Modelling the Mutation Process in the HIV Genome

By Hildete Prisco Pinheiro *
State University of Campinas - Brazil
Françoise Seillier-Moiseiwitsch
University of North Carolina at Chapel Hill
Pranab Kumar Sen
University of North Carolina at Chapel Hill

Abstract

In modelling the mutational process in DNA sequences of the human immunodeficiency virus (HIV) one cannot assume independence among positions along the sequences. Sites are analyzed, at the nucleotide or amino-acid level, by comparing each sequence to the consensus sequence. The state at each site is regarded as a binary event (i.e., there is a mutation or not). Two families of models are considered. Each sequence can be thought of as a degenerate lattice, then autologistic models are applicable. The probability of mutation at a specific site, given all others, has an exponential form with, as predictor, a function of neighboring sites. Since there is no closed form for the likelihood function, a Markov-chain Monte-Carlo procedure is called upon. The second model is based on the Bahadur representation for the joint distribution of dichotomous responses and assumes only pairwise dependence among sites. Parameter estimation is performed via the maximum likelihood method.

*This research was funded in part by CAPES (Brazilian Institution), the National Science Foundation, the American Foundation for AIDS Research and the National Institutes of Health.

Key words and phrases: Autologistic Models, Bahadur representation, binary data, categorical data, MCMC methods

1 Introduction

Our primary interest is in modelling whether there is a mutation or not at each site. We compare each sequence to the consensus sequence and look for differences. The response is thus binary and two alternative models are formulated: an *autologistic* model (Section 2), and a Bahadur representation for the joint distribution of Bernoulli trials (Section 3). We assume only pairwise dependence among sites. Parameter estimation is discussed in each case and a data analysis is developed in Section .

2 The Autologistic Model

The autologistic model was introduced by Besag (1972, 1974, 1975) and is widely suited to spatial binary data (Cressie, 1993). This parametric family can handle situations involving both spatial correlation and dependence on covariates. Applications include modelling the distribution of plant species in terms of climate variables like temperature and rainfall (Huffer & Wu, 1995a), the effect of soil variables on disease incidence in plants (Gumpertz & Graham, 1995) and network autocorrelation data (Smith, Calloway & Morrisey, 1995). It is important to point out that in these papers the whole data set consists of a single lattice, but in our case we have a number of independent lattices (i.e., sequences).

Let

$$y(i) = \begin{cases} 0 \text{ if there is no mutation at site } i \\ 1 \text{ if there is a mutation at site } i \end{cases}$$

Construct $1 \times n$ vectors, where n is the total number of sites along a sequence:

$$\mathbf{y} = (y(1) \ \ldots \ y(n)) \quad \text{and} \quad \mathbf{0} = (0\ 0\ \ldots\ 0).$$

The model formulation requires that we define the concept of *neighborhood* and introduce the *positivity condition*.

Definition 1
A site j is a *neighbor* of site i if the conditional distribution of $Y(i)$, given all other site values, depends functionally on $y(j)$, for $j \neq i$. Also define

$$N_i = \{j : j \textit{ is a neighbor of } i\} \tag{2.1}$$

to be the *neighborhood* of site i. ■

Positivity Condition
Let Y be a discrete variable associated with n sites. Define $\zeta = \{\mathbf{y} : \Pr(\mathbf{y}) > 0\}$ and $\zeta_i = \{y(i) : \Pr(y(i)) > 0\},\ i = 1, \ldots, n$. Then the *positivity condition* is satisfied if $\zeta = \zeta_1 \times \ldots \times \zeta_n$. For a continuous variable, the same definition applies except that $\Pr(\cdot)$ is replaced by $f(\cdot)$. ■

This condition states that the support of the joint distribution is the cartesian product of the supports for the marginal distributions. It implies that considering the elements $\{y(i) :\ i, \ldots, n\}$ jointly does not rule out combinations allowed in the set $\{y(1)\} \times \{y(2)\} \times \ldots \times \{y(n)\}$. For instance, this condition is invalidated for an infectious disease model. Because of the one-way nature of infection, the following situation is not allowed: $y(i) = 1$ when $\{y(j) = 0,\ j \in N_i\}$, i.e., $y(i)$ is affected when all the neighbors of i are not affected.

Without loss of generality, assume that 0 can occur at each site. Let

$$Q(\mathbf{y}) = \log\{\Pr(\mathbf{y})/\Pr(\mathbf{0})\}, \qquad \mathbf{y} \in \zeta, \tag{2.2}$$

where ζ is the support of the distribution of $\mathbf{Y}$. Then the knowledge of $Q(.)$ is equivalent to the knowledge of $\Pr(.)$, because

$$\Pr(\mathbf{y}) = \exp(Q(\mathbf{y}))/\sum_{\mathbf{y}\in\zeta} \exp(Q(\mathbf{y}))$$

in the case of discrete y's. A similar formula, with integrals replacing summations, applies for continuous y's.

Proposition 1 (Cressie, 1993)
The function Q satisfies the following two properties:

i.

$$\frac{\Pr(y(i) \mid \{y(j) : j \neq i\})}{\Pr(0(i) \mid \{y(j) : j \neq i\})} = \frac{\Pr(\mathbf{y})}{\Pr(\mathbf{y}_i)} = \exp(Q(\mathbf{y}) - Q(\mathbf{y}_i)) \tag{2.3}$$

where $0(i)$ denotes the event $Y(i) = 0$ and $\mathbf{y}_i \equiv (y(1), \ldots, y(i-1), 0, y(i+1), \ldots, y(n))$

ii. Q can be expanded uniquely on ζ as

$$\begin{aligned} Q(\mathbf{y}) \quad = \quad & \sum_{1\le i\le n} y(i)\, \mathrm{G}_i(y(i)) + \sum_{1\le i<j\le n} y(i)\, y(j)\, \mathrm{G}_{ij}(y(i),\, y(j)) \\ & + \sum_{1\le i<j<k\le n} y(i)\, y(j)\, y(k)\, \mathrm{G}_{ijk}\,(y(i),\, y(j),\, y(k)) \ + \ \ldots \\ & + y(1) \ \ldots \ y(n)\, \mathrm{G}_{1\ldots n}(y(1),\, \ldots,\, y(n)), \qquad \mathbf{y} \in \zeta. \end{aligned} \tag{2.4}$$

■

Note that although the expansion (2.4) is unique, the function $\{G_{ij\ldots}\}$ are not uniquely specified. By defining $G_{ij\ldots}(y(i),\, y(j),\, \ldots) \equiv 0$ whenever one of the arguments is 0, uniqueness is obtained.

