classifications: approximation of attained significance levels. Psychometrika **44:** 75–84.

Armitage, P., 1955 Tests for linear trends in proportions and frequencies. Biometrics **11:** 375–386.

Bayarri, M. J., and J. O. Berger, 1999 P-values for composite null models. Discussion Paper, Institute of Statistics and Decision Science, Duke University, Durham, NC.

Best, N. G., D. J. Spiegelhalter, A. Thomas and C. E. G. Brayne, 1996 Bayesian analysis of realistically complex models. J. R. Stat. Soc. B **159:** 323–342.

Brackley, M. E., J. G. De Boer and B. W. Glickman, 1999 Use of log-linear analysis to construct explanatory models for TDBP- and AFB1-induced mutation spectra in *lacI* transgenic animals. Mutat. Res. **425:** 55–69.

Cariello, N. F., G. R. Douglas, M. J. Dycaico, N. J. Gorelick, G. S. Provost *et al.*, 1997 Databases and software for the analysis of mutations in the human *p53* gene, the human *hprt* gene and both the *lacI* and the *lacZ* gene in transgenic rodents. Nucleic Acids Res. **25:** 136–137.

Carr, G. J., and N. J. Gorelick, 1994 Statistical tests of significance in transgenic mutation assays: considerations on the experimental unit. Environ. Mol. Mutagen. **24:** 276–283.

Carr, G. J., and N. J. Gorelick, 1995 Statistical design and analysis of mutation studies in transgenic mice. Environ. Mol. Mutagen. **25:** 246–255.

Carr, G. J., and N. J. Gorelick, 1996 Mutational spectra in transgenic animal research: data analysis and study design based upon the mutant or mutation frequency. Environ. Mol. Mutagen. **28:** 405–413.

Chib, S., and E. Greenberg, 1998 Analysis of multivariate probit models. Biometrika **85:** 347–361.

Cochran, W. G., 1954 Some methods of strengthening the common χ^2 tests. Biometrics **10:** 417–451.

De Boer, J. G., J. C. Mirsalis, G. S. Provost, K. R. Tindall and B. W. Glickman, 1996 Spectrum of mutations in kidney, stomach, and liver from *lacI* transgenic mice recovered after treatment with tris(2,3-dibromopropyl)phosphate. Environ. Mol. Mutagen. **28:** 418–423.

Dunson, D. B., 2000 Bayesian latent trait models for clustered mixed outcomes. J. R. Stat. Soc. B **62:** 355–366.

Foster, P. L., E. Eisenstadt and J. Cairns, 1982 Random components in mutagenesis. Nature **299:** 365–367.

Fung, K. Y., D. Krewski, J. N. K. Rao and A. J. Scott, 1994 Tests for trend in developmental toxicity experiments with correlated binary data. Risk Anal. **14:** 639–648.

Fung, K. Y., D. Krewski and R. T. Smythe, 1996 A comparison of tests for trend with historical control in carcinogen bioassay. Can. J. Stat. **24:** 431–454.

Fung, K. Y., X. Lin and D. Krewski, 1998 Use of generalized linear mixed models in analyzing mutant frequency data from the transgenic mouse assay. Environ. Mol. Mutagen. **31:** 48–54.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1996 *Bayesian Data Analysis.* Chapman & Hall, London.

Haseman, J. K., J. Huff and G. A. Boorman, 1984 Use of historical control data in carcinogenicity studies in rodents. Toxicol. Pathol. **12:** 126–135.

Hutchison, F., and J. E. Donellan, Jr., 1997 A mutation spectra database for bacterial and mammalian genes. Nucleic Acids Res. **25:** 192–195.

Ibrahim, J. G., L. M. Ryan and M.-H. Chen, 1998 Using historical controls to adjust for covariates in trend tests for binary data. J. Am. Stat. Assoc. **93:** 1282–1293.

Nishino, H., D. J. Schaid, V. L. Buettner, J. Haavik and S. S. Sommer, 1996 Mutation frequencies but not mutant frequencies in Big Blue mice fit a Poisson distribution. Environ. Mol. Mutagen. **28:** 414–417.

Piegorsch, W. W., and A. J. Bailer, 1994 Statistical approaches for analyzing mutational spectra: some recommendations for categorical data. Genetics **136:** 403–416.

Piegorsch, W. W., A. C. Lockhart, B. H. Margolin, K. R. Tindall, N. J. Gorelick *et al.*, 1994 Sources of variability in data from a *lacI* transgenic mouse mutation assay. Environ. Mol. Mutagen. **23:** 17–31.

Piegorsch, W. W., A. C. Lockhart, G. J. Carr, B. H. Margolin, T. Brooks *et al.*, 1997 Sources of variability in data from a positive selection lacZ transgenic mouse mutation assay: an interlaboratory study. Mutat. Res. **388:** 249–289.

Prentice, R. L., R. T. Smythe, D. Krewski and M. Mason, 1992 On the use of historical control data to estimate dose response trends in quantal bioassay. Biometrics **48:** 459–478.

Roff, D. A., and P. Bentzen, 1989 The statistical analysis of mitochondrial DNA polymorphisms—χ^2 and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

Tarone, R. E., 1982 The use of historical control information in testing for a trend in proportions. Biometrics **38:** 215–220.

Communicating editor: H. Ochman