

Characterizing the Progression of Viral Mutations Over Time

A. S. FOULKES and V. DE GRUTTOLA

Development and spread of resistance of human immunodeficiency virus type 1 to antiretroviral therapies is a serious medical and public health concern. A wide variety of mutations have been identified that either singly or in combination reduce the susceptibility of the virus to available therapies. This paper describes methods for understanding the genetic pathways that lead to high-level drug resistance under selective drug pressure, as well as for estimating the rates at which viral populations progress along these pathways. These methods can be used to determine whether the presence of certain mutations among drug-sensitive viruses predispose a patient under a particular treatment to develop patterns of mutations that confer high-level drug resistance. Our approach assumes that viral genotypes can be characterized as belonging to discrete states, defined by patterns of viral mutations, and considers two approaches to modeling the rates of transition between these states. The first approach treats the state at a given time point as known, whereas the second treats this as a latent variable. We apply our methods to genetic sequences of viruses cloned from the plasma of 170 patients who participated in three phase II clinical studies of efavirenz combination therapy (DMP 266-003, -004, -005). Multiple viral clones are available from each plasma sample at each time of measurement, allowing for consideration of the effect of minority species on the evolution of the viral populations infecting patients; the availability of such information motivates the second analytic approach. The sequences can be found in the Stanford HIV RT and Protease Sequence Database.

KEY WORDS: Clones; Cluster analysis; EM; Genotype; HIV-1; Latent variable; Markov chain; Repeated measures.

1. INTRODUCTION

Over time patients receiving antiretroviral therapies (ART's) to suppress infection with human immunodeficiency virus type 1 (HIV-1) may stop responding favorably to treatment. One reason for treatment failure is development of antiviral resistance, which occurs either as a result of development of mutations in the viral genome under selective drug pressure or as a result of naturally occurring polymorphisms. Knowledge about the specific genotypic and phenotypic characteristics of the viral populations that infect patients can help in selecting antiviral regimens as well as in the development of new treatments. In particular, understanding how viral populations "evolve" under selective drug pressure (or its absence) can be helpful in predicting how a patient is likely to respond to a given therapy. The term *evolve* appears in quotation marks because the detection of new genetic variants may reflect actual viral evolution or just an improved ability to detect these variants after the sensitive ones are suppressed. The goal of this paper is to investigate the genetic pathways that characterize the transition from sensitive to resistant virus and the rates of progression along these pathways.

Our investigation starts by defining genetic states that are characterized by patterns of viral mutations and then considers two approaches to modeling the rates of transition between these states. The first approach treats the state at a given time point as known, whereas the later treats this state as an unobservable latent variable. The latter approach is motivated by the availability of data from multiple viral clones derived from single plasma samples. Because there may be considerable variability in genetic sequences among the different clones, it is useful to have a method that accommodates this variability in modeling state transitions. Note that the uncertainty regarding state is not simply a measurement problem. If the risk over time

of developing new mutations does not depend on the prevalence of a particular mutation, then neither should the state depend on its prevalence. In other words, if it is the minority species that dominates the evolutionary potential of a viral population, then it should properly determine the state. On the other hand, mutations that have no effect on viral evolution toward more (or less) resistant states should not affect the state, even if they occur in the majority species. Our goal is to classify patients according to risk of developing resistance mutations so that appropriate treatment decisions can be made, just as recipients of blood products are classified to ensure they receive the appropriate blood type. As demonstrated in this paper, we best accomplish this goal by basing our classification on data from multiple clonal sequences measured over time.

Our method is illustrated using data collected from 170 patients who participated in three different phase II clinical studies of efavirenz combination therapy (DMP 266-003, -004, -005). Viral clones were obtained from these patients during the course of the studies and sequences were obtained for each of these clones. These sequences can be found in the Stanford HIV RT and Protease Sequence Database, which is available to the public (Shafer 2001). A sample of sequences from each patient's viral population is observed at each time point; however, these clonal sequences cannot be linked by an identifier. So for example, for a given patient we may observe 10 clonal sequences at week 0 and 12 sequences at week 4. We do not know, however, which, if any, of the sequences observed at week 4 were derived from the sequences observed at week 0. Several recent papers considered how to use genotypic characteristics at a single time point to make predictions about phenotypic response(s) (Foulkes and De Gruttola 2002; Segal, Cummings, and Hubbard 2001). The methods developed in this paper permit estimation of the probability of the virus evolving along different pathways when multiple clonal sequences are available across several time points.