Example (Liang, Zeger & Qaqish, 1992)

The representation of $Q(.)$ in (2.4) has been used to represent the probability distribution for a vector $\mathbf{x}$ of binary responses in a saturated log-linear model:

$$\Pr(\mathbf{x}) = \exp\{u_0 + \sum_{j=1}^{n} u_j\, x_j + \sum_{j<k} u_{jk}\, x_j\, x_k + \ldots + u_{12\ldots n}\, x_1\, \ldots\, x_n\} \tag{2.5}$$

where there are $2^n - 1$ parameters $u = (u_1,\, \ldots,\, u_n, u_{11},\, u_{12},\, \ldots,\, u_{n-1,n},\, \ldots,\, u_{12\ldots n})'$. These have straightforward interpretations in terms of conditional probabilities. For example,

$$\begin{aligned} u_j &= \text{logit}\{\Pr(x_j = 1 \mid x_k = 0,\, k \neq j)\}, \quad j = 1, \ldots, n, \\ u_{jk} &= \log \text{OR}(x_j,\, x_k \mid x_l = 0,\, l \neq j,\, k), \quad j < k = 1, \ldots, n, \end{aligned}$$

and

$$u_{123} = \log \text{OR}(x_1,\, x_2 \mid x_3 = 1,\, x_l = 0,\, l > 3) - \log \text{OR}(x_1,\, x_2 \mid x_3 = 0,\, x_l = 0,\, l > 3) \tag{2.6}$$

where

$$\text{OR}(v,\, w) = \frac{\Pr(v = w = 1)\Pr(v = w = 0)}{\Pr(v = 1,\, w = 0)\Pr(v = 0,\, w = 1)}$$

The implication of properties (i) and (ii) is that the expansion (2.4) for $Q(\mathbf{y})$ is actually made up of conditional probabilities. For instance,

$$y(i)\, G_i(y(i)) = \log\left[\frac{\Pr(y(i) \mid \{0(j) : j \neq i\})}{\Pr(0(i) \mid \{0(j) : j \neq i\})}\right]$$

$$\begin{aligned} &y(i)\, y(j)\, G_{ij}(y(i),\, y(j)) \\ &\quad = \log\left\{\frac{\Pr(y(i) \mid y(j),\, \{0(l) :\, l \neq i, j\})}{\Pr(0(i) \mid y(j),\, \{0(l) :\, l \neq i, j\}} \times \frac{\Pr(0(i) \mid \{0(l) :\, l \neq i\})}{\Pr(y(i) \mid \{0(l) :\, l \neq i\}}\right\} \end{aligned}$$

■

From (2.2), the joint probability distribution of $\mathbf{y}$, $\Pr(\mathbf{y})$, is proportional to $\exp(Q(\mathbf{y}))$. Finding the proportionality constant as a function of the parameters enables us to write down the full likelihood and to obtain the maximum likelihood estimates of the parameters. Unfortunately, this is not always possible. Further, there is a powerful theorem regarding the form the function Q must take so that the conditional expressions combine consistently into a proper joint distribution. We must first define a *clique*.

Definition 2
A *clique* is a set of sites that consists either of a single site or of sites that are all neighbors of each other. ■

Theorem 1 (Hammersley-Clifford, 1971)
Suppose that $\mathbf{Y}$ is distributed according to a Markov random field on ζ that satisfies the positivity condition. Then, the negpotential function $Q(\cdot)$ given by (2.4) must satisfy the property that if sites $i, j, \ldots, n$ do not form a clique, then $G_{ij\ldots n}(\cdot) = 0$, where cliques are defined by the neighborhood structure $\{N_i : \ i = 1, \ldots, n\}$. ■

Assuming only pairwise dependence between sites,

$$Q(\mathbf{y}) = \sum_{1 \le i \le n} y(i)\, \mathrm{G}_i(y(i)) + \sum_{1 \le i < j \le n} y(i)\, y(j)\, \mathrm{G}_{ij}(y(i),\, y(j)) \tag{2.7}$$

Since $y(i)$ is either 1 or 0, the only values for the functions G needed in (2.7) are $\mathrm{G}_i(1) \equiv \alpha_i$ and $\mathrm{G}_{ij}(1,1) \equiv \gamma_{ij}$. Thus,

$$Q(\mathbf{y}) = \sum_{i=1}^{n} \alpha_i\, y(i) + \sum_{1 \le i < j \le n} \gamma_{ij}\, y(i)\, y(j) = \log\{\Pr(\mathbf{y})/\Pr(\mathbf{0})\} \tag{2.8}$$

where $\gamma_{ij} = 0$ unless sites i and j are neighbors. Then,

$$Q(\mathbf{y}) - Q(\mathbf{y}_i) = \alpha_i\, y(i) + \sum_{j=1}^{n} \gamma_{ij}\, y(i)\, y(j)$$

where $\gamma_{ij} = \gamma_{ji}$ and, to maintain identifiability of the parameters, $\gamma_{ii} = 0$. Therefore, from (2.3),

$$\frac{\Pr(y(i) \mid \{y(j) : j \neq i\})}{\Pr(0(i) \mid \{y(j) : j \neq i\})} = \exp\{\alpha_i\, y(i) \ + \ \sum_{j=1}^{n} \gamma_{ij}\, y(i) y(j)\}$$

Because $y(i) = 0$ or 1,

$$\Pr(0(i) \mid \{y(j) : j \neq i\}) = \frac{1}{1 + \exp(\alpha_i + \sum_{j=1}^{n} \gamma_{ij}\, y(j))}$$

and

$$P(y(i) \mid \{y(j) : j \neq i\}) = \frac{\exp(\alpha_i\, y(i) + \sum_{j=1}^{n} \gamma_{ij}\, y(i)\, y(j))}{1 + \exp(\alpha_i + \sum_{j=1}^{n} \gamma_{ij}\, y(j))} \tag{2.9}$$

Even with just pairwise dependence this formulation may involve too many parameters for data sets of moderate size and we need to reduce the number of parameters by imposing some constraints. For instance,

- For the α's,
  1. α_i's all different;
  2. $\alpha_i = \alpha,\ \forall i$;
  3. a different α_i for each specific amino acid (or nucleotide) in the consensus sequence.
- For the γ's,
  1. γ's all different;
  2. $\gamma_{ij} = 0$ if $\mid i - j \mid > d$, where d is some chosen threshold;
  3. $\gamma_{ij} = \gamma \rho^{|i-j|}$;
  4. equal γ_{ij} for each pair of specific amino acids (or nucleotide) in the consensus sequence;
  5. $\gamma_{ij} = \gamma,\ \forall\, i,\, j$.

