

Exploratory Bayesian Model Selection for Serial Genetics Data

Jing X. Zhao,^{1,*} Andrea S. Foulkes,² and Edward I. George³

¹Division of Biostatistics, Department of Biostatistics and Epidemiology,
University of Pennsylvania School of Medicine, 503 Blockley Hall, 423 Guardian Dr.,
Philadelphia, Pennsylvania 19104, U.S.A.

²Department of Biostatistics, University of Massachusetts, Amherst, Massachusetts 01003, U.S.A.

³Department of Statistics, University of Pennsylvania Wharton School, Philadelphia,
Pennsylvania 19104, U.S.A.

**email*: jzhao@cceb.upenn.edu

SUMMARY. Characterizing the process by which molecular and cellular level changes occur over time will have broad implications for clinical decision making and help further our knowledge of disease etiology across many complex diseases. However, this presents an analytic challenge due to the large number of potentially relevant biomarkers and the complex, uncharacterized relationships among them. We propose an exploratory Bayesian model selection procedure that searches for model simplicity through independence testing of multiple discrete biomarkers measured over time. Bayes' factor calculations are used to identify and compare models that are best supported by the data. For large model spaces, i.e., a large number of multileveled biomarkers, we propose a Markov Chain Monte Carlo (MCMC) stochastic search algorithm for finding promising models. We apply our procedure to explore the extent to which HIV-1 genetic changes occur independently over time.

KEY WORDS: Bayes' factor; Bayesian model selection; HIV mutations; Markov chain Monte Carlo simulation; Markov process; Metropolis–Hasting algorithm.

1. Introduction

A wide variety of mutations in the viral genome of human immunodeficiency virus type-1 (HIV-1) are associated with reduced susceptibility to antiretroviral therapies (ARTs) (Shafer, 2001). These mutations occur either under selective drug pressure or as a result of naturally occurring polymorphisms. Identifying pathways in which viral populations progress genotypically in the presence (or absence) of particular therapies may provide insight into the mechanism of resistance and help make more informed treatment decisions. However, characterizing the genetic progression to resistance over time continues to present an analytic challenge due to the large number of possible mutations at each site on the viral sequence and the complex, uncharacterized interactions among them.

This manuscript aims to characterize statistically the dependency among viral genetic loci measured over time. We propose an exploratory Bayesian model selection procedure to identify Markov models that best describe this mutation process. This procedure is built on a three-stage hierarchical mixture model for the Markov model uncertainty. Because the number of potential models is large for even a modest number of loci, we propose a Metropolis–Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970; Chib and Greenberg, 1995), a Markov chain Monte Carlo (MCMC) procedure, which stochastically searches for promising models by sampling from the posterior distribution over the set of mod-

els. The potential of such a stochastic search follows from the simple observation that higher probability models are more likely to be sampled (George and McCullough, 1993). This exploratory approach to characterizing the viral genetic pathways can provide interesting biological and clinical insights.

The application of a hidden Markov modeling framework to serial viral genetics data is described in Foulkes and DeGruttola (2003). This approach involves assigning patients to states at each observed time point based on the genetic characteristics of their viral populations and modeling the transition rates between these states assuming a first-order stationary continuous-time Markov process. The existence of multiple clonal sequences for an individual at each time point motivates the use of an expectation–maximization algorithm to account for missingness in state assignment. In this manuscript, we assume a similar Markov process. However, the method we describe allows us to search among multiple nonnested models to arrive at the one that provides the best representation of our data.

Consider for example three amino acid (AA) sites on the protease region, indexed by M46, I50, and I84, where each number represents the location on the viral genome and the letters represent the corresponding consensus (wild-type) AA. For simplicity, let us assume that there are two possible AAs at each site, denoted mutant and wild-type. (In general, this is not true and our method does not require dichotomous biomarkers.) In this case, there are $2^3 = 8$ possible states

(each defined by a distinct AA sequence) and the method described in Foulkes and DeGruttola (2003) would estimate each of the $64 - 8 = 56$ transition rates between these states. Our method searches for independence structure, i.e., whether the rates of mutating (or reverting back to wild-type) are independent across sites. So, for example, our approach would allow us to discover that changes in the AAs at sites M46, I50, and I84 occur independently. This would imply that $2 + 2 + 2 = 6$ parameters could fully characterize the 56 possible transitions. This approach is a natural analytic technique to explore biological relationships that can later be confirmed through rigorous experimentation and further analyses.

Our work is motivated by research suggesting that certain mutations on the viral genome subsist only in the presence of substitutions at other positions and that these linked mutations may be beneficial to the virus (Karnoub, Seillier-Moiseiwitsch, and Sen, 1999). Karnoub et al. (1999) propose a binomial independence test based on a 2×2 contingency table. Their approach allows for testing the mutational linkage of two sites by treating each site as a binary variable (i.e., mutant or not) and is described for cross-sectional data. Hoffman, Schiffer, and Swanstrom (2002) similarly characterize the correlations between pairs of positions in the HIV-1 protease sequence of patients with and without PI exposure based on the Shannon entropy. Our approach has the advantage of allowing for testing of independence across a large number of sites and additionally accounts for changes over time. Furthermore, incorporating covariates is straightforward in this modeling context.

While the likelihood ratio test (LRT) is a natural choice for comparing models, it is inadequate in our setting due to the large number of genetic loci, the multiple AA observed at each site and the nonnested nature of the models to be compared. For these reasons, we propose a Bayesian model selection procedure. This approach has the advantage of accommodating many parameters with relatively small sample sizes. Furthermore, it can be applied to sparse data and allows for testing multiple, competing nonnested hypotheses. The Bayes factors (BFs) comparing any given pair of models are calculated for hypothesis testing. When the model space becomes large, MCMC search algorithms can be used to ease the computational burden of the exhaustive pairwise comparisons between all of the models.