To illustrate our approach in a simple example, consider a setting in which, over time, patients acquire one of two

A. S. Foulkes is Assistant Professor, Division of Biostatistics, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 (E-mail: afoulkes@cceb.upenn.edu). V. De Gruttola is Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Supported by NIH/NIAID grants R01AI51164 and 5R01AI28076. The authors thank Dr. Wing Wong and Dr. Mark Segal for many helpful discussions and Dr. Lee Bacheler and Dr. Robert Shafer for making the data available.

possible sets of mutations, $C_2 = \{L10I, A71V, L90M\}$ or $C_3 = \{D30N, N88D\}$. Here each number represents an amino acid site in the protease region of the viral genome. The letter preceding the number is the most prevalent amino acid (AA) at the corresponding site (commonly referred to as the wild-type or consensus AA) and the letter following each number indicates the observed AA. A mutation is defined by the observed AA differing from the consensus AA. Patients in our simple example all start in a “wild-type state” with no mutations, C_1 ; C_2 and C_3 can be thought of as states to which patients belong if their viral populations exhibit the corresponding mutations at the time of failure. Here failure is defined as the presence of viral rebound, indicating loss of activity of treatment. Based on previous findings, patients in C_2 are expected to show high levels of resistance to a drug from the protease inhibitor (PI) class, indinavir (IDV), whereas patients in C_3 are likely to be partially sensitive to IDV yet highly resistant to another PI, nelfinavir (NFV) (Foulkes and De Gruttola 2002). This paper considers whether it is possible to transition from C_2 to C_3 as well as from C_1 to these states, and provides a method for estimating the rates of transition from each state to the others (some transitions may not be possible). In reality, patients are generally not all in a wild-type state at baseline even if they are naive to treatment, and many more patterns of resistance mutations are possible, so that additional states may be defined. The goal remains to investigate rates of transition among different states.

In addition to describing methods for estimating transition rates, we present an approach to testing hypotheses regarding these rates. As an example, we test whether all transition rates from one state to another are equal. Identifying departures from this particular null hypothesis is important for two reasons. First, knowledge that certain transitions are more common than others may impact treatment choices; for example, this information may be used to avoid exposing a patient to a treatment that is associated with rapid transition from the patient's initial state to a highly resistant state. Second, knowledge of differences in the rates at which viral populations move along genetic pathways to resistance may provide insight into biological mechanisms of disease progression and drug action. As described in subsequent sections, our analyses suggest that mutations at site 82 occur at a significantly faster rate among viral populations with mutations at sites 10, 71, and 77 compared to those with a mutation only at site 63. The fact that mutations at site 82 are strongly associated with resistance to one class of drugs (protease inhibitors) makes it important for both patient management and drug development to identify early predictors of development of this mutation.

The data described in this paper can be thought of as arising from a hidden Markov model (HMM) as described in Durbin, Eddy, Krogh, and Mitchison (1997). In HMM's, the true state sequence or path is unobserved, but assumed to follow a simple Markov chain. One approach to estimate transition probabilities between states makes use of a version of the expectation-maximization (EM) algorithm. The usefulness of this approach in the context of biological sequence analysis, in particular the analysis of a single observation of a very long path, was described by Durbin et al. (1997). In our setting, the hidden trajectory for a patient is the concatenation of sequences obtained by selecting one of the observed sequences from each observed

time point. The observed states are all of the clonal sequences across all time points. As described above, we do not observe the links between clonal sequences over time, that is, we do not know the paths connecting clonal sequences between time points. In our case, the paths are relatively short, with a median number of states for an individual equal to 5. The hidden Markov approach allows us to “draw strength” from the information available on clonal sequences from the entire population to estimate model parameters in settings where the individual paths may not be precisely known.