2.1 Estimation Procedure

Suppose the data consist of m independent sequences with n sites. In fact, we have $y_k(i)$, where $i = 1, \ldots, n$ and $k = 1, \ldots, m$. Then, $\mathbf{y}(i)$ is a $m \times 1$ vector.

We can get approximate estimates by using the *pseudolikelihood* function L_P (Besag, 1975), i.e. the product of the conditional probabilities, which specifies the model assuming independence among sites.

$$L_P(\boldsymbol{\theta}; \mathbf{y}(1), \ldots, \mathbf{y}(n)) = \prod_{i=1}^{n} \Pr(\mathbf{y}(i) \mid \{\mathbf{y}(j) : j \neq i\}; \boldsymbol{\theta})$$

where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_n, \gamma_{1\,2}, \ldots, \gamma_{n-1\,n})'$ is the parameter vector. Since the sequences are independent, we can write

$$\begin{aligned} L_P(\boldsymbol{\theta}; \mathbf{y}(1), \ldots, \mathbf{y}(n)) &= \prod_{k=1}^{m} \prod_{i=1}^{n} \Pr(y_k(i) \mid \{y_k(j) : j \neq i\}) \\ &= \prod_{k=1}^{m} \prod_{i=1}^{n} \left\{ \frac{\exp[\alpha_i\, y_k(i) + \sum_{j=1}^{n} \gamma_{ik}\, y_k(i)\, y_k(j)]}{1 + \exp[\alpha_i + \sum_{j=1}^{n} \gamma_{ij}\, y_k(j)]} \right\} \end{aligned} \tag{2.10}$$

The maximum pseudolikelihood estimate (MPLE) is the value of $\boldsymbol{\theta}$ which maximizes L_P, i.e., (2.10). The MPLE is consistent and asymptotically normal (Comets, 1992; Comets & Gidas, 1992). It has some disadvantages: its efficiency is unknown and expected to be inferior to that of the maximum likelihood estimate (MLE). Also, there is no direct way of obtaining standard errors for the estimates.

In order to calculate the MLE we need to proceed by Markov-chain Monte-Carlo simulation as proposed by Geyer & Thompson (1992). We can write the joint distribution of $\mathbf{y} = (\mathbf{y}(1), \ldots, \mathbf{y}(n))$ as

$$\Pr\{\mathbf{y}(1), \ldots, \mathbf{y}(n); \boldsymbol{\theta}\} = C(\boldsymbol{\theta}) F(\mathbf{y}(1), \ldots, \mathbf{y}(n); \boldsymbol{\theta}).$$

where F is an explicitly computable function and $C(\boldsymbol{\theta}) = \Pr\{\mathbf{0}(1), \ldots, \mathbf{0}(n); \boldsymbol{\theta}\}$. For the autologistic model,

$$F = \exp\left\{ \sum_{k=1}^{m} \sum_{i=1}^{n} \alpha_i \, y_k(i) \; + \sum_{k=1}^{m} \sum_{1 \le i < j \le n} \gamma_{ij} \, y_k(i) \, y_k(j) \right\} \tag{2.11}$$

Fix some specific value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$ say. Suppose we can generate a random vector $(\mathbf{Y}'(1), \ldots, \mathbf{Y}'(n))$ from the distribution with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The random variable

$$\frac{F(\mathbf{Y}'(1), \ldots, \mathbf{Y}'(n); \boldsymbol{\theta})}{F(\mathbf{Y}'(1), \ldots, \mathbf{Y}'(n); \boldsymbol{\theta}_0)}$$

has mean

$$\begin{aligned} \sum_{\mathbf{y}'(1), \ldots, \mathbf{y}'(n)} & \frac{F(\mathbf{y}'(1), \ldots, \mathbf{y}'(n); \boldsymbol{\theta})}{F(\mathbf{y}'(1), \ldots, \mathbf{y}'(n); \boldsymbol{\theta}_0)} \, C(\boldsymbol{\theta}_0) \, F(\mathbf{y}'(1), \ldots, \mathbf{y}'(n); \boldsymbol{\theta}_0) \\ &= C(\boldsymbol{\theta}_0) \sum_{\mathbf{y}'(1), \ldots, \mathbf{y}'(n)} F(\mathbf{y}'(1), \ldots, \mathbf{y}'(n); \boldsymbol{\theta}) \\ &= \frac{C(\boldsymbol{\theta}_0)}{C(\boldsymbol{\theta})}, \qquad \text{since} \sum_{\mathbf{y}'(1), \ldots, \mathbf{y}'(n)} C(\boldsymbol{\theta}) F(\mathbf{y}'(1), \ldots, \mathbf{y}'(n); \boldsymbol{\theta}) = 1. \end{aligned}$$

Then, suppose we have a (not necessarily independent) set of random vectors $(\mathbf{Y}^{(r)}(1), \ldots, \mathbf{Y}^{(r)}(n))$, $r = 1, 2, \ldots, R$, each drawn from the distribution $C(\boldsymbol{\theta}_0) F(.\,; \boldsymbol{\theta}_0)$ for some fixed $\boldsymbol{\theta}_0$. Denote the observations by $(\mathbf{Y}(1), \ldots, \mathbf{Y}(n))$. The function

$$H(\boldsymbol{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \frac{F(\mathbf{Y}^{(r)}(1), \ldots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta})}{F(\mathbf{Y}^{(r)}(1), \ldots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta}_0)} \times \frac{F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta})} \tag{2.12}$$

is an unbiased estimator of the likelihood ratio of $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$,

$$\frac{C(\boldsymbol{\theta}_0) \, F(\mathbf{Y}(1) \ldots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}) \, F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta})}$$

since

$$\begin{aligned} E[H(\boldsymbol{\theta})] &= \frac{1}{R} \frac{F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta})} \sum_{r=1}^{R} E\left[\frac{F(\mathbf{Y}^{(r)}(1), \ldots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta})}{F(\mathbf{Y}^{(r)}(1), \ldots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta}_0)} \right] \\ &= \frac{C(\boldsymbol{\theta}_0) F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}) F(\mathbf{Y}(1), \ldots, \mathbf{Y}(n); \boldsymbol{\theta})} \end{aligned}$$

To obtain the MLE of $\boldsymbol{\theta}$, use equation (2.12), based on a single simulated sequence, and minimize it with respect to $\boldsymbol{\theta}$ (Geyer & Thompson, 1992).