Our investigation focuses on the viral genetics setting; however, the methods we describe can be applied to other settings in which multiple discrete biomarkers are measured over time. The data motivating this research were generated from 170 HIV patients in three phase II clinical studies of Efavirenz (EFV) combination therapy and is available for public use in the Stanford HIV RT and Protease Sequence Database (Shafer, 2001). A detailed description of these data can be found in Bachelier et al. (2000, 2001). Briefly, patients were randomized to receive EFV or placebo plus Indinavir or AZT and 3TC or two nucleoside reverse transcriptase inhibitors (NRTIs). Viral protease and reverse transcriptase sequences (consisting of 99 and 229 AA sites, respectively) from multiple HIV-1 clones are available for each individual at each time point. A sample of the data is given in Table 1.

Table 1
A sample of protease sequences data

ID	Week	Clone	Protease Sequences (site 1–99)																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																								
10	0	1	P	Q	I	T	L	W	Q	R	P	I	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	I	V	T	...	T	Q	L	G	C	T	L	N	F	4	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	5	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	K	I	G	C	T	L	N	F	6	P	Q	I	T	L	W	Q	R	P	I	V	T	...	T	Q	L	G	C	T	L	N	F	7	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	11	0	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	32	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	72	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	12	0	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	V	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	13	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	A	T	...	T	Q	I	G	C	T	L	N	F	24	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	13	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	10	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	37	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	A	T	...	T	Q	I	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	14	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	9	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F
		11	0	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	32	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	72	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	12	0	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	V	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L			G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	13	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q		I	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	A	T	...	T	Q	I	G	C	T	L	N	F	24	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T		Q	I	G	C	T	L	N	F	13	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T			...	T	Q	I	G	C	T	L	N	F	10	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V		T	...	T	Q	I	G	C	T	L	N	F	37	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L		A	T	...	T	Q	I	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L			V	T	...	T	Q	I	G	C	T	L	N	F	14	0	1	P	Q	V	T	L	W	Q	R		P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	9	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q		R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F																																																																																																																																																					
				12	0	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	V	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L		N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	13	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T		L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	A	T	...	T	Q	I	G	C	T			L	N	F	24	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	13	0	1	P	Q	V	T	L	W	Q	R	P	L			V	T	...	T	Q	I	G	C	T	L	N	F	10	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q		L	G	C	T	L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I		G	C	T	L	N	F	37	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I			G	C	T	L	N	F	2	P	Q	I	T	L	W			Q	R	P	L	A	T	...	T	Q	I		G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	14	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q		I	G	C	T	L	N	F	9	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F																																																																																																																																																																																																																																																																								
						24	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	13	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	10	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T		L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G			C	T	L		N	F	37	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T			L	N	F	2	P	Q	I	T	L	W	Q			R	P	L	A	T	...	T	Q	I	G	C	T		L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	14	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T		L	N	F	9	1	P	Q	I	T	L	W	Q	R	P	L	V	T			...	T	Q	L	G	C	T	L	N	F	2	P	Q			V	T	L	W	Q	R	P	L	V	T		...	T	Q	I	G	C	T	L	N	F																																																																																																																																																																																																																																																																																																																																																																																											
							13	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L			N	F	10	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T		L	N	F	2	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	3	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	37	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T			Q	I	G	C	T	L		N	F	2	P	Q	I	T	L	W	Q	R	P	L	A	T	...	T	Q	I	G			C	T	L	N	F	3	P	Q	I	T	L			W	Q	R	P	L	V	T	...	T	Q	I	G		C	T	L	N	F	14	0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C			T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F	9	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T		Q	L	G		C	T	L	N	F	2	P	Q	V	T	L	W	Q			R	P	L	V	T	...	T	Q	I	G	C	T	L	N	F																																																																																																																																																																																																																																																																																																																																																																																																																
						14			0	1	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T			L	N		F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C		T	L	N	F	9	1	P	Q	I	T	L	W	Q	R	P	L	V	T	...	T	Q	L	G	C	T	L	N	F	2	P	Q	V	T	L	W	Q	R	P	L	V	T	...	T	Q	I	G	C	T	L		N	F																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																

In Section 2, we describe Markov models for the evolution of multiple serial genetic indicators. We then introduce our Bayesian model selection procedure and the MCMC search algorithm for large model spaces. In Section 3, we illustrate our method using the data example described above. In Section 4, we explore further a couple of issues including the time homogeneity assumption and treatment effects. Finally in Section 5, we summarize our findings and describe areas of future methodological and applied research.

2. Bayesian Model Selection

2.1 The Markov Model

Suppose each state in a Markov process is represented by a unique pattern of AAs across R sites in the viral genome. For example, if $R = 3$ and each site has four possible AA, then there are $4^3 = 64$ possible states. We denote these states G_1, \dots, G_N , where $G_i = (g_{i1}, g_{i2}, \dots, g_{iR})'$ is a vector of length R with AAs at the corresponding sites as elements. The Markov process can now be described by (1) where the state space is $\mathcal{S} = (G_1, G_2, \dots, G_N)$ and $X(t)$ is a random variable for state membership at time t :

$$p_{ij} = \Pr(G_i \rightarrow G_j) = \Pr(X(t+1) = G_j | X(t) = G_i). \quad (1)$$

To summarize the data, we use Y to denote the $N \times N$ matrix of transition counts, where the (i, j) th element, y_{ij} , is the number of transitions from G_i to G_j across all patients over time. For example, for the simple case of $R = 2$ sites, let us consider the possible states for the combination of site 3 and site 10 in Table 1. There are two possible AAs at site 3: wild-type (I) and mutant (V), and two possible AAs at site 10: wild-type (L) and mutant (I). Thus, there are four possible states for these two sites: (IL), (II), (VL), and (VI). For our complete data set, the 4×4 matrix Y is then given by

$$Y = \begin{array}{cccc|cccc} & \text{IL} & \text{II} & \text{VL} & \text{VI} & & & & & \\ \text{IL} & y_{11} & y_{12} & y_{13} & y_{14} & \text{IL} & 343 & 20 & 1 & 0 \\ \text{II} & y_{21} & y_{22} & y_{23} & y_{24} & \text{II} & 5 & 90 & 0 & 0 \\ \text{VL} & y_{31} & y_{32} & y_{33} & y_{34} & \text{VL} & 2 & 0 & 0 & 0 \\ \text{VI} & y_{41} & y_{42} & y_{43} & y_{44} & \text{VI} & 0 & 0 & 0 & 0, \end{array} \quad (2)$$

where $y_{11} = 343$ is the number of transitions from (IL) to (IL), $y_{12} = 20$ is the number of transitions from (IL) to (II), etc. Note that under the Markov model, the transition counts $\{y_{ij}; i, j = 1, \dots, N\}$ have a product multinomial distribution

$$\prod_{i=1}^N \left(\frac{y_{i.}!}{\prod_{j=1}^N y_{ij}!} \right) \prod_j (p_{ij})^{y_{ij}}, \quad (3)$$

where p_{ij} is the transition probability in (1) and $y_{i.} = \sum_j y_{ij}$ is the i th row sum.

