Clustering Mutational Spectra via Classification Likelihood and Markov Chain Monte Carlo Algorithms

Mario MEDVEDOVIC, Paul SUCCOP, Rakesh SHUKLA, and Kathleen DIXON

We have analyzed a set of 39 mutational spectra of the supF gene that were generated by different mutagenic agents and under different experimental conditions. The cluster analyses was performed using a newly developed clustering procedure. The clustering criterion used in the procedure was developed by applying the classification likelihood approach to multinomial observations. We also developed a Gibbs sampling-based optimization procedure that outperformed previously developed methods in a comparative simulation study. The results of the cluster analysis showed that our clustering procedure was able to recreate natural grouping of the mutational spectra with respect to the characteristics of mutagenic agents used to generate them and with respect to experimental conditions applied in the process of generating spectra. These results are an important confirmation of the relevance of mutational spectra in characterizing mutagenic mechanisms of different carcinogens.

Key Words: Classification expectation maximization algorithm; Cluster analysis; Gibbs sampling; Multinomial distribution; Stochastic expectation maximization algorithm.

1. INTRODUCTION

Characterizing mutagenic mechanisms of different carcinogens is one of the greatest challenges of modern toxicology. One of the general approaches to studying the mutagenesis of different agents relies on examining the type and the distribution of mutations induced in a target gene by specific mutagenic treatments. In such experiments, the stretch of DNA containing the target gene is first exposed to a mutagenic treatment that will cause certain types of premutagenic lesions to form. In the second step, such treated DNA is introduced into a selected cell line where some of the lesions induced in the first step will cause

Mario Medvedovic is Senior Research Associate, Department of Environmental Health, University of Cincinnati Medical Center, 3223 Eden Avenue, Cincinnati, OH 45267-0056 (E-mail: medvedm@email.uc.edu). Paul Succop is Associate Professor, Department of Environmental Health, University of Cincinnati. Rakesh Shukla is Professor, Department of Environmental Health, University of Cincinnati. Kathleen Dixon is Associate Professor, Department of Environmental Health, University of Cincinnati.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 6, Number 1, Pages 19–37

mutations to be generated during replication of the damaged DNA. The distribution of induced point mutations within the target gene is referred to as the mutational spectrum of the mutagenic agent used to cause the initial damage to DNA.

In numerous studies involving different mutagenic agents, it has been confirmed that different agents generally give rise to different mutational spectra. It is generally assumed that the type and the distribution of initial lesions induced by a mutagen and characteristics of replication and repair machinery present in the cell line used in the experiment are both reflected in the observed mutational spectrum. Hence, differences and similarities of mutational spectra can be assumed to reflect differences and similarities in the mechanisms of action of corresponding mutagens and cell lines used in experiments.

One of the systems frequently used to study mechanisms of mutagenesis, developed by Seidman, Dixon, Razzaque, Zagursky, and Berman (1985), uses the bacterial tyrosine suppressor tRNA gene (supF) as the mutational target. We have collected and analyzed 39 published mutational spectra generated after treating the supF target gene with different types of mutagens. Similarities and differences between different mutational spectra were assessed using a newly developed clustering procedure. The goal was to confirm that mutational spectra contain sufficient information on the type of initial damage to allow for identifying groups of spectra generated by mutagens with similar mechanisms of action and under similar experimental conditions. This was done by examining clusters of mutational spectra generated by our clustering procedure. The optimal number of clusters was selected based on the interpretability of observed clusters and the statistical significance of the comparison between the model with the optimal number of clusters and models with fewer clusters (Aitkin, Anderson, and Hinde 1981).

The mutational spectrum observed in an experiment is characterized by counts of point mutations observed at each of the positions (bases) in the target gene. If the target gene is M bases long and y_k denotes the number of point mutations observed at the kth position, $k = 1, \ldots, M$, the mutational spectrum is defined by the vector $\mathrm{y} = (y_1, \ldots, y_M)$.

The statistical model for mutational spectra has been described by Adams and Skopek (1987). In their model, y is assumed to be an observation from a multinomial random variable $\mathrm{Y} \sim \mathrm{mult}(p_1, \ldots, p_M, N)$, where $N = y_1 + \cdots + y_M$ and p_k represents the probability that an observed point mutation occurs at the kth position. In general, the goal of the statistical analysis of such data is to make inferences about the parameters of the distribution $\mathrm{p} = (p_1, \ldots, p_M)$ based on the observed data y. Two mutational spectra, generated by multinomial random variables with parameters p_1 and p_2, are considered to be equal if $\mathrm{p}_1 = \mathrm{p}_2$. Otherwise, they are considered to be different. According to the model we used for the cluster analysis, mutational spectra from a single cluster were generated by multinomial distributions with the same probability vector.

The clustering procedure we used for clustering mutational spectra is based on the classification likelihood as the clustering criterion (McLachlan and Basford 1987). Maximizing the classification likelihood criterion is generally a difficult discrete optimization problem. Celleux and Govaert (1992) described a general maximization algorithm for this problem that they called the classification expectation maximization (CEM) algorithm. CEM is an iterative algorithm that will converge to the global solution of the maximization problem if started from the initial values sufficiently close to the global solution. Otherwise, it will converge to a local maximum of the clustering criterion. Celleux and Govaert (1992) also proposed two approaches to combining CEM with the stochastic expectation maximization (SEM) algorithm in order to alleviate this problem. In our initial tests, none of these three algorithms offered satisfactory results. Therefore, we also developed an algorithm for maximizing the classification likelihood based on the Gibbs sampler (Gelfand and Smith 1990) and the CEM algorithm. We compared its performance in a comparative simulation study to the performance of three algorithms developed by Celleux and Govaert. Based on results of this simulation study, we decided to use the Gibbs sampler-based algorithm for the analysis of our data.

2. CLASSIFICATION LIKELIHOOD

Suppose that T mutational spectra $\mathrm{X} = (\mathrm{x}_1', \ldots, \mathrm{x}_T')$ were generated as independent observations from Q $(Q < T)$ different multinomial random variables defined with parameters $\mathrm{p}_1, \ldots, \mathrm{p}_Q$, i.e., for each $i \in \{1, \ldots, T\}$, there exists a unique $j \in \{1, \ldots, Q\}$ such that x_i is a realization of the multinomial random variable with the probability density function

$$f_{\mathrm{mult}}(\mathrm{y} \mid \mathrm{p}_j) + \frac{N_i!}{y_1! \cdots y_M!} p_{j1}^{y_1} \cdots p_{jM}^{y_M}, \tag{2.1}$$

where $N_i = x_{i1} + \cdots + x_{iM} = y_1 + \cdots + y_M$ represents the total number of observed mutations in the ith spectrum. Each of the parameters $\mathrm{p}_1, \ldots, \mathrm{p}_Q$ of probability density functions in (2.1) defines a cluster of mutually similar mutational spectra. Let z_{ij} be the indicator variable stating whether the ith spectrum belongs to the jth cluster; $z_{ij} = 1$ if the ith spectrum belongs to the jth cluster and $z_{ij} = 0$ otherwise. The assignment vector $\mathrm{z}_i = (z_{i1}, \ldots, z_{iQ})'$ defines completely to which cluster the ith vector belongs, and the classification matrix $\mathrm{Z} = (\mathrm{z}_1, \ldots, \mathrm{z}_T)$ completely defines the distribution of all T spectra among Q clusters. Let π_j denote the proportion of mutational spectra coming from the jth cluster, $\pi_j = (\Sigma_{i=1}^T z_{ij})/T$. Prior to taking into account x_i, the probability of observing the assignment vector z_i is $p(\mathrm{z}_i) = \Pi_{j=1}^Q \pi_j^{z_{ij}}$. It is further assumed that $\mathrm{x}_1, \ldots, \mathrm{x}_T$ given $\mathrm{z}_1, \ldots, \mathrm{z}_T$, respectively, are conditionally independent and

$$p(\mathrm{x}_i \mid \mathrm{z}_i) = \sum_{j=1}^{Q} \mathrm{z}_{ij} f_{\mathrm{mult}}(\mathrm{x}_i \mid \mathrm{p}_j) \quad \text{for any } i = 1, \ldots, T.$$