Here is a summary of the procedure:

1. Choose a model, as in equation (2.9).
2. Use the MPLE to generate an initial point estimate ($\boldsymbol{\theta}_0$) for the unknown parameter vector.
3. Use Gibbs sampling or the Metropolis algorithm to generate a simulated sequence of random vectors from the distribution with parameter vector $\boldsymbol{\theta}_0$. Discard an initial warm-up sample, then generate random vectors $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$ .... Select R vectors (R should be at least 1000) by sampling every k (50, say) because of the dependence between successive vectors.
4. Use equation (2.12) to define $H(\boldsymbol{\theta})$, the simulated likelihood ratio of $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$.
5. Using numerical optimization, maximize the function $-\log H(\boldsymbol{\theta})$ to obtain the MLE $\hat{\boldsymbol{\theta}}$, and use the equivalent form of the observed information matrix to evaluate standard errors for these estimates.

Alternatively in (3), instead of sampling every k, R samplers can be initiated at different values, discarding an initial warm-up at each. ■

The Gibbs Sampler algorithm (Geman & Geman, 1984; Gelfand & Smith, 1990) and the Metropolis algorithm (Metropolis et al., 1953) are special cases of the Hastings algorithm (Hastings, 1970; Peskun, 1973). An explanation of how these algorithms work and of the differences between them follows.

Hastings' Algorithm

Suppose we want to generate a discrete or continuous random variable (scalar or vector) Y which takes values in a space τ with density $\pi(t)$. We assume that $\pi(.)$ is completely specified up to a normalizing constant. We also select a Markov transition kernel $q(s,t)$, s, $t \in \tau$, which is used to generate trial values of Y. In the discrete case, the interpretation is that if the current values of Y is s, the next trial value is t with probability $q(s,t)$. The choice of $q(.,.)$ is almost arbitrary and the performance of the algorithm may vary according to this choice. The algorithm follows.

Step 0: Take an arbitrary starting value $Y = t_0$ and set $i = 0$.

Step 1: Given $Y = t_i = s$, choose a new trial value according to the probability distribution $q(s,t)$, s, $t \in \tau$, where $\tau = \{0, 1\}$ and

$$q(s,t) = \Pr(Y_1 = t \mid Y_0 = t_0 = s)$$

is an arbitrary Markov kernel (generating an irreducible and aperiodic chain).

Step 2: Calculate

$$\alpha(s,t) = \min\left\{\frac{\pi(t)\,q(t,s)}{\pi(s)\,q(s,t)},\, 1\right\},$$

where $\pi(t)$ is a distribution specified up to a normalizing constant. Here $\pi(.) = F(.;\boldsymbol{\theta}_0)$, where F is given by (2.11).

Step 3: Generate a random variable

$$U = \begin{cases} 1 \text{ with probability } \alpha(s,t) \\ 0 \text{ with probability } 1-\alpha(s,t) \end{cases}$$

Step 4: If $U = 1$, move to t, i.e., $t_{i+1} = t$. Otherwise $t_{i+1} = t_i$.

Step 5: Set $i := i + 1$. ■

Metropolis' Algorithm

q is taken to be a symmetric random walk, i.e., $q(s,t) = q(t,s)$. Then, in step 2:

$$\alpha(s,t) = \min\left\{\frac{\pi(t)}{\pi(s)},\, 1\right\}$$

■

The Gibbs Sampler Algorithm

For the Gibbs Sampler, each of the updating steps corresponds to generating a new value from a particular conditional distribution, i.e., we always accept the new value. Then, in step 2: $\alpha(s,t) = 1$. In our specific case, we are generating m independent sequences of length n and the Gibbs sampler proceeds as follows:

Step 0: Take arbitrary starting values, say $(y_1, y_2, \ldots, y_n)$. Define $y_j^{(0)} = y_j,\ j = 1, \ldots, n$. Let $i = 0$.

Step 1: Given the current values of $y_1^{(i)}$, $y_2^{(i)}$, $\ldots$, $y_n^{(i)}$, generate a new pseudo-random value from the conditional distribution of Y_1 given $Y_j = y_j^{(i)},\ j = 2, \ldots, n$, and call it $y_1^{(i+1)}$. This probability distribution is given in equation (2.9), i.e.,

$$P(y(i) \mid \{y(j) : j \neq i\}) = \frac{\exp(\alpha_i\, y(i) + \sum_{j=1}^{n} \gamma_{ij}\, y(i)\, y(j))}{1 + \exp(\alpha_i + \sum_{j=1}^{n} \gamma_{ij}\, y(j))}$$

Step 2: Given the current values of $y_1^{(i+1)}$, $y_3^{(i)}$, $\ldots$, $y_n^{(i)}$, generate a pseudo-random value from the conditional distribution of Y_2 given $Y_1 = y_1^{(i+1)}$, $Y_j^{(i)},\ j = 3, \ldots, n$, and call it $y_2^{(i+1)}$.

Continue updating one component at a time until...

Step n: Given the current values of $y_1^{(i+1)}$, $y_2^{(i+1)}$, ..., $y_{n-1}^{(i+1)}$, generate a pseudo-random value from the conditional distribution of Y_n given $Y_j = y_j^{(i+1)}$, $j = 1, \ldots, n-1$, and call it $y_n^{(i+1)}$.

Step n+1: Set $i := i + 1$ and return to Step 1.