Our aim is to identify the model that best fits our data among all possible models for the dependence structure across sites. Returning to our simple example, if $R = 3$ sites are under consideration we might want to determine whether these three sites mutate independently or mutate together. We can test formally, $H_0: P(G_i \rightarrow G_j) = \Pr(g_{i1} \rightarrow g_{j1}) * \Pr(g_{i2} \rightarrow g_{j2}) * \Pr(g_{i3} \rightarrow g_{j3})$ versus the alternative $H_A: P(G_i \rightarrow G_j)$. In general, a LRT can be used to test for the independence

of sites under this Markov model. However, as the number of sites increases, the number of possible models will also escalate. In addition to the two extreme cases, dependence among all sites and independence among all sites, intermediate dependence models must also be considered. For instance, in our three site example, these intermediate models are: (1) $(r_1)(r_2, r_3)$ where sites r_2 and r_3 change together but they mutate independently of site r_1 ; (2) $(r_1, r_2)(r_3)$ where sites r_1 and r_2 change together but they mutate independently of site r_3 ; and (3) $(r_1, r_3)(r_2)$ where sites r_1 and r_3 change together but they mutate independently of site r_2 . In general, the number of possible models is given by the Bell number, $\text{Bell}_R = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^R}{k!}$.

The LRT has limited effectiveness in this setting for many reasons. First, the test may not be powerful enough in relatively small sample size settings with so many parameters, and so may tend to favor the null. Second, the asymptotic approximations may be inapplicable because the data are often sparse and unbalanced. For example, at a given site, the vast majority of observed AAs are often wild-type, and therefore few transitions between other states are observed. We avoid such limitations of the LRT by using a Bayesian framework that accounts for all postdata model uncertainty with the posterior distribution of model probabilities. A further advantage of the Bayesian framework is that it facilitates the construction of an MCMC model search procedure to find promising models in large problems when it is not computationally feasible to make all pairwise comparisons.

2.2 A Three-Stage Hierarchical Mixture Model

Suppose a set of K models $\mathcal{M} = \{M_1, \dots, M_K\}$ are under consideration and our objective is to select the most appropriate model in \mathcal{M} for the data. Bayesian model selection focuses on the model posterior distribution $p(M_1 | Y), \dots, p(M_K | Y)$, which summarizes all of the relevant information in the data Y for distinguishing between the proposed models in \mathcal{M} . The probability that M_k is the correct underlying model for our data conditional on observing Y is given by (4). Note that the marginal distribution of the data $p(Y | M_k)$ can be expressed in terms of $p(Y | \theta_k, M_k)$ and $p(\theta_k | M_k)$ as described in (5), where θ_k is the parameter vector (of transition probabilities in our example) for M_k :

$$p(M_k | Y) = \frac{p(Y | M_k)p(M_k)}{\sum_k p(Y | M_k)p(M_k)}, \quad (4)$$

$$p(Y | M_k) = \int p(Y | \theta_k, M_k)p(\theta_k | M_k) d\theta_k. \quad (5)$$

We now specify the three-stage hierarchical mixture model as described by Chipman, George, and McCulloch (2001) that includes the following components: (1) a prior for each model, given by $p(M_k)$, (2) a prior for the parameters of each model, given by $p(\theta_k | M_k, \alpha_k)$, where θ_k is a vector of unknown parameters that indexes the members of M_k , and (3) a density for the observed data, Y given by $p(Y | \theta_k, M_k)$. In the absence of reliable prior information, the most common and practical approach to prior specification is to construct ‘‘non-informative,’’ semiautomatic formulations (Chipman et al., 2001). Thus, we consider the following default specifications:

1. Model space prior $p(M_k) = 1/K$ (Uniform). Under this prior, the model posterior is proportional to the marginal likelihood ($p(M_k | Y) \propto p(Y | M_k)$) allowing for the posterior odds simplification described in the next section.
2. Parameter priors. We propose the same distribution for the parameters in all models in our model space and can therefore exclude the subscript k . Conjugate priors are chosen in order to obtain a rapidly computable closed-form solution for the marginal distribution. Because our data have product multinomial distributions, we use a Dirichlet distribution (beta distribution in binomial case) for the transition probabilities. This is given in (6). For simplicity, we consider a single prior distribution across all start states (i.e., we can omit the row index i here). The prior specification is then completed by choosing values for the Dirichlet hyperparameter $\alpha = \{\alpha_1, \dots, \alpha_N\}$. By using this Dirichlet prior, conjugate prior for the multinomial, we lessen the computational burden dramatically:

$$p(\theta | \alpha) = \frac{\Gamma(\alpha_1 + \dots + \alpha_N)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_N)} \theta_1^{\alpha_1 - 1} \dots \theta_N^{\alpha_N - 1}. \quad (6)$$

3. Distribution of the observed data. As previously described, the probability distribution of the observed data is a product multinomial distribution. Thus, $p(Y | \theta_k, M_k)$ is of the form (3), where θ_k is the set of transition probabilities $\{p_{ij}\}$ corresponding to the set of states defined by model M_k .