Hence, the classification likelihood for the data is given by

$$L_C(\mathrm{Z}, \mathrm{P}) = \prod_{i=1}^{T} \sum_{j=1}^{Q} \mathrm{z}_{ij} \pi_j f_{\mathrm{mult}}(\mathrm{x}_i \mid \mathrm{p}_j),$$

where $\mathrm{P} = (\mathrm{p}'_1, \ldots, \mathrm{p}'_Q)$, and the classification log likelihood is

$$l_C(\mathrm{Z}, \mathrm{P}) = \sum_{i=1}^{Q} \sum_{j=1}^{T} \mathrm{z}_{ij}(\log \pi_j + \log f_{\mathrm{mult}}(\mathrm{x}_i \mid \mathrm{p}_j)).$$

In terms of the EM algorithm for the finite mixture (Dempster, Laird, and Rubin 1977), (Z, X) represents the complete data and $L_C(\cdot)$ is the probability distribution of the complete data. $l_C(\cdot)$ as the clustering criterion in the context of clustering Gaussian data was introduced by Symons (1981). Bryant (1991) refers to it as the penalized classification maximum likelihood (CML) criterion. It can be shown that, for Gaussian data, under certain conditions, maximizing $l_C(\cdot)$ is equivalent to maximizing the traditional variance criterion (Celleux and Govaert 1992) and to maximizing the information criterion in the context of latent class models for discrete data (Celleux and Govaert 1991). The clustering procedure consists of identifying $(\mathrm{Z}^*, \mathrm{P}^*)$ maximizing $l_C(\cdot)$.

3. MAXIMIZATION ALGORITHMS

3.1 CLASSIFICATION EXPECTATION MAXIMIZATION (CEM) ALGORITHM

Celleux and Govaert (1992) described three different algorithms for maximizing $l_C(\cdot)$. The first method they called the classification expectation and maximization algorithm (CEM). Following the general EM approach as described by Dempster et al. (1977), the algorithm consists of alternating expectation (E), classification (C), and maximization (M) steps. The algorithm generates the sequence $(\mathrm{Z}_n, \mathrm{P}_n)$, $n = 0, 1, \ldots$, which, under certain conditions, converges to the set of optimal parameters $(\mathrm{Z}^*, \mathrm{P}^*)$ that maximize $l_C(\Omega)$. The algorithm proceeds as follows.

E Step. Given current parameters $(\mathrm{Z}_n, \mathrm{P}_n)$, the current posterior probabilities that x_i belongs to the jth cluster are calculated as

$$\begin{aligned} t_{jn}(\mathrm{x}_i) &= p(z_{ij} = 1, z_{ik} = 0, k = 1, \ldots, j-1, j+1, \ldots, G \mid \mathrm{Z}_n, \mathrm{P}_n, x_i) \\ &= \frac{\pi_{jn} f_{\mathrm{mult}}(\mathrm{x}_i \mid p_{jn})}{\sum_{l=1}^{Q} \pi_{ln} f_{\mathrm{mult}}(\mathrm{x}_i \mid p_{ln})}, \end{aligned}$$

where

$$\pi_{jn} = \left(\sum_{i=1}^{T} z_{ijn} \right) \Big/ T. \qquad (3.1)$$

C Step. Assign each x_i to the cluster k that provides the maximum posterior probability, i.e., set $z_{ik,n+1} = 1$ and $z_{il,n+1} = 0$ for all $l \neq k$.

M Step. Given the current assignments, calculate the maximum likelihood estimates (MLE) of parameters for multinomial random variables defining each cluster based on the data currently classified in the corresponding cluster,

$$\mathrm{p}_{j,n+1} = \frac{\sum_{i=1}^{T} z_{ij,n+1}\mathrm{x}_i}{\sum_{k=1}^{M}\sum_{i=1}^{T} z_{ij,n+1}x_{ik}}. \tag{3.2}$$

Since distribution parameters P_n are a function of the current classification matrix Z_n, the sequence $(\mathrm{Z}_n, \mathrm{P}_n)$ generated by the CEM algorithm can be represented with the corresponding sequence of classification matrices Z_n.

It turns out that the CEM algorithm shares advantages and shortcomings of the general EM algorithm (Celleux and Govaert 1992). The sequence $l_C(\mathrm{Z}_n)$ is increasing in each step of the algorithm, i.e., $l_C(\mathrm{Z}_n) \leq l_C(\mathrm{Z}_{n+1})$ for each $n > 0$. Since there is a finite number of different classification matrices, this implies that, for any initial classification matrix Z_0, there exists m such that $\mathrm{Z}_{m+1} = \mathrm{Z}_m$. In other words, P_m is a stationary point for the sequence Z_n. Furthermore, if the initial point of the algorithm is sufficiently close (details are given by Celleux and Govaert (1992)) to Z^*, then the sequence Z_n generated by the CEM algorithm will converge to Z^*. On the other hand, the number of stationary points can be large and to which one the algorithm will converge depends on the choice of the initial classification matrix Z_0.

3.2 Stochastic Expectation Maximization (SEM) Algorithm

In order to alleviate the problem of severe dependence on the initial values, Celleux and Govaert (1992) suggested the use of the stochastic expectation maximization (SEM) algorithm for generating initial parameters for the CEM algorithm. The SEM algorithm has been proposed by Celleux and Diebolt (1985) as an alternative to the EM algorithm for estimating parameters of finite mixtures. In the SEM algorithm, the C step of the CEM algorithm is replaced by the stochastic (S) step.

S Step. Assign x_i to the jth cluster with probability $t_{j,n}(\mathrm{x}_i)$ for $j = 1, \ldots, Q$ and $i = 1, \ldots, T$. In this way, x_i is not necessarily classified into the cluster with the highest posterior probability. It can be shown that the sequence of mixture parameter estimates generated by the SEM algorithm forms a homogeneous Markov chain for which ergodicity holds (Celleux and Diebolt 1985). Furthermore, if ML estimates of mixture parameters are the only stable fixed point for the corresponding EM algorithm, then means of random samples obtained from the stationary distribution of such Markov chains are asymptotically normally distributed with the mean equal to the ML estimates of the mixture parameters. However, since ML estimates are usually not the only fixed point of the corresponding EM algorithm, the properties of the stationary distribution remain unclear. The hybrid algorithm suggested by Celleux and Govaert (1992) uses the SEM algorithm to generate a number of assignment matrices. The assignment matrix yielding the highest l_C among those generated by the SEM algorithm is then used as the starting position for the CEM algorithm. The ability of the SEM algorithm to generate a starting position close enough to the optimal solution after a reasonable number of iterations remains questionable.