A single cycle through steps 1 to n+1 completes one iteration of the algorithm. We will discard an initial warm-up sample, then generate R random vectors by sampling the chain every k cycles. ■

3 Model based on the Bahadur Representation

Let $\mathbf{y}_k = (y_k(1) \ \ldots \ y_k(n))$ be the $1 \times n$ vector representing the binary responses for the n sites along sequence k, i.e., whether there is a mutation or not at each site along the sequence. Since each $y_k(i)$ $(i = 1, \ldots, n)$ assumes the value 1 or 0, the random variable $Y_k(i)$ is a Bernoulli trial with parameter $\xi_i = \Pr\{Y_k(i) = 1\}$. Then, $\mathrm{E}[Y_k(i)] = \xi_i$ and $\mathrm{Var}[Y_k(i)] = \xi_i(1 - \xi_i)$.

Let

$$Z_k(i) = \frac{Y_k(i) - \xi_i}{\sqrt{\xi_i(1 - \xi_i)}}$$

and

$$r_{ij} = \mathrm{E}[Z_k(i)\, Z_k(j)],\ r_{ijl} = \mathrm{E}[Z_k(i)\, Z_k(j)\, Z_k(l)],\ \ldots, r_{1\,2\,\ldots\,n} = \mathrm{E}[Z_k(1)\, Z_k(2)\, \ldots\, Z_k(n)]$$

So, r_{ij} are second-order correlations, r_{ijl} third-order correlations, and so on. Hence, there are $\binom{n}{2} + \binom{n}{3} + \ldots + \binom{n}{n} = 2^n - n - 1$ correlation parameters.

Denote by $\mathbf{P}_{[1]k}$ the joint distribution of $y_k(i)$'s when they are independent, i.e.,

$$\mathbf{P}_{[1]k}(y_k(1), \ldots, y_k(n)) = \prod_{i=1}^{n} \xi_i^{y_k(i)}\, (1 - \xi_i)^{1 - y_k(i)} \tag{3.1}$$

Let $\mathbf{P}(\mathbf{y}_k)$ be the distribution of $\mathbf{Y}_k$, where $\mathbf{Y}_k = (Y_k(1) \ \ldots \ Y_k(n))$ is a $1 \times n$ vector denoting the response random vector for the k-th sequence.

Proposition (Bahadur, 1961)
For every $\mathbf{y}_k = (y_k(1), \ldots, y_k(n))$,

$$\mathbf{P}(\mathbf{y}_k) = \mathbf{P}_{[1]k}(\mathbf{y}_k)\, f(\mathbf{y}_k), \tag{3.2}$$

where

$$f(\mathbf{y}_k) = 1 + \sum_{i<j} r_{ij}\, z_k(i)\, z_k(j) + \ldots + r_{1\,2\,\ldots\,n}\, z_k(1)\, z_k(2)\, \ldots\, z_k(n) \tag{3.3}$$

■

If we assume pairwise dependence only, the distribution of $\mathbf{Y}_k$ is

$$\begin{aligned}
\mathbf{P}(\mathbf{y}_k) &= \mathbf{P}_{[1]k}(\mathbf{y}_k)\,[1+\sum_{i<j} r_{ij}\,z_k(i)\,z_k(j)] \\
&= \prod_{i=1}^{n} \xi_i^{y_k(i)}\,(1-\xi_i)^{1-y_k(i)}\,[1+\sum_{i<j} r_{ij}\,z_k(i)\,z_k(j)] \\
&= \prod_{i=1}^{n} \xi_i^{y_k(i)}\,(1-\xi_i)^{1-y_k(i)} \\
&\quad \times \left[1+\sum_{i<j} r_{ij}\,\left(\frac{y_k(i)-\xi_i}{\sqrt{\xi_i(1-\xi_i)}}\right)\left(\frac{y_k(j)-\xi_j}{\sqrt{\xi_j(1-\xi_j)}}\right)\right] \qquad (3.4)
\end{aligned}$$

3.1 Estimation Procedure

Let $\boldsymbol{\theta} = (\xi_1,\ldots,\xi_n, r_{12},\ldots,r_{1n}, r_{23},\ldots,r_{2n},\ldots,r_{n-1\,n})'$. An estimate of $\boldsymbol{\theta}$ can be obtained by maximum likelihood. The likelihood function for sequence k is given by (3.4). For m independent sequences, the likelihood function is

$$\begin{aligned}
&\mathbf{L}(\boldsymbol{\theta};\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_m) \\
&\qquad = \prod_{k=1}^{m}\prod_{i=1}^{n} \xi_i^{y_k(i)}\,(1-\xi_i)^{1-y_k(i)} \\
&\qquad \times \left[1+\sum_{i<j} r_{ij}\,\left(\frac{y_k(i)-\xi_i}{\sqrt{\xi_i(1-\xi_i)}}\right)\left(\frac{y_k(j)-\xi_j}{\sqrt{\xi_j(1-\xi_j)}}\right)\right] \qquad (3.5)
\end{aligned}$$

4 Data Analysis

As an example, we consider a set of HIV-1 sequences from LaRosa et al. (1990) and Myers et al. (1993) which span the V3 loop of the envelope gene. The data consist of 87 sequences with 35 amino acids each. They are all from the B clade (North America, Western Europe, Brazil and Thailand). Within the V3 loop are found determinants for T-cell-adaptation and macrophage tropism (the wild-type phenotype). These strains (T-cell adapted and macrophage tropic) are subject to different selection pressures and therefore linkage patterns within the V3 loop are most likely to differ as well.

4.1 The Autologistic Model

Recall from Section 2 that $y(i) = 0$ or 1,

$$\Pr(0(i) \mid \{y(j) : j \in N_i\}) = \frac{1}{1 + \exp\left(\alpha_i + \sum_{j \in N_i} \gamma_{ij}\, y(j)\right)}$$

and

$$P(y(i) \mid \{y(j) : j \in N_i\}) = \frac{\exp\left(\alpha_i\, y(i) + \sum_{j \in N_i} \gamma_{ij}\, y(i)\, y(j)\right)}{1 + \exp\left(\alpha_i + \sum_{j \in N_i} \gamma_{ij}\, y(j)\right)} \tag{4.1}$$

where $N_i = \{j : j \text{ is a neighbor of } i\}$ is the neighborhood of site i.

Since the data set is of moderate size and this model involves too many parameters, we reduce the number of parameters by imposing some restrictions.

Model 1

• The neighborhood is defined as the immediate positions, i.e, the neighborhood of site i is $i-1$ and $i+1$. At the two extremities of the sequences, the neighborhood is only the next one for the left extremity and only the previous one for the right extremity, i.e., the neighborhood of site 1 is site 2 and the neighborhood of site n is $n-1$.