2.3 The Bayes Factors

Pairwise comparison of two models, M_1 and M_2 , are summarized by the posterior odds given in (7). This expression reveals how the data, through the BF $B_{12} = p(Y | M_1) / p(Y | M_2)$ updates the prior odds $p(M_1) / p(M_2)$ to yield the posterior odds. Model selection proceeds by choosing M_k corresponding to the highest $p(M_k | Y)$:

$$\frac{p(M_1 | Y)}{p(M_2 | Y)} = \frac{p(Y | M_1)}{p(Y | M_2)} \times \frac{p(M_1)}{p(M_2)}. \quad (7)$$

Under our uniform prior on the model space, the prior odds are identically 1, and so the posterior odds will be identical to the BF. Thus, pairwise comparisons of models here can be completely summarized by the BF comparisons. Furthermore, the BF calculations are easily obtained, using the fact that here each marginal distribution of the data is given by:

$$p(Y | M) = \left[\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_N)} \right]^N \times \prod_i \left[\left(\frac{y_i!}{\prod_j y_{ij}!} \right) \frac{\Gamma(y_{i1} + \alpha_1) \dots \Gamma(y_{iN} + \alpha_N)}{\Gamma(y_i + \alpha_0)} \right]. \quad (8)$$

2.4 MCMC Stochastic Search

As shown above, rapidly computable closed-form expressions for the posterior odds are available under our model setup. Although extremely useful for comparing any given pair, ex-

haustive comparison for all models is simply not feasible when the model space is large. This quickly occurs in our problem because the size of the model space increases sharply as the number of sites increases. To mitigate this difficulty in larger problems, we propose an MCMC stochastic search procedure based on a MH algorithm (Metropolis et al., 1953; Hastings, 1970; Chib and Greenberg, 1995). This procedure simulates a (correlated) sample of models from the posterior distribution (4) over \mathcal{M} . As it stochastically moves around the model space, it tends to locate more rapidly the higher posterior probability models because these are more likely to be visited. The visited models can then be easily compared using the BF calculation just described. Essentially, the MCMC algorithm is a heuristic method for reducing attention to a manageable subset of models.

Our MCMC procedure will simulate a Markov chain of models, say $M^{(1)}, M^{(2)}, \dots$, that converges to the posterior distribution $p(M | Y)$ over \mathcal{M} (Chipman et al., 2001). Our MH algorithm does this by successive sampling from a transition kernel or proposal density $q(M | M^{(j)})$ (a probability distribution over \mathcal{M} for each given $M^{(j)}$), imposing a particular accept-reject decision step at each transition. To define an MH algorithm for our problem, we simply need to specify the model space \mathcal{M} and a transition kernel $q(M | M^{(j)})$.

Because our models differ from each other by the independence relationship among the group of examined sites, we consider a transition kernel that randomly picks one pair of sites and then moves from the model in which these are dependent to the model in which they are independent or vice versa. Thus, our transition kernel q (=constant) is symmetric, and our algorithm is a special case of the original Metropolis algorithm (Metropolis et al., 1953). The following step-by-step procedure summarizes our algorithm:

1. Randomly select a pair of sites $\{r_1, r_2\}$ from the group of sites in the current model M .
2. If r_1 and r_2 are independent in M , choose M' to be the model in which these two are dependent. If r_1 and r_2 are dependent in M , choose M' to be the model in which these two are independent.
3. Calculate the posterior odds: $p(M' | Y) / p(M | Y)$.
4. Accept the transition to M' with probability: $\min\{1, p(M' | Y) / p(M | Y)\}$. Otherwise do not move and remain at M .
5. Repeat steps 1–4 B times (where B is a large number) to obtain a (correlated) sample from \mathcal{M} .

In problems where it is not feasible to exhaustively compute the full posterior distribution, this algorithm can be used to quickly identify many of the high probability models. The Bayes factor calculations can then be used to compare all the visited models. Note that this Bayes factor calculation is part of the algorithm in step 3, and so can simply be recorded as the algorithm proceeds.

3. Example and Results

For the purpose of illustration, we apply our methods to four sets of sites with our data on viral sequences from 170 HIV-1 patients. The four sets of sites we consider are $\{46, 50, 84\}$, $\{32, 46, 50, 84\}$, $\{32, 46, 50, 82, 84\}$, and $\{32, 46, 50, 54, 82,$

84}, all of which are associated with PI resistance (Shafer, 2001). And we apply our methods with the simple hyperparameter values specification $\alpha = (1, \dots, 1)^T$, under which the Dirichlet distribution for the transition probability parameters θ is uniform, $\text{Dirichlet}(1, 1, 1, \dots) \sim \Gamma(N)/1 \cdots 1 = \Gamma(N)$.

3.1 Single Random Selection (SRS) Analysis

In our data, 120 of the 170 HIV-1 patients are measured at more than one time point. Furthermore, multiple clonal sequences are observed at each of these time points. To obtain ordered sequences of clones for the application of our method, we begin by randomly selecting a single clone from the multiple clones at each time point for each of the 120 patients. We refer to the analysis of a single draw of such sequences as a single random selection (SRS) analysis.

In the case of the three sites {46, 50, 84}, the model space consists of five possible models. Because the number of models is small, all Bayes' factors can be easily computed. Column 3 of Table 2 (Panel A) presents the BF's for each model compared to the independence model (46)(50)(84), which itself has a BF of 1. Here, this BF of 1 is the largest, suggesting that the model in which these three sites change independently is most favored by the data. The intermediate model, (46)(50, 84) is the second highest posterior probability model with a BF of 0.869. Note also that all the BF's provide decisive evidence against the fully dependent model (46, 50, 84), according to the calibration suggested by Jeffreys (1961, Appendix B).