In addition to the challenges described previously, we must address the interval censoring of the exact times of occurrence of mutations. Viruses can be sequenced only in patients with sufficiently high viral burden, but such sequencing is expensive and therefore infrequent. We assume that viruses that remain below the level of detection do not change states. We also assume that mutations occur when they are detected and that the length of follow-up is independent of the processes we consider. As discussed in Section 4, the shortness of the intervals between measurement justifies ignoring the interval censoring.

Our example considers the protease region of the viral genome, because all patients were protease naive when the studies were initiated and because many different mutations (or naturally occurring polymorphisms) appear to be related to resistance. Although the data were generated during clinical studies of EFV, we chose to consider the protease region, because EFV resistance is dominated by a single mutation. In the investigation of antiviral resistance, the protease and reverse transcriptase regions are of particular interest because they code for enzymes targeted by most currently available therapies. The protease region consists of 99 amino acid sites and the observed residues at these sites vary both between and within individuals. To generate the data for our example, sequencing was done over several time points and, at each time point, multiple clones were sequenced. Genotyping can occur only when the virus is above 50 copies/mm and is generally done at the time that the virus is observed to rebound. A sample of the data is given in Table 1.

Section 2 presents an approach to clustering similar clonal sequences to reduce dimensionality, and then describes the Markov model and methods for estimation for both situations where state membership is treated as known and situations where state membership is treated as a latent variable. Section 3 illustrates the approaches with one data example, and Section 4 presents a discussion and future research directions.

2. METHODS

The proposed approach begins by grouping sequences with similar patterns of mutations. The use of standard clustering techniques to reduce the dimensionality of large arrays of correlated data was described in several publications; see, for example, Eisen, Spellman, Brown, and Botstein (1998). In the example presented in Section 3, each clonal sequence is first transformed to a binary sequence of indicators for the presence of a mutation at the corresponding location. Hierarchical clustering based on a Manhattan distance and average linkage is used to arrive at initial cluster centers. K -means clustering based on a Euclidean distance is then employed to arrive at

Table 1. Sample Data

ID	Week	Clone	Protease sequence (sites 1–99)			Cluster	
1	0	1	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
		2	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
		3	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
		4	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
		5	PQITLWQRPLVTIK	...	STQLGCTLNF	4	
		6	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
		7	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
		8	PQITLWQRPLVTIK	...	LTQLGCTLNF	4	
2	0	1	PQITLWQRPLVTIK	...	LTQLGCTLNF	1	
		2	PQVTLWQRPLVTIK	...	LTQLGCTLNF	1	
		3	PQITLWQRPLVTIK	...	LTQLGCTLNF	1	
		4	PQITLWQRPLVTIK	...	LTQLGCTLNF	1	
		5	PQITLWQRPLVTIK	...	LTQLGCTLNF	1	
		6	PQITLWQRPLVTIK	...	LTQLGCTLNF	1	
	31	1	1	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			2	PQITLWQRPLATIK	...	LTQLGCTLNF	1
			3	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			4	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			5	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			6	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			7	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
	72	1	1	PQITLWQRPLVTIR	...	LTQLGCTLNF	1
			2	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			3	PQITLWQRPLVTIK	...	LTQLGCTLNF	1
			4	PQISLWQRPLVTIR	...	LTQLGCTLNF	1
			5	PQITLWQRPLVTIR	...	LTQLGCTLNF	1
			6	PQITLWQRPLVTIR	...	LTQLGCTLNF	1
			7	PQITLWQRPLVTIR	...	LTQLGCTLNF	1
8			PQITLWQRPLVTIR	...	LTQLGCTLNF	1	
3	0	1	PQITLWQRPLVTVK	...	LTQLGCTLNF	4	
		2	PQITLWQRPLVAVK	...	LTQLGCTLNF	4	
		3	PQITLWQRPLVTVK	...	LTQLGCTLNF	2	
	13	1	1	PQITLWQRPLVTVK	...	LTQLGCTLNF	4
			2	PQITLWQRPLVAVK	...	LTQLGCTLNF	4
	24	1	PQITLWQRPLVTVK	...	LTQLGCTLNF	4	