• $\alpha_i = \alpha,\ \forall i$
• $\gamma_{ij} = \gamma,\ \forall\, i, j.$

The model is then,

$$P(y(i) \mid \{y(j) : 0 < \mid i - j \mid \leq 1\}) = \frac{\exp\left(\alpha\, y(i) + \sum_{j \in N_i} \gamma\, y(i)\, y(j)\right)}{1 + \exp\left(\alpha + \sum_{j \in N_i} \gamma\, y(j)\right)} \tag{4.2}$$

The maximum pseudo-likelihood estimates (MPLE) and the simulated maximum likelihood estimates (SMLE) are shown in Table 2. We used the Metropolis algorithm to obtain the MCMC maximum-likelihood estimates: 1000 samplers were generated with different initial values, discarding an initial warm-up of 500 iterations in each.

To verify if we reached convergence after a burn-in of 500, we considered the first 100 samplers and compared at each of the 33 positions the empirical probabilities with the probabilities under model (4.2).

For each position we computed the statistic Z

$$Z_i = \frac{\hat{p}_i - p_i}{\sqrt{p_i(1-p_i)/m}}, \quad i = 1, \ldots, 33$$

where,

$$p_i = P(y(i) \mid \{y(j) : 0 < \mid i - j \mid \leq 1\}) = \frac{\exp\left(\tilde{\alpha}\, y(i) + \sum_{j \in N_i} \tilde{\gamma}\, y(i)\, y(j)\right)}{1 + \exp\left(\tilde{\alpha} + \sum_{j \in N_i} \tilde{\gamma}\, y(j)\right)},$$

$\tilde{\alpha}$ and $\tilde{\gamma}$ are the MPLE of α and γ, respectively, and $\hat{p}_i$ is the empirical distribution computed from the first 100 samplers.

For each position we compute 4 probabilities:

$$P(y(i) = 0 \mid y(i-1) = 0, y(i+1) = 0) \quad , \quad P(y(i) = 0 \mid y(i-1) = 0, y(i+1) = 1)$$
$$P(y(i) = 0 \mid y(i-1) = 1, y(i+1) = 0) \text{ and } P(y(i) = 0 \mid y(i-1) = 1, y(i+1) = 1)$$

For the empirical distribution we have,

$$\begin{aligned}
\hat{p}_{i1} &= \frac{(\#y(i) = 0, y(i-1) = 0, y(i+1) = 0)}{(\#y(i-1) = 0, y(i+1) = 0)}, \\
\hat{p}_{i2} &= \frac{(\#y(i) = 0, y(i-1) = 0, y(i+1) = 1)}{(\#y(i) - 1 = 0, y(i+1) = 1)}, \\
\hat{p}_{i3} &= \frac{(\#y(i) = 0, y(i-1) = 1, y(i+1) = 0)}{(\#y(i-1) = 1, y(i+1) = 0)} \quad \text{and} \\
\hat{p}_{i1} &= \frac{(\#y(i) = 0, y(i-1) = 1, y(i+1) = 1)}{(\#y(i-1) = 1, y(i+1) = 1)}
\end{aligned}$$

We only looked at 33 positions, because the first and last positions remain constant. So, we then have 4×33 statistics to compute and the results in Table 1 show that only 6 out of 132 (4%) of those statistics have a significant result. Hence, if we adopt a level of significance of 5%, a burn-in of 500 is satisfactory.

For this model,

$$\log\left\{\frac{P(y(i) = 1 \mid y(i-1) = 0,\, y(i+1) = 0)}{1 - P(y(i) = 1 \mid y(i-1) = 0,\, y(i+1) = 0)}\right\} = \alpha$$

Table 1: Convergence Results

	Positions										
Statistic	2	3	4	5	6	7	8	9	10	11	12
Z_{i1}	1.41	0.06	0.88	0.30	1.89	-1.44	0.05	0.98	-0.42	-1.15	1.65
Z_{i2}	-0.36	-0.87	1.62	-1.29	0.82	-0.38	-1.65	1.25	0.59	-1.03	-0.27
Z_{i3}	0.65	-0.83	-0.62	-0.35	1.37	-1.19	0.03	-0.63	0.08	0.78	-0.82
Z_{i4}	-0.80	-1.47	-0.55	2.62*	0.54	-0.88	0.07	-0.50	1.34	-0.78	0.77
Statistic	13	14	15	16	17	18	19	20	21	22	23
Z_{i1}	0.65	-0.27	0.71	0.14	2.45*	-0.89	0.83	-0.83	1.06	-0.43	0.55
Z_{i2}	-0.89	0.44	-0.10	-1.59	1.03	-0.25	0.12	0.12	-0.32	0.09	-1.54
Z_{i3}	0.56	-1.77	0.61	-0.50	0.06	0.47	-1.13	0.83	0.47	-0.80	-0.33
Z_{i4}	-0.95	-0.67	0.04	0.26	0.54	-2.34*	0.96	-0.64	-0.40	-1.57	-0.40
Statistic	24	25	26	27	28	29	30	31	32	33	34
Z_{i1}	0.92	-2.01*	-0.55	1.01	-0.17	0.89	0.10	-1.55	-0.17	-2.26*	-0.69
Z_{i2}	1.52	-0.22	-0.57	1.32	0.09	-0.10	0.99	0.79	0.82	-1.72	0.88
Z_{i3}	-0.74	0.21	0.47	0.37	0.07	0.85	0.20	0.27	-1.17	1.27	0.02
Z_{i4}	-0.28	-0.76	-1.19	-0.24	-0.25	0.27	-1.02	-1.20	3.91*	1.22	-1.43

$\star \mid Z_i \mid > 1.96$.

Table 2: Model 1

	MPLE	SMLE
α	-1.7117	-1.7117
γ	0.7813	0.8204

Hence, α is the log-odds of mutation at a certain site when there is no mutation in its neighborhood. If α is negative, the probability of mutation at site i is less than the probability of no mutation at that site, given no mutation at its neighborhood. Since we are assuming that $\alpha_i = \alpha$, the probability of no mutation is 5.53 (exp(1.71)) times higher than the probability of mutation given that there is no mutation in the neighborhood. When there is mutation in the neighborhood of site i, the log odds of mutation is a function of both α and γ. So, when there is mutation in the neighborhood of site i, if γ is large and positive the log-odds of mutation at this site increases, and if γ is negative the log-odds of mutation at this site decreases.