3.3 Classification Annealing Expectation Maximization (CAEM) Algorithm

An alternative approach to the idea of combining SEM and CEM algorithms, suggested by Celleux and Govaert (1992), is the classification annealing EM algorithm (CAEM). In this algorithm, the current posterior probabilities that x_i belongs to the jth group are calculated as

$$t_{jn}(\mathrm{x}_i) = \frac{\{\pi_{jn} f_{\mathrm{mult}}(\mathrm{x}_i, p_{jn})\}^{1/\tau_n}}{\sum_{l=1}^{Q} \{\pi_{ln} f_{\mathrm{mult}}(\mathrm{x}_i, p_{jn})\}^{1/\tau_n}},$$

where $\tau_n \to 0$ as $n \to \infty$ and τ_n is a monotonously decreasing sequence. The classification of observations to clusters is again performed by random assignment according to their corresponding posterior probabilities. It is easy to see that, when setting $\tau_0 = 1$, the algorithm starts out as the SEM algorithm and, in the limit, turns into the CEM algorithm. Using the simulated annealing terminology (Van Laarhoven and Aarts 1987), the sequence τ_n is referred to as a sequence of temperatures. The CAEM algorithm does share certain similarities with a general simulated annealing algorithm, i.e., it starts out as a random algorithm and, in the limit as $\tau_n \to 0$, turns into a deterministic optimization algorithm that increases the objective function in each step. As Celleux and Govaert concluded in their article, CAEM seems to have a certain advantage over the traditional simulated annealing approach since it "accepts" all new parameters, unlike usual simulated annealing algorithms. Unfortunately, the convergence properties of the CAEM algorithm are generally unknown.

3.4 Gibbs Sampling-Based Algorithms

Since the previously described algorithms have unsatisfactory convergence properties and performed poorly in initial tests, we have developed two optimizational algorithms based on the Gibbs sampler. Consider the following formulation of the clustering likelihood in which parameters $(\mathrm{p}_j;\, j = 1, \ldots, Q)$ are replaced by corresponding estimates based only on the current assignments:

$$L_{CG}(\mathrm{Z}) = \prod_{i=1}^{T} \prod_{j=1}^{Q} z_{ij} \pi_j f_{\mathrm{mult}}(\mathrm{x}_i \mid \mathrm{p}_j(\mathrm{Z})),$$

where $\mathrm{p}_j(\mathrm{Z})$ is defined by Equation (3.2). The equivalence of maximizing $L_{CG}(\cdot)$ and $L_C(\cdot)$ follows directly from the fact that the $(\mathrm{Z}^*, \mathrm{P}^*)$ is the stationary point for the CEM algorithm. Now, if we define the constant

$$C = \sum_{\text{all possible } Z\text{'s}} L_{CG}(\mathrm{Z}),$$

then

$$g(\mathrm{Z}) = \frac{L_{CG}(\mathrm{Z})}{C}$$

defines a probability distribution function on the set of all possible assignments Z. Furthermore, finding the assignment Z^* that maximizes $L_{CG}(\cdot)$ now corresponds to finding the assignment Z^* with the highest probability $g(\mathrm{Z}^*)$. The basic idea of the newly proposed approach to maximizing $L_{CG}(\cdot)$ is to use Gibbs sampling (Gilks, Richardson, and Spiegelhalter 1996) to generate a sample of observations from the probability density function (p.d.f.) $g(\cdot)$ and then to identify the assignments with high probabilities of occurring by evaluating the obtained sample. If a sufficiently large sample of assignments is created, it is reasonable to assume that the assignment Z^* with the highest probability of occurring will be present in the sample. The Gibbs sampling approach to generating samples from complex multivariate distributions proceeds as follows. Suppose that $Z_{-i} = (\mathrm{z}_1, \ldots, \mathrm{z}_{i-1}, \mathrm{z}_{i+1}, \ldots, \mathrm{z}_T)$ and

$$\begin{aligned} g_i(\mathrm{z}_i \mid \mathrm{Z}_{-i}) = p(\mathrm{z}_i \mid \mathrm{Z}_{-i}) &= \frac{p(\mathrm{z}_i, \mathrm{Z}_{-i})}{p(\mathrm{Z}_{-i})} \\ &= \frac{g(\mathrm{z}_i, \mathrm{Z}_{-1})}{\sum_{j=1}^{Q} g(z_{ij} = 1, \mathrm{Z}_{-i})} = \frac{L_{CG}(\mathrm{z}_i, \mathrm{Z}_{-i})}{\sum_{j=1}^{Q} L_{CG}(z_{ij} = 1, \mathrm{Z}_{-i})} \end{aligned}$$

defines the conditional distribution of the assignment for the ith spectrum given assignments of all other spectra. Given a current assignment Z, the iterated assignment is produced by generating updated assignments for individual spectra from their conditional distributions, given the current assignments for the rest of the spectra; i.e., given the rest of the current assignments Z_{-i}, the ith spectrum is assigned to the jth cluster with probability $g_i(z_{ij} = 1 \mid \mathrm{Z}_{-i})$. The sequence of assignments Z_n generated in such a way represent observations from an ergodic Markov chain with the stationary distribution being equal to $g(\cdot)$. Consequently, $\mathrm{prob}(\mathrm{Z}_n = \mathrm{Z}) \rightarrow g(\mathrm{Z})$ as $n \rightarrow \infty$, regardless of the initial assignment Z_0. Therefore, the assignments generated by this approach, after an appropriate burn-in period, can be assumed to form a sample from the distribution $g(\cdot)$. The larger $L_{CG}(\mathrm{Z})$, the more likely it is that the assignment Z will be included in the generated sample. The burn-in period refers to the number of assignments needed to be generated before the Markov chain converges to its stationary distribution. The advantage of the Gibbs sampling approach over SEM and CAEM algorithms is in its appropriate convergence properties. The algorithm using the Gibbs sampler to find optimal assignment we call the Gibbs sampling maximization (GSM) algorithm.

The number of distinct partitions of T observations into Q nonempty groups can be calculated using the following (Everitt 1993):

$$N(T, Q) = \frac{1}{Q!} \sum_{j=1}^{Q} (-1)^{Q-j} \binom{Q}{j} j^T.$$

For example, $N(40, 4) \approx 5.0 \times 10^{22}$. It is clear that, if $\mathrm{Z}_K^* = \mathrm{argmax}\{L_{CG}(\mathrm{Z}_n), n = 1, \ldots, K\}$, where $\mathrm{Z}_1, \ldots, \mathrm{Z}_n$ are generated by the Gibbs sampler starting from Z_0, then $\mathrm{prob}(\mathrm{Z}_K^* = \mathrm{Z}^*) \to 1$ as $K \to \infty$ for any starting initial point Z_0. However, due to the huge size of the sample space for the random process generating Z's, it is not clear that any sample generated in a reasonable time will be sufficiently large to give us a high probability of $\mathrm{Z}_K^* = \mathrm{Z}^*$. One way of assessing whether the sample size K is large enough is to monitor Z_K^* for different initial points Z_0. If the algorithm has the same Z_K^* for each of a number of different initial points, it is a good indication that K is large enough and that $\mathrm{Z}^* = \mathrm{Z}_K^*$.

The alternative approach to the use of the Gibbs sampler in finding Z^* is to use assignments generated by the Gibbs sampler as starting points for the CEM algorithm. In this approach, the ability of the Gibbs sampler to generate a representative sample from the distribution $g(\cdot)$ is combined with the ability of the CEM algorithm to identify the assignment with the highest likelihood in the neighborhood of the initial assignment. The primary objective of the Gibbs sampler is no longer to identify the most likely assignment, but to generate good (i.e., proximal to Z^*) initial assignments for the CEM algorithm. This is based on the fact that the CEM algorithm will converge to the global maximum if the initial classification matrix is close to Z^*. The idea behind the expectation that GSM will offer a good starting point is based on the expectation that the global maximum is surrounded by an area of classifications with relatively high likelihoods. While this would be an obvious fact for a continuous likelihood function, it is not necessarily the case for the classification likelihood.