When we add site 32 and consider {32, 46, 50, 84}, the model space increases to include 15 elements. The 15 BF's comparing each model to the independence model are given in column 3 of Table 2 (Panel B). The two most favored models, (32, 46)(50)(84) and (32, 46)(50, 84), both suggest dependence between changes at sites 32 and 46, and that changes at sites 50 and 84 are independent of those at sites 32 and 46. The most favored model additionally suggests independent changes at sites 50 and 84. When we add site 82 and consider {32, 46, 50, 82, 84}, the model space increases to 52 elements. The 52 BF comparisons with the independence model are displayed in the top left-hand plot of Figure 1. In this case, the highest BF model turned out to be (32, 46)(50, 82)(84). Note that this finding is consistent with the highest BF four-site model, and suggests further dependence between changes at sites 50 and 82. Finally, when we add site 54 and consider {32, 46, 50, 54, 82, 84}, it is still computationally feasible to calculate all 203 BF comparisons with the independence model (results not shown). In this case, the two models with the highest Bayes' factors are the full model (32, 46, 50, 54, 82, 84) and model (32, 46, 50, 54, 82)(84). These two models are very different from the "best" five-site model, (32, 46)(50, 82)(84). This lack of consistency with the five-site results suggest that there may be some mixed structure in the particular SRS data considered here.

Although MCMC model search is not needed in these smaller problems, we applied it here to compare its performance with the exact BF results. For the four-site case above, Figure 2 shows that the visiting frequencies resulting from running the simulation 50,000 times are completely consistent with the exact BF results. Indeed, the two most visited models are (32, 46)(50, 84) and (32, 46)(50)(84). For the five-site case, the left-hand side of Figure 1 compares the BF's with the relative frequency of visits by the MCMC algorithm. Again, the algorithm tends to visit the high posterior probability models more often. In particular, the highest BF model (32, 46)(50, 82)(84) was also the most visited model.

Table 2

SRS and mean RRS Bayes' factors (BFs) for three-site and four-site models

Model number	Model	Bayes' factors	
		SRS	Mean RRS
Three sites {46, 50, 84}			
1 (Full)	(46, 50, 84)	0.002	0.002
2	(46, 50)(84)	0.028	0.069
3	(46)(50, 84)	0.869*	1.020*
4	(46, 84)(50)	0.032	0.009
5 (Independent)	(46)(50)(84)	1*	1*
Four sites {32,46,50,84}			
1 (Full)	(32, 46, 50, 84)	0.186	0.001
2	(32, 46, 50)(84)	0.595	0.005
3	(32, 46, 84)(50)	1.221	0.001
4	(32, 46)(50, 84)	4.759**	0.031
5	(32, 46)(50)(84)	5.474**	0.030
6	(32, 50, 84)(46)	0.468	0.458
7	(32, 50)(46, 84)	0.020	0.006
8	(32, 50)(46)(84)	1.649	0.607
9	(32, 84)(46, 50)	0.021	0.040
10	(32)(46, 50, 84)	0.002	0.002
11	(32)(46, 50)(84)	0.028	0.069
12	(32, 84)(46)(50)	0.741	0.577
13	(32)(46, 84)(50)	0.033	0.009
14	(32)(46)(50, 84)	0.869	1.020**
15 (Independent)	(32)(46)(50)(84)	1	1**

*The null model is the independence model (46)(50)(84).

**The null model is the independence model (32)(46)(50)(84).

3.2 Repeated Random Selection (RRS) Analysis

A weakness of the SRS analysis is its dependence on a single random draw. A more comprehensive analysis would entail analysis of all possible sequences. Although such an analysis is computationally infeasible, it can be approximated by averaging independently repetitions of an SRS analysis, which we refer to as a repeated random selection (RRS) analysis. Below we apply an RRS analysis to the same sites considered in Section 3.1 by taking the geometric mean of the BF's for 25 independent repetitions of an SRS analysis. The results of this RRS analysis are presented in column 4 of Table 2.

For the three sites {46, 50, 84}, the intermediate model (46)(50, 84) has the highest mean BF value under our RRS analysis, and the independence model has the second highest mean BF value. These two models were also the two highest posterior probability models found by our previous SRS analysis. When we add site 32 and consider {32, 46, 50, 84}, the RRS analysis yields a different ordering of the models than did our previous SRS analysis. In particular, the two largest mean BF's correspond to model (32)(46)(50, 84) and the independence model (32)(46)(50)(84). The RRS analysis suggests