Model 2

- The neighborhood is defined as in model 1.
- $\alpha_i = \alpha$, $\forall i$
- $\gamma_{ij} = \gamma \rho^{|i-j|}$.

The model is then,

$$P(y(i) \mid \{y(j) : 0 < \mid i-j \mid \leq 1\}) = \frac{\exp(\alpha\, y(i) + \sum_{j=1}^{n} \gamma \rho^{|i-j|}\, y(i)\, y(j))}{1 + \exp(\alpha + \sum_{j=1}^{n} \gamma \rho^{|i-j|}\, y(j))} \tag{4.3}$$

Table 3: Model 2

	MPLE	SMLE
α	-0.0582	-0.0582
γ	1.5715	1.5715
ρ	0.4679	0.4679

There is no difference between the pseudolikelihood estimates and the simulated maximum-likelihood estimates, indicating that the dependency among neighboring positions is weak. Again we are assuming that $\alpha_i = \alpha$ and the probability of no mutation is 1.05 (exp(0.06)) times higher than the probability of mutation given that there is no mutation in the neighborhood. When there is mutation in the neighborhood of site i, the log odds of mutation is a function of α, γ, ρ and the distance between sites i and j (since ρ has power $\mid i-j \mid$). So, when there is mutation in the neighborhood of site i, the log-odds of mutation at this site increases, since γ and ρ are positive.

4.2 Model based on the Bahadur Representation

From Section 2.3, $\mathbf{y}_k = (y_k(1), \ldots, y_k(n))$ is a $n \times 1$ vector representing the binary responses for the n sites along sequence k, i.e, whether there is a mutation from the consensus or not at each

site along the sequence. Here, we consider $k = 87$ sequences and $n = 35$ sites.

Probabilistic model

$$\begin{aligned} \mathbf{P}(\mathbf{y}_k) &= \prod_{i=1}^{n} \xi_i^{y_k(i)} (1-\xi_i)^{1-y_k(i)} \\ &\times \left[1 + \sum_{i<j} r_{ij} \left(\frac{y_k(i) - \xi_i}{\sqrt{\xi_i(1-\xi_i)}} \right) \left(\frac{y_k(j) - \xi_j}{\sqrt{\xi_j(1-\xi_j)}} \right) \right] \end{aligned} \tag{4.4}$$

The number of parameters to be estimated in the above model is too big $(35 + \binom{35}{2})$ compared with the number of observations (87) in the data set. Therefore, we cannot work with this general model. To get reliable estimates we need to reduce the parameter space. First, let $r_{ij} = 0$ if $\mid i - j \mid \geq 1$ and $\xi_i = \xi, \forall i$. Then, for illustration purposes, let us discard all positions with less than 20% of mutation and still consider $r_{ij} = 0$ if $\mid i - j \mid \geq 1$. In this case we have 12 positions and 23 (12+11) parameters. The 12 positions we are now dealing with are: 9, 10, 11, 13, 14, 19, 20, 22, 23, 24, 25 and 32. The results are shown in Tables 4 and 5.

Table 4: Bahadur Representation Model (35 positions)

Parameter	ξ	$r_{1,2}$	$r_{2,3}$	$r_{3,4}$	$r_{4,5}$	$r_{5,6}$	$r_{6,7}$	$r_{7,8}$	$r_{8,9}$
MLE	0.24	0.57	0.47	0.75	0.48	0.65	0.32	0.40	-0.02
Parameter	$r_{9,10}$	$r_{10,11}$	$r_{11,12}$	$r_{12,13}$	$r_{13,14}$	$r_{14,15}$	$r_{15,16}$	$r_{16,17}$	$r_{17,18}$
MLE	0.11	0.14	-0.11	-0.05	0.25	0.01	0.53	0.70	0.46
Parameter	$r_{18,19}$	$r_{19,20}$	$r_{20,21}$	$r_{21,22}$	$r_{22,23}$	$r_{23,24}$	$r_{24,25}$	$r_{25,26}$	$r_{26,27}$
MLE	-0.07	0.36	-0.21	0.13	0.11	0.31	0.26	-0.20	0.22
Parameter	$r_{27,28}$	$r_{28,29}$	$r_{29,30}$	$r_{30,31}$	$r_{31,32}$	$r_{32,33}$	$r_{33,34}$	$r_{34,35}$	
MLE	-0.01	0.49	0.00	0.49	-0.11	-0.26	0.42	0.47	

Comparing the results of Tables 4 and 5, we see that the assumption of equal probability of mutation for all positions (model 1) is not tenable. For instance, in Table 5, the estimated probability of mutation at position 14 is 0.18 while at position 13 it is 0.73, despite the fact that $r_{13,14}$ is 0.77, indicating high correlation between these adjacent sites. A negative estimate of r_{ij} means that the mutation rate at position i is lower than at position j.

Table 5: Bahadur Representation Model (12 positions)

Parameter	ξ_9	ξ_{10}	ξ_{11}	ξ_{13}	ξ_{14}	ξ_{19}	ξ_{20}	ξ_{22}	ξ_{23}
MLE	0.30	0.35	0.53	0.73	0.18	0.27	0.55	0.40	0.26
Parameter	ξ_{24}	ξ_{25}	ξ_{32}	$r_{9,10}$	$r_{10,11}$	$r_{11,13}$	$r_{13,14}$	$r_{14,19}$	$r_{19,20}$
MLE	0.58	0.73	0.30	-0.15	0.16	-0.16	0.77	0.49	0.30
Parameter	$r_{20,22}$	$r_{22,23}$	$r_{23,24}$	$r_{24,25}$	$r_{25,32}$				
MLE	-0.22	0.09	0.31	-0.27	0.14				

5 Concluding Summary

The autologistic model formulated in Section 2 is able to handle situations involving spatial binary data and dependence on covariates. One problem is that the estimation procedure is not straightforward and computer-intesnsive MCMC procedures are needed. Also, the general model formulation involves too many parameters and when applied to data sets of moderate size, we need to reduce the number of parameters by imposing restrictions. For the data set we use, the number of sequences is small compared to the number of positions, and we end up with a very simple model because of the restrictions we impose on the parameter space. For instance, we assume equal mutation rate and equal correlation structure over all positions, which may not be true in reality.