In a case when there are many different stationary points with relatively large likelihoods, it is important for the Gibbs sampler to generate initial positions close to all such assignments. However, the Gibbs sampler can generally experience difficulties in moving between high probability areas of the state space (in our case, assignments with high likelihoods) when such areas are separated by regions of very low probabilities. This can cause the Gibbs sampler to move slowly through the sample space and fail to identify all areas of high probability, causing it to generate a sample strongly dependent on the initial values. A simple way to alleviate this problem is to modify the sampling distribution $g(\cdot)$ in such a way that differences in probabilities between regions of high probabilities and regions with low probabilities are reduced while the rankings of assignments with respect to their probabilities (likelihoods) remain the same. One way to achieve this is by transforming $g(\cdot)$ (Jennison 1993) as

$$g^*(\mathrm{Z}) = \frac{g(\mathrm{Z})^\xi}{C(\xi)},$$

where $C(\xi) = \Sigma_{\text{all possible } Z\text{'s}}\, g(\mathrm{Z})^\xi$ and $\xi < 1$. It is easy to show that the corresponding conditional distributions for $g^*(\cdot)$ are given by

$$g_i^*(\mathrm{z}_i \mid Z_{-i}) = \frac{g(\mathrm{z}_i \mid Z_{-1})^\xi}{\sum_{j=1}^{Q} g(z_{ij} = 1 \mid Z_{-i})^\xi}.$$

Obviously, for any two assignments Z' and Z'', $g(Z') \leq g(Z'')$ will imply $g^*(Z') \leq g^*(Z'')$. Hence, finding the assignment that maximizes $g(\cdot)$ is equivalent to finding the assignment that maximizes $g^*(\cdot)$. On the other hand, for $\xi < 1$, the Gibbs sampler based on $g^*(\cdot)$ is generally able to move between different areas of high probabilities quicker than the Gibbs sampler based on $g(\cdot)$ (Jennison 1993). For $\xi = 0$, the Gibbs sampler algorithm turns into a procedure for random generation of possible assignments with each assignment having the same probability of being generated. An appropriate $\xi < 1$ might significantly increase the speed with which the Gibbs sampler moves from one to another area of assignments with high probabilities while keeping the difference in probabilities between assignments with high likelihood and those with small likelihood large enough in order to generate as few as possible assignments with small likelihoods. In our experimentations with different ξ's, the $\xi = 0.1$ seemed to be appropriate for the clustering problem considered. The algorithm combining the Gibbs sampler and the CEM algorithm we call the Gibbs classification expectation maximization (GCEM) algorithm.

4. SIMULATION STUDY

Performances of the five algorithms for maximizing the clustering criterion and the performance of the classification likelihood as the clustering criterion were examined through a simulation study. T observed mutational spectra were generated as observations from Q different multinomial distributions representing different groups of similar mutagenic mechanisms. If n_j denotes the number of observations from the jth group represented by $\mathrm{mult}(p_{1j}, \ldots, p_{Mj}; N_{ji})$, then $\Sigma\, n_j = T$ and N_{ji} represent the number of mutations generated in the ith observed spectrum from the jth group. For the whole simulation study, the number of positions at which a mutation could occur, representing the dimension of the multinomial distributions used to generate the data, was fixed to be $M = 10$. This decision greatly reduced the computational burden in the simulation study. Other parameters specifying simulated data were varied in order to ensure a relatively wide applicability of the conclusions. For each combination of parameters, 10 different data sets and 10 random initial arrangements of T generated spectra in Q clusters were generated. Each of the five algorithms were then applied to each data set starting from each of the 10 initial arrangements. Parameters defining the type of the data generated were the number of clusters ($Q = 2, 3, 4$), the number of spectra in each cluster ($n_1, \ldots, n_Q = 5, 10, 20$), and the number of mutations in each spectrum ($N_{ji} = 5, 10, 20$, $j = 1, \ldots, Q$; $i = 1, \ldots, n_j$). The relative probabilities for multinomial distribution functions corresponding to the each of the clusters $(p_{1j}, \ldots, p_{Mj})$ were chosen to define different degrees of separation between the Q clusters. The degree of separation, or to what degree the multinomial distributions defining different clusters were different, was based on the effect size index e for contingency tables defined by Cohen (1969). For the two multinomial distributions $\mathrm{mult}(p_{1j}, \ldots, p_{Mj}; N)$, $j = 1, 2$, with the same number of M-dimensional Bernoulli experiments N, e is defined as

$$e((p_{11},\ldots,p_{M1};N)(p_{12},\ldots,p_{M2};N)) = \sum_{i=1}^{M}\sum_{j=1}^{2}\frac{(p_{ij}-p_i^*)^2}{p_i^*},$$

where

$$p_i^* = (p_{i1}+p_{i2})/2.$$

In his book, Cohen defines the difference between two multinomial distributions to be small if $e \leq 0.05$, medium if $e \approx 0.1$, and large if $e \approx 0.2$. The relative probabilities for different clusters in this simulation study were chosen in such a way that the effect size for the pairwise difference of two clusters was $e \approx 0.05, 0.1, 0.2$. Furthermore, the relative probabilities were chosen in such a way that one category was assumed to be a "hot spot" and had higher than average relative probability of a mutation while the remaining nine positions were "cold spots," with equal but smaller relative probabilities. The differences in multinomial distributions for different clusters were due to differences in the location of the hot spot. The hot-spot and cold-spot probabilities for corresponding e are given in Table 1.

The number of GSM, GCEM, SEM, and CAEM samples used in determining the optimal assignment was 1,000. The number of iterations for the CEM algorithm was limited to 1,000. However, the CEM algorithm converged to a stationary point in less than 20 iterations in all cases considered in this simulation study. The sequence of temperatures τ_i, $i = 0, 1, \ldots,$ for the CAEM algorithm was defined as in Celleux and Govaert (1992), $\tau_0 = 1$, and $\tau_{i+1} = a\tau_i$, where $a = 0.97$. The adjustment of $\xi = 0.1$ was used in the GCEM algorithm. One problem in estimating multinomial probability parameters using the maximum likelihood estimators as given in (3.2) is the fact that $p_{jk,n}$ will be zero if none of the spectra currently classified in the jth cluster had any observed mutations at the kth position. This will then prevent any spectrum having at least one mutation at the kth position from being classified into the jth cluster regardless of how well the number of mutations at other positions in this spectrum agree with the $\mathrm{p}_{j,n}$. One way around this problem is to use the Bayesian estimates assuming a uniform prior. In this case, the estimates for p's are given by

$$\mathrm{p}_{j,n+1} = \frac{\left(\sum_{i=1}^{T} z_{ij,n+1}\mathrm{x}_i\right) + 1_M}{\left(\sum_{k=1}^{M}\sum_{i=1}^{T} z_{ij,n+1}x_{ik}\right) + M}, \tag{4.1}$$

where 1_M is an M-dimensional row vector of ones. In the same way, the estimates of proportions of spectra belonging to a cluster can be modified in order to avoid the sequence Z_n being stuck within the part of the sample space for which one of the π_j's is zero, i.e., (3.1) can be replaced by

$$\pi_{jn} = \left(1 + \sum_{i=1}^{T} z_{ijn}\right) \Big/ (T+Q). \tag{4.2}$$

Table 1. Relative Probabilities in the Simulation Study Corresponding to Different Levels of the Separation Between Clusters

e	p_{hot}	p_{cold}
0.05	0.19	0.09
0.1	0.28	0.08
0.2	0.37	0.07

In the initial investigation, the algorithms using probability estimates as given in (4.1) and (4.2) performed better than equivalent algorithms using the maximum likelihood estimates. Therefore, adjusted estimates were used in all algorithms.