cluster assignments for each clonal sequence. Hartigan’s criteria with a cutoff of 50 is used to determine the appropriate number of clusters (Hartigan 1975). Table 1 reports cluster assignments for a sample of data. Clustering can result in multiple cluster assignments for each patient/time combination. An example is provided in Table 1 by patient 3, who, at week 0, has two clones that belong to cluster 4 and one clone that belongs to cluster 2. Other methods for grouping sequences also could be considered. For example, a cluster could comprise sequences that have a specific pattern of known resistance mutations or known resistance mutations could be up-weighted in the distance calculations. We consider an alternative approach to clustering that incorporates knowledge of resistance mutations and we discuss the sensitivity of our results to the choice of clusters in Section 4. By pooling all clonal sequences to create clusters, we may be unduly weighting patients who have more clones or more visits. For this reason, we also consider grouping patients based on the numbers of accessory mutations and mutations known to be associated with intermediate or high level resistance (Shafer 2001). This later approach does not require use of a clustering algorithm. These alternative approaches and the results of applying them to our data setting are discussed in more detail in Section 4.

Using information about how clonal sequences cluster, patients can be assigned to states at each observed time. Let

S_1, S_2, \dots, S_{N_s} represent these states. The following sections describe methods for estimating transition rates between these states and testing hypotheses related to these rates. In Section 2.1, states are assumed to be known at all time points, whereas in Section 2.2, states are treated as latent variables. In both estimation approaches, all transitions are assumed to be observed. For example, suppose a patient is in S_1 at t_1 and S_2 at t_2 . For the purpose of estimating the transition rates, it is assumed that this patient did not change states during the interval (t_1, t_2) and transitioned to S_2 at time t_2 . We discuss the reasonableness of this assumption as well as alternatives to it in Section 4.

2.1 Estimation and Testing Assuming a Markov Process and Known States

Suppose N_s states are observed and the state to which a patient belongs at a given time point is known. In the example provided in Section 3.1, this situation is taken to be the cluster in which the majority of a patient’s clones belongs. Returning to the data described in Table 1, patient 1 is assigned to S_4 , patient 2 is assigned to S_1 at all three observed time points, and patient 3 is assigned to S_4 at all observed time points. Using the notation of Albert (1962), let $\mathbf{P}(t)$ be a matrix of transition probabilities so that the (i, j) element of $\mathbf{P}(t)$ is the probability of transitioning from S_i to S_j in time t . Assume the probability

of transitioning from S_i at time t to S_j at time s depends only on $|s - t|$, that is, the process has stationary transition probabilities. Under this assumption, $\mathbf{P}(t)$ can be written

$$\mathbf{P}(t) = \exp(t\mathbf{Q}) = \sum_{n=0}^{\infty} \frac{t^n \mathbf{Q}^n}{n!}, \quad (1)$$

where \mathbf{Q} is an $N_s \times N_s$ matrix referred to as the infinitesimal generator. Assessment of the assumption of stationary transition probabilities is discussed in Section 3.1.

Let $q(i, j)$ be the (i, j) element of \mathbf{Q} , let K equal the number of patients, let $Z_k(t)$ be the state at time t for patient k , let T be the total amount of follow-up time over all patients, let $N(i, j)$ equal the number of transitions from S_i to S_j , and let $A_T(i)$ equal the total time S_i is occupied over all K patients. Albert (1962) showed that the $q(i, j)$'s can be estimated consistently by

$$\hat{q}(i, j) = \frac{N(i, j)}{A_T(i)}. \quad (2)$$

We now consider tests for the hypothesis that all transitions from one state to another occur at the same rate; that is, tests for departures from $H_0: q(i, j) = q(i', j')$, $i \neq j$ and $i' \neq j'$. Using the large sample properties of $\hat{q}(i, j)$ described in Albert (1962), it is straightforward to show that for $i \neq j$, statistics for testing that each transition rate is equal to an overall rate q_0 are given by