Model Searching, # of sites=5
site 32 46 50 82 84

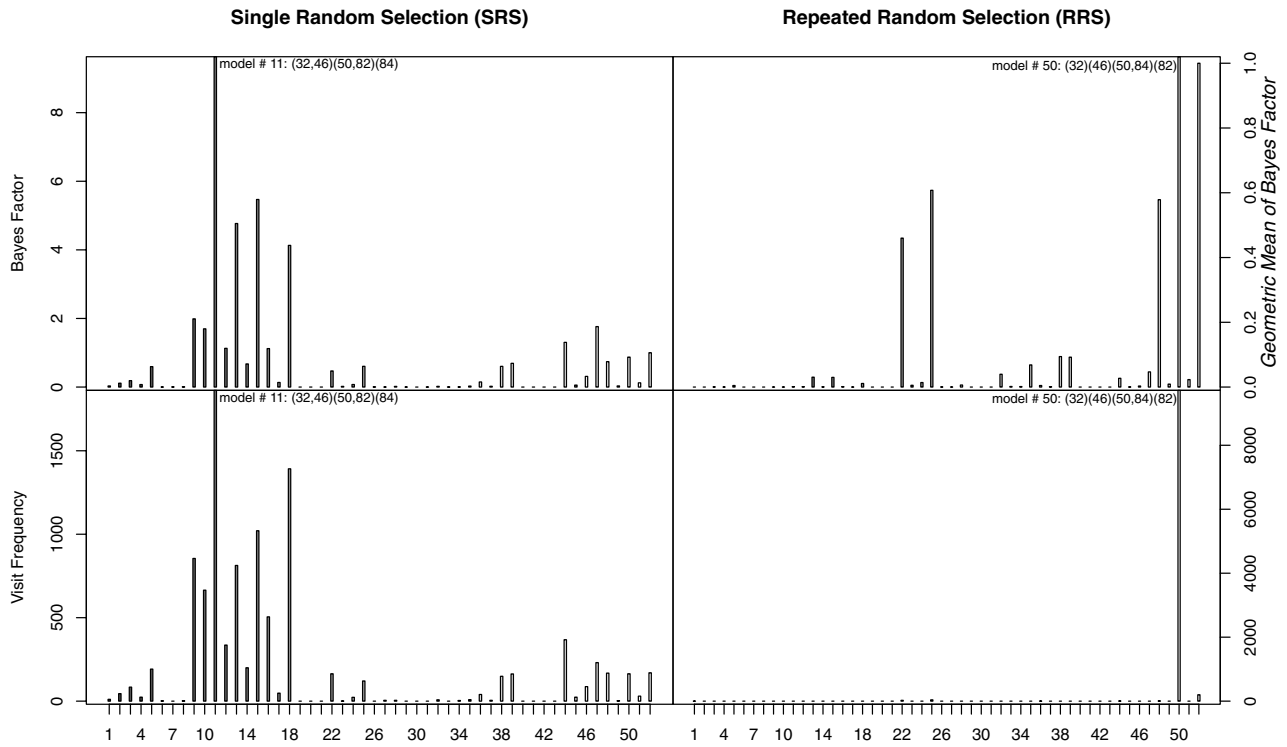


Figure 1. Left: SRS analysis of five-site models (upper left: Bayes' factors (BFs); lower left: MCMC visit frequencies with 10,000 iterations). Model no. 11: (32, 46)(50, 82)(84) had both the largest BF and the highest frequency of visits. Right: RRS analysis of five-site models (upper right: Geometric mean BFs; lower right: MCMC visit frequencies with 10,000 iterations). Model no. 50: (32)(46)(50, 84)(82) had both the largest mean BF and the highest frequency of visits.

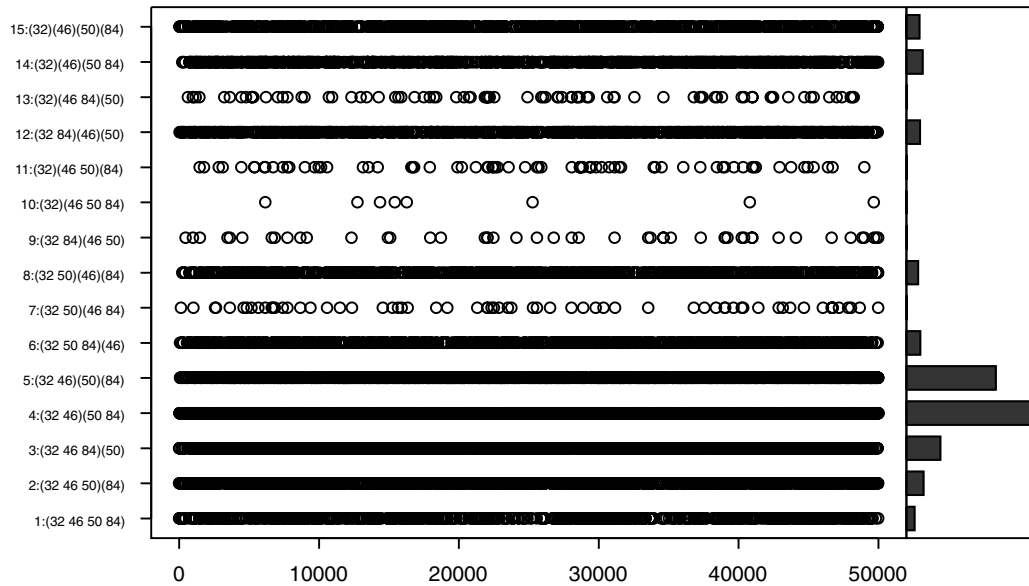


Figure 2. MCMC model search of four-site models for an SRS analysis. (Center: MCMC iteration results; right: Marginal histogram of visit frequencies with 50,000 iterations.) Models (32, 46)(50, 84) and (32, 46)(50)(84) are the first and second most visited models.

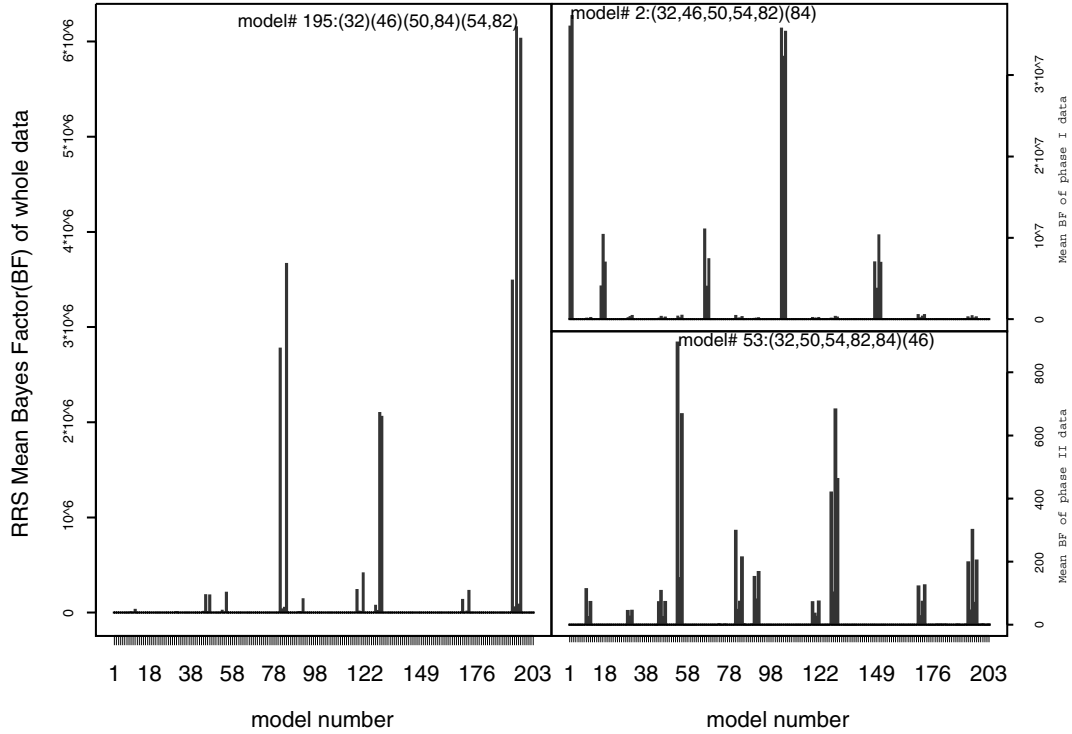


Figure 3. Mean Bayes’ factors for six-site models (left: Overall highest mean BF model is no. 195: (32)(46)(50, 84)(54, 82); upper right: Phase I highest mean BF model is no. 2: (32, 46, 50, 54, 82)(84); lower right: Phase II highest mean BF model is no. 53: (32, 50, 54, 82, 84)(46)).