The results of the simulation study comparing performances of the five algorithms are given in Table 2. The average number of times the algorithm reached $Z_{\max}$ for each of 10 generated data sets, starting from 10 different initial positions, reflects the ability of the algorithm to consistently find $Z_{\max}$ regardless of the starting position. The performance of the GCEM algorithm is clearly superior to the other four algorithms, with this algorithm producing the same or a better result than any other algorithm in all 81 scenarios. The GSM algorithm produced the same or a better result than CEM, SEM, and CAEM in 63 out of 81 scenarios. The performances of CAEM and SEM were rather similar, while CEM, as expected, showed the worst performance of the five. While such a comparison, without taking into account the computational complexities of the algorithms, might seem unfair, it did clearly identify the best algorithm to use, given that the computational burden is manageable. Therefore, we decided to use the GCEM algorithm in the analysis of the actual data.

The results pertaining to the ability of the classification likelihood to recreate the original structure of the data are given in Table 3. These results suggest that $Z_{\max}$ will offer a reasonable approximation for the underlying data structure when clusters are rather well separated and there is a relatively large number of mutations in each spectrum.

When comparing the performance of the GCEM algorithm across different simulation scenarios, several factors seem to be influencing its ability to consistently identify optimal assignments. The most influential and consistent effect seems to be the pairwise level of separation between clusters (e). The better the separation of clusters (larger e), the more consistent is GCEM in identifying $Z_{\max}$. In a few situations for which the average $Z_{\max}$ was larger for lower levels of separation, keeping all other parameters fixed, the differences were minimal and likely to have arisen due to chance. The other factor influencing the performance of the GCEM algorithm is the size of the parameter space, which depends on two factors: the number of clusters and the number of spectra per cluster. Out of these two parameters, the number of clusters had a more transparent effect in our simulation study. For all other parameters fixed, GCEM performed better for smaller numbers of clusters in all situations. The effect of the number of spectra per cluster seems to be less straightforward. This is probably due to the fact that, while increasing the number of spectra per cluster is increasing the size of the parameter space and in this way makes it more difficult for GCEM

Table 2. Results of the Simulation Study Pertaining to the Ability of Algorithms to Find $Z_{\max}$ Out of 10 Trials for Each of 10 Simulated Data Sets; Average Number of Times the Algorithm Reached $Z_{\max}$

			Effect size index (e)								
			2.0			*1.0*			*0.5*		
			Number of mutations per spectrum								
Number of clusters	*Number of spectra per cluster*	*Algorithm*	*20*	*10*	*5*	*20*	*10*	*5*	*20*	*10*	*5*
2	5	GCEM	10	10	10	10	10	10	10	10	10
		GSM	10	10	9.5	9	9.2	9.9	7.9	8.4	9
		SEM	7.1	4.1	2.4	4.1	3.6	1.6	3.2	4	1.1
		CAEM	6.1	3.4	2.1	3.3	2.5	1.9	2.7	3.1	1.2
		CEM	5.4	2	1.1	2.2	1.3	0.5	0.1	0.5	0.7
	10	GCEM	10	10	10	10	10	10	10	10	10
		GSM	9.4	7	7.4	8.2	2	6.9	10	8.3	4.4
		SEM	7.5	3.1	2.4	3.5	1.7	1.6	2.5	1.8	1
		CAEM	7.4	2.9	2.1	3.6	1.3	1.4	2	1.4	1
		CEM	6.5	2.2	0.7	2.6	0.5	0.1	0.4	0	0.2
	20	GCEM	10	10	9	10	10	9.8	8.9	9	9
		GSM	10	6	2.2	5.2	5	6	7.8	5.1	2.3
		SEM	6.8	2.2	1.6	3.3	2.4	0.9	3	1.7	1.1
		CAEM	6.8	2.3	1.4	3.2	1.6	1.1	2.2	1.1	1.4
		CEM	7.4	3.1	0.8	4.5	0.4	0	0	0	0.1
3	5	GCEM	10	10	10	10	10	10	8.7	9.7	9.6
		GSM	10	6.3	5.7	8.9	7.9	4.5	8.4	5.6	2.3
		SEM	5.5	3	0.2	2.5	1.2	0.9	4	1.5	0
		CAEM	3.9	1.4	0.7	1.2	0.6	0.8	2.2	0.7	0.1
		CEM	2.5	0	0.3	0.3	0	0.1	0.7	0	0.1
	10	GCEM	10	8.7	9.1	9.6	6.7	8.4	6	4.8	8
		GSM	8.5	2.4	2.5	5.5	1.3	0	3.2	2.1	1.1
		SEM	6.8	1.9	0.4	2.7	1.3	0.2	1.7	0.1	0.5
		CAEM	5	1.1	0.2	1.9	0.3	0.2	0.2	0.3	0.2
		CEM	2.5	0.1	0.2	0.5	0	0	0.1	0	0
	20	GCEM	10	10	8.4	10	6.3	4	1.5	3.2	2.4
		GSM	8.6	1.6	1.1	6.2	0.3	0	0	0	0
		SEM	8.5	2.6	0.1	4.1	0.5	0.1	0.7	0	0
		CAEM	6.6	1.1	0	2.3	0.6	0.1	0	0.1	0.1
		CEM	5.5	0	0	1.8	0	0	0	0	0
4	5	GCEM	10	9.9	7.4	8.7	6	6.2	3.8	5.5	6
		GSM	8.6	6.5	1.3	7.8	3.1	1.5	3.4	2.3	0.3
		SEM	4.8	2	0.3	2.4	0.8	0	1.5	1.1	0
		CAEM	2.2	0.7	0.2	0.3	0	0.2	0	0.4	0
		CEM	1	0	0	0	0	0	0	0	0
	10	GCEM	10	6.7	2.5	7.7	1.5	2.5	0.9	1.6	5.1
		GSM	7.9	2.8	0.2	2.3	0	0.2	0.7	0.4	0
		SEM	6.5	2.9	0	2.4	0.4	0	0.5	0	0
		CAEM	3.4	0.5	0.1	0.4	0	0	0	0	0
		CEM	1.7	0	0	0	0	0	0	0	0
	20	GCEM	10	8.4	2.3	8.7	0.8	1.3	0.7	1.3	1.2
		GSM	7.9	0	0	0	0	0	0	0	0
		SEM	8.7	1.1	0	0.7	0	0	0.6	0	0
		CAEM	5.8	0.9	0	0.9	0.4	0	0	0.1	0
		CEM	2.9	0	0	0	0	0	0	0	0

Table 3. Results of the Simulation Study Pertaining to the Ability of the Clustering Criteria to Recreate the Original Structure of the Data; Average Number of Misclassified Spectra in the $Z_{\max}$ Classification as the Percentage of the Total Number of Spectra