$$T_{ij} = \frac{\hat{q}(i, j) - q_0}{\sqrt{q_0/E A_T(i)}} \stackrel{\text{iid}}{\sim} \mathbf{N}(0, 1), \quad (3)$$

where

$$E A_T(i) = \sum_{k=1}^K \int_0^{t_k} \Pr[Z(t) = S_i] dt. \quad (4)$$

Under this null, $E A_T(i)$ is the i th diagonal element, of the matrix given by

$$\begin{aligned} E A_T &= \sum_{k=1}^K \int_0^{t_k} \left\{ pI \sum_{n=0}^{\infty} \frac{t^n \mathbf{Q}^n}{n!} \right. \\ &\quad \left. + (N_s - 1)^{-1} \left(I - \sum_{n=0}^{\infty} \frac{t^n \mathbf{Q}^n}{n!} \right) (I - pI) \right\} dt \\ &= \sum_{k=1}^K \left\{ pI \mathbf{Q}^{-1} (\exp(t_k \mathbf{Q}) - I) + \frac{1}{N_s - 1} (I - pI) \right. \\ &\quad \left. \times (t_k I - \mathbf{Q}^{-1} \exp(t_k \mathbf{Q}) + \mathbf{Q}^{-1}) \right\}, \end{aligned} \quad (5)$$

where pI is a diagonal matrix with the i th diagonal element equal to the probability of starting in S_i and t_k is the length of time patient k is observed.

When assessing the significance of each T_{ij} , adjustments need to be made for the number of tests $N_s(N_s - 1)$ under consideration. A Bonferroni adjustment provides valid results, but is conservative; a less conservative alternative is provided by adjustment based on the false discovery rate (FDR) (Benjamini and Hochberg 1995). Both corrections are considered in the example provided in Section 3.

2.2 Treating State Membership as Unobservable

This section considers methods for estimating the transition rates when state membership is treated as unknown. The idea behind the approach is that each possible pattern of state assignments (over time) for an individual has a probability associated with it. The estimation procedure, a version of the EM algorithm (Dempster, Laird, and Rubin 1977), iterates between two steps: (1) probabilities are calculated from current estimates of the transition rates and (2) estimated probabilities are used to update estimates of the transition rates.

The expectation of the complete data log-likelihood conditional on the observed cluster memberships \mathbf{u} and the current estimate of \mathbf{Q} , $\hat{\mathbf{Q}}^{(m)}$, is given by

$$\begin{aligned} E(\log L | \hat{\mathbf{Q}}^{(m)}, \mathbf{u}) &= \sum_{k=1}^K \sum_{l \in V_k} \hat{p}_{kl}^{(m)} \left\{ \text{Const.} \right. \\ &\quad \left. + \sum_{i, j \neq i} N_{kl}(i, j) \log[q(i, j)] - \sum_{i, j \neq i} A_{kl}(i) q(i, j) \right\}, \end{aligned} \quad (6)$$

where V_k is the set of all possible sequences of state membership for person k , $N_{kl}(i, j)$ is the number of transitions from S_i to S_j observed in the l th sequence of states for person k , and $A_{kl}(i)$ is the corresponding length of time in S_i . The probability associated with the l th pattern for person k given at least one of the patterns in V_k is observed is estimated by $\hat{p}_{kl}^{(m)}$, a function of the current estimate $\hat{\mathbf{Q}}^{(m)}$. More formally,

$$\hat{p}_{kl}^{(m)} = \Pr(V_k = v_{kl} | \hat{\mathbf{Q}}^{(m)}, \mathbf{u}). \quad (7)$$

Letting $\tilde{N}^{(m)}(i, j) = \sum_k \sum_{l \in V_k} \hat{p}_{kl}^{(m)} N_{kl}(i, j)$ and $\tilde{A}^{(m)}(i) = \sum_k \sum_{l \in V_k} \hat{p}_{kl}^{(m)} A_{kl}(i)$, it is straightforward to show that the conditional expectation given in (6) reaches an extrema at $\hat{q}_{ij}^{(m+1)}$ given by

$$\hat{q}_{ij}^{(m+1)} = \frac{\tilde{N}^{(m)}(i, j)}{\tilde{A}^{(m)}(i)}. \quad (8)$$

If the matrix of second derivatives of the expected log-likelihood conditional on the observed data is negative definite, then the estimate at the final iteration of the EM is at least a local maximum (Dempster et al. 1977).