The model based on the Bahadur representation (Section 3) can also handle dependent binary data, but the inclusion of dependent covariates is not as easy as in the autologistic model. An advantage of this model is that the likelihood function has a closed form and the parameter estimates are obtained by the maximum-likelihood approach using numerical optimization methods, such as the Newton-Raphson procedure. Again, we have to reduce the parameter space when applying this model to our data set, because the sample size (the number of independent sequences) is small compared to the number of positions. Ideally, the ratio between the number of sequences and the number of positions should be at least 5. When we discard all positions with frequency of mutation less than 20%, we see that the mutation rate varies a lot from one position to the other, confirming the fact that the assumption of equal mutation rate over all positions does not hold.

References

[1] K K Amfoh, R F Shaw, and G E Bonney. The use of logistic models for the analysis of codon frequencies of DNA sequences in terms of explanatory variables. *Biometrics*, 50:1054–1063,

1994.

[2] R R Bahadur. A representation of the joint distribution of responses to n dichotomous items. In H Solomon, editor, *Studies in Item Analysis and Prediction*, pages 158–176. Stanford University Press, 1961.

[3] J O Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, second edition, 1980.

[4] J Besag. Nearest-neighbor systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B*, 34:75–83, 1972.

[5] J Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[6] J Besag. Statistical of non-lattice data. *The Statistician*, 24:179–195, 1975.

[7] J Besag. On the statisitcal analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, pages 259–302, 1986.

[8] J Besag. Digital image processing towards Bayesian image analysis. *Journal of Applied Statistics*, 16:395–407, 1989.

[9] R C Bose, I M Chakravarti, P C Mahalanobis, C R Rao, and K J C Smith. A representation of the joint distribution of responses to n dichotomous items. In R C Bose, editor, *Essays in Probability and Statistics*, pages 111–132. The University of North Carolina Press, 1970.

[10] L D Broemeling. *Bayesian Analysis of Linear Models.* Marcel Dekker, INC, 1985.

[11] G Casella and E I George. Explaining the Gibbs sampler. *The American Statistician*, 46:167–174, 1992.

[12] A D Cliff and J K Ord. *Spatial Processes - Models and Applications.* Pion, 1981.

[13] N A C Cressie. *Statistics for Spatial Data.* John Wiley & Sons, INC, second edition, 1993.

[14] A P Dawid. Introduction to Bayesian statistics. In S Kotz and N L Johnson, editors, *Encyclopedia of Statistics 4*, pages 89–105. John Wiley & Sons, 1983.

[15] A E Gelfand, S E Hills, A Racine-Poon, and A F M Smith. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85:972–985, 1990.

[16] A E Gelfand and A F M Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[17] S Geman and D Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[18] C J Geyer and E A Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54:657–699, 1992.

[19] J M Hammersley and P Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, Oxford University, 1971.

[20] W K Hastings. Monte Carlo sampling methods using Markov chain and their applications. *Biometrika*, 57:97–109, 1970.

[21] W Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. of Math. Stat.*, 19:293–325, 1948.

[22] F W Huffer and H Wu. Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Not published yet, 1995.

[23] F W Huffer and H Wu. Variable selection in auto-logistic models. Not published yet, 1995.

[24] S Kullback. Kullback information. In S Kotz and N L Johnson, editors, *Encyclopedia of Statistics 4*, pages 421–425. John Wiley and Sons, 1983.

[25] G J LaRosa, J P Davide, K Weinhold, J A Waterbury, A T Profy, J A Lewis, A J Langlois, G R Dreesman, R N Boswell, P Shadduck, L H Holley, M Karplus, D P Bolgnesi, T J Matthews, E A Emini, and S D Putney. Conserved sequence and structural elements in the HIV-1 principal neutralizing determinant. *Science*, 249(4971):932–935, 1990. published erratum appears in Science 1991 Feb 15, **251**(4995), 811.

[26] K Liang, S L Zeger, and B Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, 54:3–40, 1992.

[27] G Myers, B T M Korber, J A Berzovsky, R F Smith, and G F Pavlakis. Human Retrovirus and AIDS. Theoretical Biology and Biophysucs Group, Los Alamos National Laboratory, New Mexico, 1991.

[28] G Myers, B T M Korber, Wain-Hobson, R F Smith, and G F Pavlakis. Human Retrovirus and AIDS, 1993 Compendium, Part III. Theoretical Biology and Biophysucs Group, Los Alamos National Laboratory, New Mexico, 1993.

[29] G Myers, B T M Korber, Wain-Hobson, R F Smith, and G F Pavlakis. Human Retrovirus and AIDS, April 1994 update to the 1993 Compendium. Theoretical Biology and Biophysucs Group, Los Alamos National Laboratory, New Mexico, 1994.

[30] G Myers, B T M Korber, S Wain-Hobson, R F Smith, and G F Pavlakis. Human Retrovirus and AIDS. Theoretical Biology and Biophysucs Group, Los Alamos National Laboratory, New Mexico, 1993.

[31] G Myers, A B Rabson, J A Berzovsky, T F Smith, and F Wong-Staal. Human Retrovirus and AIDS. Los Alamos: Los Alamos National Laboratory, 1990.

[32] E Parzen. *Stochastic Processes*. Holden-Day, INC, 1962.

[33] A E Raftery. A model for high-order markov chains. *Journal of the Royal Statistical Society, Ser.* **B**, 47:528–539, 1985.

[34] A E Raftery and S Tavaré. Estimation and modeling repeated patterns in high order Markov chains with the mixture transition distribution model. *Applied Statistics*, 43:179–199, 1994.

[35] R L Smith. A tutorial on Markov chain Monte Carlo. Unpublished Manuscript, 1994.

[36] R L Smith, M O Calloway, and J P Morrissey. Network autocorrelation with a binary dependent variable: A method and an application. Not yet published, 1994.

[37] D J Strauss. Clustering on coloured lattices. *Journal of Applied Probability*, 14:135–143, 1977.

[38] M A Tanner. *Tools for Statistical Inference*. Springer-Verlag, second edition, 1993.

[39] H Wu and F W Huffer. Modeling the distribution of plant species using the autologistic regression model. Not published yet, 1995.