Number of clusters	Number of spectra per cluster	*Effect size index (e)*								
		2.0			*1.0*			*0.5*		
		Number of mutations per spectrum								
		20	*10*	*5*	*20*	*10*	*5*	*20*	*10*	*5*
2	5	1.0	12.0	27.0	18.0	25.0	31.0	32.0	41.0	37.0
	10	4.5	10.5	22.0	12.5	33.5	37.0	36.0	41.5	40.5
	20	1.8	10.5	24.3	7.0	31.3	37.5	40.3	43.3	43.8
3	5	6.7	25.3	33.3	23.3	34.7	34.7	45.3	48.7	52.7
	10	4.7	30.0	37.3	22.0	40.0	45.3	51.7	52.7	54.0
	20	3.0	15.5	33.0	10.5	41.3	53.7	53.0	58.3	58.8
4	5	10.5	31.5	37.0	34.5	42.5	52.0	52.0	54.5	54.5
	10	5.8	29.0	49.8	29.3	53.5	55.0	60.0	61.5	59.5
	20	4.1	17.9	42.9	19.3	52.9	59.5	61.0	66.5	66.0

to find $Z_{\max}$, the increase in the number of spectra is also adding more information to the data, resulting in better defined cluster centers. Similar effects of the parameters defining the size of the parameter space on the effectiveness of the clustering criterion in recreating the original structure of the data is evident in Table 3; i.e., an increase in the number of clusters, while keeping other parameters fixed, always resulted in the increase in the percentage of the misclassified spectra. On the other hand, there was no consistency in the effect of the number of spectra per cluster. Finally, the effect of the number of mutations per spectrum on the performance of the GCEM algorithm seems to be more pronounced for large and medium separation levels than it is for the low separation level. For example, the average number of times when GCEM was able to find $Z_{\max}$ was always higher than or equal for the high number of mutations (20) than for the low number of mutations (4.2) whenever $e = 2$ or $e = 1$. For poorly separated clusters ($e = 0.5$), the effect of the number of mutations per spectra was smaller and the trend seemed to be have been reversed. Since it seems counterintuitive that more information per spectrum can consistently result in a poorer performance of the maximization algorithm, this particular trend is likely a result of random fluctuations in the shape of the likelihood function for poorly separated clusters.

5. CLUSTER ANALYSIS OF SELECTED OBSERVED MUTATIONAL SPECTRA

The clustering procedure based on the classification likelihood and the GCEM algorithm was applied to a set of 39 observed mutational spectra identified through a literature search. The first group of mutagens consists of chemicals that, after interacting with DNA, create bulky adducts that could inhibit proper base pairing and/or interfere with DNA replication. Based on their affinities for reacting with a certain type of DNA nucleotides, these mutagens are further separated into four subgroups: chemicals reacting primarily with guanine bases (G), chemicals predominantly interacting with guanine but also interacting with adenine bases (Ga), chemicals with similar affinities for interacting with guanine and adenine bases (AG), and chemicals predominantly interacting with adenine bases but also interacting with guanine (Ag). The subgroup for a mutagen was determined from the literature prior to the analysis. The second group of mutational spectra used in the analysis consists of spectra generated after the target gene was exposed to ultraviolet light.

In the treatment of the clustering problem described so far, the number of clusters Q was assumed to be known. We performed a statistical assessment of the number of clusters present in the data by testing hypotheses that the model with $Q+1$ clusters fits data better than the model with Q clusters, where $Q \in \{1, 2, 3, \ldots\}$. The statistically optimal number of clusters $Q_{\max}$ was the integer such that models with $Q+1$ clusters fit data significantly better that the model with Q clusters for all $Q = 1, \ldots, Q_{\max} - 1$ and the model with $Q_{\max} + 1$ clusters did not fit data better than the model with $Q_{\max}$ clusters. The test statistic for testing the null hypothesis that the model with Q clusters is appropriate versus the alternative hypothesis that the model with $Q+1$ clusters was the log (classification) likelihood statistic

$$\lambda^* = -2[l_C(Z_Q^*) - l_C(Z_{Q+1}^*)],$$

where $l_C(Z_Q^*)$ is the maximum of $l_C(\cdot)$ assuming Q clusters and $l_C(Z_{Q+1}^*)$ assuming $Q+1$ clusters. The rejection region for this test of hypothesis can be constructed using the parametric bootstrap approach described for the mixture problem by Aitkin et al. (1981). In our problem, the null hypothesis of Q clusters was rejected in favor of the alternative hypothesis of $Q+1$ clusters whenever $\lambda^* > \lambda_{\max}$, where $\lambda_{\max} = \max(\lambda^t, t = 1, \ldots, 100)$ and λ^t is the log-likelihood statistic calculated for the tth data set simulated under the null hypothesis (assuming $Z = Z_Q^*$); i.e., the null hypothesis was rejected whenever the estimated p-value was less than 0.01. Each of the 100 simulated data sets was created by generating T random observations x_{iQ}^t, $i = 1, \ldots, T$, from corresponding multinomial distributions $\mathrm{mult}(\hat{\mathrm{p}}_{jQ}; N_i)$, where $\hat{\mathrm{p}}_{jQ}$ represents the estimated relative probabilities of mutations for spectra in the jth cluster, $j = 1, \ldots, Q$, and $z_{ij} = 1$. Parameters $\hat{\mathrm{p}}_{jQ}$ of Q multinomial distributions were estimated using the observed optimal clustering when assuming Q clusters, i.e., $\hat{\mathrm{p}}_{jQ}$ were calculated as

$$\hat{\mathrm{p}}_{jQ} = \frac{\left(\sum_{i=1}^{T} z_{ij}\mathrm{x}_i\right) + 1_M}{\left(\sum_{k=1}^{M}\sum_{i=1}^{T} z_{ij}x_{ik}\right) + M}, \qquad j = 1, \ldots, Q.$$

Some theoretical justification of this approach, in the context of normal mixtures, is provided by Feng and McClulloch (1996). λ^* and $\lambda_{\max}$ for $Q = 1, 2, 3, 4, 5$ are given in Table 4. The results of the cluster analysis for the model with five (C5) and six clusters (C6) are given in Table 5.

Table 4. Test Statistics and Critical Values for Testing the Hypothesis That the Number of Clusters Is Q Versus the Hypothesis That the Number of Clusters Is $Q + 1$

Number of clusters $(Q+1)$	*Log-likelihood ratio statistic* (λ^*)	*Maximum simulated statistics* $(\lambda_{\max})$
2	453	69
3	316	74
4	266	78
5	147	79
6	113	69

There were two reasons for considering the model with five clusters as optimal for describing our data, although there was an indication that more than five clusters were present ($\lambda^* = 113$ and $\lambda_{\max} = 69$ for $Q = 5$). The first reason was that this model produced very intuitive and easy-to-explain clusters, while the six-cluster model produced an additional cluster for which it was difficult to find an appropriate biological explanation. The other reason was that the number of mutations in each spectrum varied greatly (Table 5) and, based on the functional form of the classification likelihood, it seems obvious that spectra with a large number of mutations would have a larger influence on the classification likelihood than spectra with a small number of mutations. In order to take into account this potential source of bias, an additional analysis was performed. In this analysis, the observed mutational spectra having less than 247 mutations (all except spectrum 44) were augmented by additional randomly generated mutations so that the total number of mutations in each spectrum was 247. Additional mutations were imputed in such a way that the originally observed pattern of mutations was preserved in the augmented spectrum. In short, probabilities of mutations at each position were estimated based on the observed mutations and $N_{\text{imputed}} = 247 - N_{\text{observed}}$ additional mutations were randomly generated according to the estimated probabilities. The cluster analysis was repeated using augmented spectra. Results of the clustering analysis for models with three, four, and five clusters were identical to the results obtained by analyzing original spectra. On the other hand, the additional cluster in the six-cluster model (Table 5, column 6R) made more biological sense than the additional cluster generated by original spectra. The newly created cluster in this case consisted of mutational spectra induced by exposing DNA to ultraviolet radiation and replicating plasmids in HeLa cell extracts. All this led us to believe that, at this point, the model with five clusters is optimal for the data at hand. However, further investigation of the method using augmented mutational spectra in situations where there is a great deal of variation in the number of mutations seems to be warranted.