We reconsider the problem of testing departures from the null hypothesis that all transition rates from one state to another are the same. This test requires estimation of the variance of the parameter estimates. Using the formula of Louis (1982), the diagonal elements of the information matrix, to be evaluated at the maximum likelihood estimator, are given by

$$\begin{aligned} &\sum_{k, l \in V_k} p_{kl} \left(\frac{N_{kl}(i, j)}{q^2(i, j)} \right) \\ &\quad - \sum_{k, l \in V_k} p_{kl} (1 - p_{kl}) \left(\frac{N_{kl}(i, j)}{q(i, j)} - A_{kl}(i) \right)^2 \\ &\quad + \sum_{l \neq l'} p_{kl} p_{kl'} \left(\frac{N_{kl}(i, j)}{q(i, j)} - A_{kl}(i) \right) \\ &\quad \times \left(\frac{N_{kl'}(i, j)}{q(i, j)} - A_{kl'}(i) \right). \end{aligned} \quad (9)$$

In some data settings, the likelihood may have multiple maxima. For example, suppose the data consist of a single patient whose clones are sequenced at two time points, one unit apart. Further suppose that at the first time point, this patient's clones all belong to cluster 1, whereas at the second time point, the clones are divided between clusters 1 and 2. The likelihood is maximized at the boundaries, that is, when the probability of transitioning from 1 to 2 equals 1 or 0. The multiple solutions indicate that the two estimates are equally good at explaining the data. We suspect that this situation will be relatively rare provided that the number of states is small compared to the number of observations and time points; we return to the issue of multiple maxima in the discussion.

3. EXAMPLE

The data used in this section were obtained during three clinical studies of efavirenz (EFV) and are available in the Stanford HIV RT and Protease Sequence Database. In all three studies patients were naive to nonnucleoside reverse transcriptase inhibitors and protease inhibitors. Patients were randomized to receive EFV or placebo in addition to indinavir or zidovudine and lamivudine or double nucleoside reverse transcriptase inhibitors. A complete description of these studies

and more characteristics of the data can be found in Bachelier et al. (2000; 2001). The protease sequences of viral clones from 170 patients are used to arrive at clusters. Sequences are available at two or more time points for 120 of these patients and are used in subsequent analyses. These patients were followed between 10 and 109 weeks, with a median length of follow-up of 54 weeks. The 25th and 75th percentiles of follow-up time are 24 and 70 weeks. Sequences were obtained between 2 and 11 time points with a median of 5 time points. At a given time point, between 1 and 21 clones are observed for a single individual. The median number of clones for a single patient/time combination is 6. Of the 581 unique patient/time combinations, 82% consist of clones that are all members of the same cluster; in 16%, clones are split between two clusters, in 1.7%, clones are split between three clusters, and in one case, clones are split between four clusters.

As discussed in Section 2, *K*-means clustering is used to cluster the sequences of multiple viral clones. A total of seven clusters of clonal sequences are observed. Two of these consist only of sequences from a single patient and are disregarded in subsequent analyses. The proportions of clones with mutations at each of 21 sites associated with PI resistance are illustrated in Figure 1. Clusters are ordered according to the average number