that site 32 changes independently of the other three sites, while the SRS analysis suggests that changes in site 32 and 46 are dependent. When we add site 82 and consider {32, 46, 50, 82, 84}, the highest mean BF model, namely (32)(46)(50, 84)(82), is consistent with the highest BF four-site model. Compared to the most favored five-site model in the SRS analysis, (32, 46)(50, 82)(84), the most favored RRS model again suggests less dependence between changes at the five sites. The mean BFs for all 52 five-site models are displayed in the top right-hand panel of Figure 1. When we add site 54 and consider {32, 46, 50, 54, 82, 84}, the highest mean BF model was found to be (32)(46)(50, 84)(54, 82), which is completely consistent with our best RRS five-site model. Recall that we did not obtain such consistency with our SRS analysis. Finally, the mean BFs for all 203 six-site models are displayed in the left-hand plot of Figure 3.

To further explore the differences between SRS and RRS analyses, we considered selection of the best models for all three-site subsets from the six sites {32, 46, 50, 54, 82, 84}. In the first two columns of Table 3, we list highest BF and the highest mean BF three-site models for SRS and RRS analyses, respectively. Note that the selected models are often different, and that the RRS models tend to exhibit less dependence between sites. The RRS models also tend to be more compatible with each other, in the sense that dependence between sites is less affected by the presence of a third site. Notably, for the RRS analysis, sites 50 and 84 are consistently dependent, as are the sites 54 and 82.

For large problems, the difficulties of computing all mean BFs will be even more formidable than computing all BFs

in an SRS analysis. Thus, some kind of heuristic search procedure is still needed. To implement such an MCMC model search for an RRS analysis, we suggest using the mean BFs instead of the BFs to drive the MH algorithm. Note, however, that the mean BF values from an RRS analysis are not bona fide BFs in the sense of corresponding to a valid probability distribution. Thus, there is no guarantee that the MH algorithm driven by the mean BFs will converge to a meaningful distribution. Nonetheless, it still may be useful to search for the large mean BF models. To explore this possibility, we applied this MCMC search to the RRS output for the five-site setup considered above. The right-hand side of Figure 1 compares the mean BFs with the relative frequency of visits by the corresponding MCMC algorithm. Although the relative frequencies do not reflect the distribution of the mean BFs well, it is promising that the algorithm spent most of its time visiting the highest BF model (32)(46)(50, 84)(82). Also, the second most often visited model was the second highest BF model. This supports the potential usefulness of such a search in large problems where exhaustive calculation of all BFs is not feasible. Note that the algorithm will still be computationally feasible in such problems, as it only requires sequential calculation of the mean BFs as it moves along.

4. Further Exploration

In the previous section, we illustrated how our Bayesian model selection procedure can be used to explore dependency relationships in a single complex biological system. We now illustrate the further flexibility of our procedure by relaxing the time homogeneity assumption of the model and by accounting

Table 3
All the selected best three-site models for six sites group {32, 46, 50, 54, 82, 84}

Three-sites subset	SRS	Mean RRS	Mean RRS	
	Combined data		With PI	Without PI
32, 46, 50	(32, 46)(50)	(32)(46)(50)	(32, 50)(46)	(32, 46)(50)
32, 46, 54	(32, 46)(54)	(32)(46)(54)	(32)(46)(54)	(32, 46)(54) ^a
32, 46, 82	(32, 46)(82)	(32)(46)(82)	(32)(46)(82)	(32, 46, 82)
32, 46, 84	(32, 46)(84)	(32)(46)(84)	(32, 84)(46)	(32, 46, 84)
32, 50, 54	(32)(50)(54)	(32)(50)(54)	(32, 50)(54)	(32)(50)(54) ^a
32, 50, 82	(32, 50, 82)	(32)(50)(82)	(32, 50)(82)	(32, 50, 82)
32, 50, 84	(32)(50)(84)	(32)(50, 84)	(32, 50, 84)	(32)(50, 84)
32, 54, 82	(32, 54, 82)	(32)(54, 82)	(32)(54, 82)	(32, 82)(54) ^a
32, 54, 84	(32)(54)(84)	(32)(54)(84)	(32, 84)(54)	(32)(54)(84)
32, 82, 84	(32)(82)(84)	(32)(82)(84)	(32, 84)(82)	(32, 82, 84)
46, 50, 54	(46)(50)(54)	(46)(50)(54)	(46)(50)(54)	(46)(50)(54) ^a
46, 50, 82	(46)(50, 82)	(46)(50)(82)	(46)(50)(82)	(46, 50, 82)
46, 50, 84	(46)(50)(84)	(46)(50, 84)	(46)(50, 84)	(46, 50, 84)
46, 54, 82	(46)(54, 82)	(46)(54, 82)	(46)(54, 82)	(46, 82)(54) ^a
46, 54, 84	(46)(54)(84)	(46)(54)(84)	(46)(54)(84)	(46, 84)(54) ^a
46, 82, 84	(46)(82)(84)	(46)(82)(84)	(46)(82)(84)	(46, 82, 84)
50, 54, 82	(50, 54, 82)	(50)(54, 82)	(50)(54, 82)	(50, 82)(54) ^a
50, 54, 84	(50)(54)(84)	(50, 84)(54)	(50, 84)(54)	(50, 84)(54) ^a
50, 82, 84	(50, 82)(84)	(50, 84)(82)	(50, 84)(82)	(50, 82, 84)
54, 82, 84	(54, 82)(84)	(54, 82)(84)	(54, 82)(84)	(54)(82, 84) ^a

^aWe obtained tied BF values for some of the three-site subsets that involved site 54. This occurred because non-PI-treated patients tended to have no mutations at site 54. In such cases, we present the more independent structure for the non-PI group.

for a treatment effect. In both cases, we do this with RRS analyses.