6. CONCLUSIONS

Results of the analysis suggest that our classification likelihood-based clustering procedure was able to recreate a natural grouping of the mutational spectra with respect to the characteristics of the mutagenic agent used to generate them and with respect to

Table 5. Results of the Cluster Analysis

Spect ID	*Target sites*	*Number of mutations*	*Mutagen type*	*Plasmid*	*Cell line*	*C5*	*C6*	*C6R*	*Reference*
24	GA	146	Adduct	PS189	Ad293	0	0	0	Bigger et al. (1992)
25	GA	86	Adduct	PS189	Ad293	0	0	0	Bigger et al. (1992)
26	GA	140	Adduct	PS189	Ad293	0	0	0	Bigger et al. (1992)
27	GA	106	Adduct	PS189	Ad293	0	0	0	Bigger et al. (1992)
28	GA	98	Adduct	PZ189	Ad293	0	0	0	Bigger et al. (1989)
6	Ga	93	Adduct	PS189	Ad293	1	1	1	Page et al. (1996b)
7	Ga	110	Adduct	PS189	Ad293	1	1	1	Page et al. (1996b)
15	G	93	Adduct	PS189	Ad293	1	1	1	Courtemanche and Anderson(1994)
17	Ga	98	Adduct	PS189	Ad293	1	1	1	Bigger et al. (1990)
18	G	85	Adduct	PS189	Ad293	1	1	1	Boldt et al. (1991)
19	G	48	Adduct	PS189	Ad293	1	1	1	Levy et al. (1992)
20	G	49	Adduct	PS189	Human fibroblasts	1	1	1	Mah et al. (1989)
22	G	68	Adduct	PZ189	Ad293	1	1	1	Yang et al. (1987)
23	G	54	Adduct	PZ189	Ad293	1	1	1	Yang et al. (1988)
29	G	75	Adduct	PS189	Ad293	1	1	1	Bigger et al. (1991)
3	G	39	Adduct	PZ189	CV-1	2	2	2	Roilides et al. (1988)
11	G	57	Adduct	PSP189	Ad293	2	2	2	Maccubin et al. (1997)
13	G	62	Adduct	PSP189	Human fibroblasts	2	2	2	Endo et al. (1994)
14	G	73	Adduct	PSP189	Human fibroblasts	2	2	2	Endo et al. (1994)
16	G	127	Adduct	PSP189	Ad293	2	2	2	Courtemanche and Anderson (1994)
30	Ga	97	Adduct	PSP189	Ad293	2	2	2	Page et al. (1996a)
1	Ag	96	Adduct	PSP189	Ad293	3	3	3	Szeliga et al. (1995)
2	Ag	105	Adduct	PSP189	Ad293	3	3	3	Szeliga et al. (1994)
4	Ag	84	Adduct	PSP189	Ad293	3	3	3	Page et al. (1995)
5	Ag	86	Adduct	PSP189	Ad293	3	3	3	Page et al. (1995)
31		71	UV	PZ189	XP-A	4	4	4	Bredberg et al. (1986)
33		28	UV	PZ189 form1	HeLa	4	4	5	Carty et al. (1995)
34		30	UV	PZ189R2	HeLa	4	4	5	Carty et al. (1995)
35		55	UV	PYZ289	HL18	4	4	4	Yagi et al. (1994)
36		84	UV	PZ189	WI38VA13	4	4	4	Yagi et al. (1991)
37		62	UV	PZ189	XP-A	4	4	4	Bredberg et al. (1986)
38		138	UV	PZ189	Monkey cells	4	4	4	Hauser et al. (1986)
39		67	UV	PZ189	XP-D	4	4	4	Seetharam et al. (1987)
40		80	UV	PZ189	GM0637	4	4	4	Bredberg et al. (1986)
42		73	UV	PZ189	XP-F	4	4	4	Yagi et al. (1991)
43		179	UV	PZ189	CV-1	4	4	4	Keyse et al. (1988)
32		44	UV	PZ189	HeLa	4	5	5	Carty et al. (1995)
41		137	UV	PSP189	XP-A	4	5	4	Parris et al. (1994)
44		247	UV	PZ189	CV-1	4	5	4	Keyse et al. (1988)

experimental conditions applied in the process of generating spectra (type of plasmid). These results are an important confirmation of the relevance of mutational spectra in characterizing mutagenic mechanisms of different carcinogens. Results of our simulation study show that classification likelihood is a good clustering criterion for clustering multinomial observations. Finally, the maximization algorithm combining Gibbs sampler and the deterministic CEM algorithm (GCEM) identified optimal clusterings more consistently than any other algorithm we examined.

[Received June 1999. Accepted March 2000.]

REFERENCES

Adams, W. T., and Skopek R. T. (1987), "Statistical Test for the Comparison of Samples From Mutational Spectra," *Journal of Molecular Biology,* 194, 391–396.

Aitkin, M., Anderson, D., and Hinde, J. (1981), "Statistical Modeling of Data on Teaching Styles," *Journal of the Royal Statistical Society, Series A,* 144, 419–461.

Bigger, C. A. H., Flickinger, A. J., St. John, J., Harvey, R. G., and Dipple A. (1991), "Preferential Mutagenesis at GC Base Pairs by the Anti 3,4-Dihydrodiol 1,2-Epoxide of 7-Methylbenz[*a*]anthracene," *Molecular Carcinogenesis,* 4, 176–179.

Bigger, C. A. H., Flickinger, D. J., Strandberg, J., Pataki, J., Harvey, R. G., and Dipple, A. (1990), "Mutational Specificity of the Anti 1,2-Dihydrodiol3,4-Epoxide of 5-Methylchrysene," *Carcinogenesis,* 11, 2263–2265.

Bigger, C. A. H., St. John, J., Yagi, H., Jerina, D. M., and Dipple, A. (1992), "Mutagenic Specificities of Four Stereoisomeric Benzo[*c*]phenanthrene Dihydrodiol Epoxides," *Proceedings of the National Academy of Science, USA,* 89, 368–372.

Bigger, C. A. H., Strandberg, J., Yagi, H., Jerina, D. M., and Dipple, A. (1989), "Mutagenic Specificity of a Potent Carcinogen, Benzo[*c*]phenanthrene (4R,3S)-Dihydrodiol (2S,1R)-Epoxide, Which Reacts With Adenine and Guanine in DNA," *Proceedings of the National Academy of Science, USA,* 86, 2291–2295.

Boldt, J., Chia-Miao Mah, M., Wang, Y.-C., Smith, B. A., Beland, F. A., Maher, V. M., and McCormick, J. J. (1991), "Kinds of Mutations Found When a Shuttle Vector Containing Adducts of 1,6-Dinitropyrene Replicates in Human Cells," *Carcinogenesis,* 12, 119–126.

Bredberg, A., Kraemer, K. H., and Seidman, M. M. (1986), "Restricted Ultraviolet Mutational Spectrum in a Shuttle Vector Propagated in Xeroderma Pigmentosum Cells," *Proceedings of the National Academy of Science, USA,* 83, 8273–8277.

Bryant, P. G. (1991), "Large-Sample Results for Optimization-Based Clustering Methods," *Journal of Classification,* 8, 31–44.

Carty, M. P., El-Saleh, S., Zernik-Kobak, M., and Dixon, K. (1995), "Analysis of Mutations Induced by Replication of UV-Damaged Plasmid DNA in HeLa Cell Extracts," *Environmental and Molecular Mutagenesis,* 26, 139–146.

Celleux, G., and Diebolt, J. (1985), "The SEM Algorithm: A Probabilistic Teacher Algorithm Derived From the EM Algorithm for the Mixture Problem," *Computational Statistics Quarter,* 2, 73–82.

Celleux, G., and Govaert, G. (1991), "Clustering Criteria for Discrete Data and Latent Class Models," *Journal of Classification,* 8, 157–176.

Celleux, G., and Govaert, G. (1992), "A Classification EM Algorithm for Clustering and Two Stochastic Versions," *Computational Statistics and Data Analysis,* 14, 315–332.

Cohen, J. (1969), *Statistical Power Analysis for the Behavioral Sciences,* New York: Academic.

Courtemanche, C., and Anderson, A. (1994), "Shuttle-Vector Mutagenesis of Aflatoxin B1 in Human Cells: Effects of Sequence Context on the supF Mutational Spectrum," *Mutation Research,* 306, 143–151.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Series B,* 39, 1–38.

Endo, H., Schut, H. A. J., and Snydewine, E. G. (1994), "Mutagenic Specificity of 2-Amino-3-Methylimidazo[4,5-f]quinoline and 2-Amino-1-Methyl-6-Phenylimidazo[4,5-b]Pyridine in the supF Shuttle Vector System," *Cancer Research,* 54, 3745–3751.

Everitt, B. S. (1993), *Cluster Analysis,* London: Edward Arnold.

Feng, Z. D., and McCulloch, C. E. (1996), "Using Bootstrap Likelihood Ratios in Finite Mixture Models," *Journal of the Royal Statistical Society, Series B,* 58, 609–617.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice,* London: Chapman and Hall.

Hauser, J., Seidman, M. M., Sidur, K., and Dixon, K. (1986), "Sequence Specificity of Point Mutations Induced During Passage of a UV-Irradiated Shuttle Vector Plasmid in Monkey Cells," *Molecular and Cellular Biology,* 6, 277–285.

Jennison, C. (1993), "Discussion on the Meeting on the Gibbs Sampler and Other Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B,* 55, 54–56.

Keyse, S. M., Amaudruz, F., and Tyrrell, R. M. (1988), "Determination of the Spectrum of Mutations Induced by Defined-Wavelength Solar UVB (313-nm) Radiation in Mammalian Cells by Use of Shuttle Vector," *Molecular and Cellular Biology,* 8, 5425–5431.

Levy, D. D., Groopman, J. D., Lim, S. E., Seidman, M. M., and Kraemer, K. H. (1992), "Sequence Specificity of Aflatoxin B1-Induced Mutations in a Plasmid Replicated in Xeroderma Pigmentosum and DNA Repair Proficient Human Cells," *Cancer Research,* 52, 5668–5673.

Maccubbin, A. E., Mudipalli, A., Nadadur, A. A., Ersing, N., and Gurtoo, H. L. (1997), "Mutations Induced in a Shuttle Vector Plasmid Exposed to Monofunctionally Activated Mitomycin C," *Environmental and Molecular Mutagenesis,* 29, 143–151.

Mah, C. M. M., Maher, V. M., Thomas, H., Reid, T. M., King, C. M., and McCormick, J. J. (1989), "Mutation Induced by Aminofluorene-DNA Adducts During Replication in Human Cells," *Carcinogenesis,* 10, 2321–2328.

McLachlan, J. G., and Basford, E. K. (1987), *Mixture Models: Inference and Applications to Clustering,* New York: Marcel Dekker.

Page, J. E., Pataki, J., Harvey, R. G., and Dipple, A. (1996a), "Mutational Specificity of the Syn 1,2-Dihydrodiol 3,4-Epoxide of 5-Methylchrysene," *Cancer Letters,* 110, 249–252.

Page, J. E., Ross, H. L., Bigger, C. A. H., and Dipple, A. (1996b), "Mutagenic Specificities and Adduct Distributions for 7-Bromomehylbenz[a]anthracenes," *Carcinogenesis,* 17, 283–288.

Page, J. E., Szeliga, J., Amin, S., Hecht, S. S., and Dipple, A. (1995), "Mutational Spectra for 5,6-Dimethylchrysene 1,2-Dihydrodiol3,4-Epoxides in the supF Gene of pSP189," *Chemical Research in Toxicology,* 8, 143–147.

Parris, N. C., Levy, D. D., Jessee, J., and Seidman, M. M. (1994), "Proximal and Distal Effects of Sequence Context on Ultraviolet Mutational Hotspots in a Shuttle Vector Replicated in Xeroderma Cells," *Journal of Molecular Biology,* 236, 491–502.

Roilides, E., Gielen, J. E., Tuteja, N., Levine, A. S., and Dixon, K. (1988), "Mutational Specificity of Benzo[a]pyrene Diolepoxide in Monkey Cells," *Mutation Research,* 198, 199–206.

Seetharam, S., Protic-Sabljic, M., Seidman, M. M., and Kraemer, K. H. (1987), "Abnormal Ultraviolet Mutagenic Spectrum in Plasmid DNA Replicated in Cultured Fibroblasts From a Patient With the Skin Cancer-Prone Disease, Xeroderma Pigmentosum," *The Journal of Clinical Investigation,* 80, 1613–1617.

Seidman, M. M., Dixon, K., Razzaque, A., Zagursky, J. R., and Berman, L. M. (1985), "A Shuttle Vector Plasmid for Studying Carcinogen-Induced Point Mutations in Mammalian Cells," *Gene,* 38, 233–237.

Symons, M. J. (1981), "Clustering Criteria and Multivariate Normal Mixture," *Biometrics,* 37, 35–43.

Szeliga, J., Lee, H., Harvey, G. R., Page, J. E., Ross, H. L., Routledge, M. N., Hilton, B. D., and Dipple, A. (1994), "Reaction With DNA and Mutagenic Specificity of Syn-Benzo[g]chrysene 11,12-Dihydrodiol13,14-Epoxide," *Chemical Research in Toxicology,* 7, 420–427.

Szeliga, J., Page, J. E., Hilton, B. D., Kiselyov, A. S., Harvey, R. G., Dunayevskiy, Y. M., Vouros, P., and Dipple, A. (1995), "Characterization of DNA Adducts Formed by Anti-Benzo[g]chrysene 11,12-Dihydrodiol 13,14-Epoxide," *Chemical Research in Toxicology,* 8, 1014–1019.

Van Laarhoven, P. J. M., and Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Applications,* Dordrecht, The Netherlands: Reidel.

Yagi, T., Sato, M., Nishigori, C., and Takebe, H. (1994), "Similarity in the Molecular Profile of Mutations Induced by UV Light in Shuttle Vector Plasmids Propagated in Mouse and Human Cells," *Mutagenesis,* 9, 73–77.

Yagi, T., Tatsumi-Miyajima, J., Sato, M., Kraemer, K. H., and Takebe, H. (1991), "Analysis of Point Mutations in an Ultraviolet-Irradiated Shuttle Vector Plasmid Propagated in Cells From Japanese Xeroderma-Pigmentosum Patients in Complementation Groups A and F," *Cancer Research,* 51, 3177–3182.

Yang, J.-L., Maher, V. M., and McCormick, J. J. (1987), "Kinds of Mutations Formed When a Shuttle Vector Containing Adducts of (±)-7β,8α-Dihdrxy-9α-Epoxy-7,8,9,10-Tetrahydro Benzo[a]pyrene Replicates in Human Cells," *Proceedings of the National Academy of Sciences, USA,* 84, 3787–3791.

Yang, J.-L., Maher, V. M., and McCormick, J. J. (1988), "Kinds and Spectrum of Mutations Induced by 1-Nitrosopyrene Adducts During Plasmid Replication in Human Cells," *Molecular and Cellular Biology,* 8, 3364–3372.