We first relax the time homogeneity assumption by dividing our data into two phases based on the median follow-up time within each patient. For each patient, we refer to the first half as Phase I and to the second half as Phase II. In Figure 3, the two mean BF plots on the right show that patients had very different mutation patterns in Phases I and II of the disease progression. The highest probability model in Phase I is (32, 46, 50, 54, 82)(84), i.e., only site 84 changes independently from the other five sites, whereas in Phase II it becomes (32, 50, 54, 82, 84)(46), i.e., only site 46 changes independently from the other five sites. Alternative relaxations of the time homogeneity assumption can be achieved by introducing time-dependent covariates into the models.

Due to our relatively small sample size, in Section 3 we treated the data as a single group, ignoring the difference between the PI and non-PI treatment groups. However, exploration of the treatment effects is ultimately warranted. To explore this issue, we divided the data into two groups, PI and non-PI exposed, and explored each group separately. In columns 3 and 4 of Table 3, we list all the highest mean BF three-site models from an RRS analysis, and note that there are substantial differences between the dependence structures for each group, which suggests a treatment effect. Future application of our methods to larger data sets will allow us to explore such comparisons more fully.

5. Remarks

We have described and illustrated a new approach for exploring statistical dependence structures for multiple genetic indicators across various sites as they evolve over time. To

deal with multiple clone measurements at each time point, two variants of our method were considered: an SRS analysis that randomly selects a single sequence of genetic transitions for each patient, and an RRS analysis that averages independent repetitions of an SRS analysis. Applying both methods to explore several sets of active sites, the RRS approach found dependence structures that remained more consistent as additional sites were added to the analysis. For larger problems, where exhaustive comparisons are not computationally feasible, we proposed an MCMC stochastic search algorithm to reduce attention to a manageable subset of models. That such an algorithm can be effective was supported by our finding of agreement between visiting frequencies and exact Bayes' factors in computationally tractable problems.

It should be noted that our model treats transition probabilities as the same between consecutive measurements, thereby ignoring the actual time between measurements. If viral sequences are measured when they first become detectable and viral replications are substantially lower between measurements, then such an assumption may be reasonable. However, in future research with larger data sets, we plan to address this assumption by elaborating our models to allow for time dependence of transition probabilities.

It should also be noted that our proposed Markov model is conceived to model transition probabilities at sites on a single clone over time. Unfortunately, our data do not contain such measurements, and because there are no identifiers among the observed clones we are unable to link clones over time. To mitigate this problem, we have considered random reconstruction of such sequences through our SRS and RRS analyses. We also plan to consider more elaborate comprehensive methods in future work. For example, Stephens, Smith, and Donnelly

(2001) propose a Bayesian method for identifying the haplotype information that is applicable to genotype data at linked loci. Analogously, we may be able to use similar methods to reconstruct the linkage between multiple clones across time points.

The flexibility of our methods is briefly illustrated in Section 4, where we adapt it to consider the effect of disease progression and the effect of treatment on dependence structure. Similar extensions of the methods will allow for the control of additional covariates, such as CD4 counts and baseline viral load. Our approach can also be applied to other contexts involving the evolution of multiple correlations among discrete variables changing over time such as, for example, the mutation structure of influenza viruses. Ultimately, our approach provides a flexible framework that allows for the discovery of potentially important clinical and biological relationships.

ACKNOWLEDGEMENTS

We thank the associate editor and the referee for their insightful and constructive suggestions. This work was supported by NSF (DMS-01-30819) and by University of Pennsylvania CFAR (NIH-1-P30-AI45008). Jing X. Zhao was supported in part by Merck BARDS fellowship.

REFERENCES

- Bacheler, L., Anton, E., Kudish, P., et al. (2000). Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy. *Antimicrobial Agents and Chemotherapy* **44**, 2475–2484.
- Bacheler, L., Jeffrey, S., Hanna, G., et al. (2001). Genotypic correlates of phenotypic resistance to efavirenz in virus isolates from patients failing nonnucleoside reverse transcriptase inhibitor therapy. *Journal of Virology* **75**, 4999–5008.
- Chib, S. S. and Greenberg, E. (1995). Understanding the Metropolis–Hasting algorithm. *The American Statistician* **49**, 327–335.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection—Institute of Mathematical Statistics Lecture Notes—Monograph Series*, Volume 38, 67–134.
- Foulkes, A. S. and De Gruttola, V. (2003). Characterizing the progression of viral mutations over time. *Journal of the American Statistical Association* **98**, 859–867.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall. **Q1**
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- Hoffman, N., Schiffer, C., and Swanstrom, R. (2002). Covariation among positions in HIV-1 protease sequences. In *Proceedings of the Third HIV DRP Symposium Antiviral Drug Resistance Session 7*. **Q2**
- Jeffreys, H. (1961). *Theory of Probability*. Cambridge, U.K.: Oxford University Press.
- Karnoub, M. C., Seillier-Moiseiwitsch, F., and Sen, P. K. (1999). A conditional approach to the detection of correlated mutations. In *Statistics in Molecular Biology and Genetics—Institute of Mathematical Statistics Lecture Notes—Monograph Series*, Volume 33, 221–235.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemistry and Physics* **21**, 1087–1091.
- Shafer, R. (2001). The Stanford HIV RT and protease sequence database. Available at <http://hivdb.stanford.edu/hiv/>.
- Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.

Received April 2004. Revised July 2004.

Accepted August 2004.

Queries

Q1 Author: Please provide the publisher location in Gelman et al. (1995).

Q2 Author: Please provide complete information in Hoffman et al. (2002).