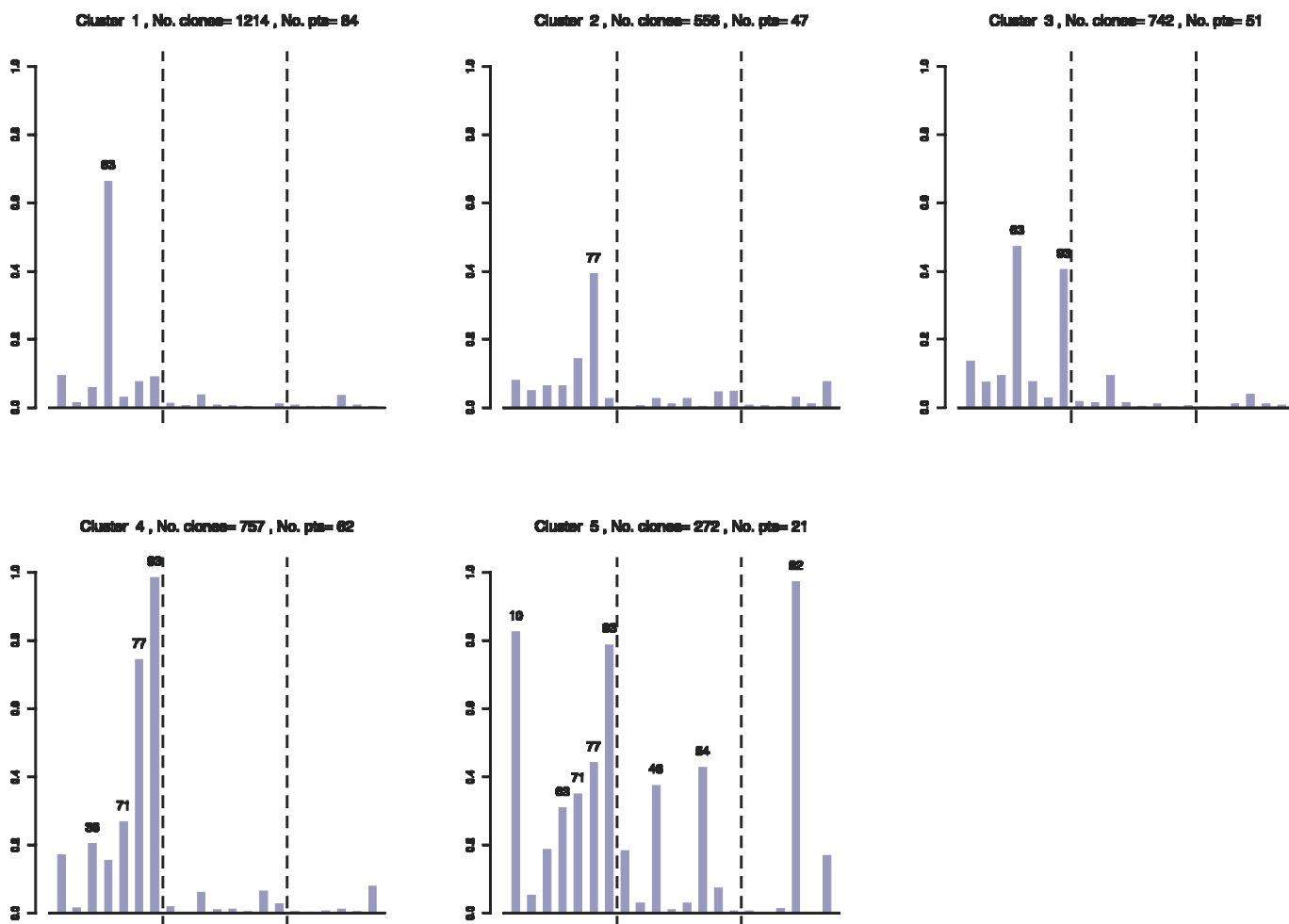


Figure 1. Proportion of Clones With Mutations at Corresponding Sites by Cluster. Sites illustrated include accessory mutations (left), mutations associated with low-level and intermediate resistance (middle), and mutations that confer high-level resistance to at least one PI (right) according to the Stanford PI Resistance Notes, April 2002.

of known resistance mutations observed. Since prior research indicates that a mutation at site 63 alone does not confer resistance (Foulkes and De Gruttola 2002), the cluster with a high prevalence of this mutation alone (cluster 1 in Fig. 1) is treated as the most nearly wild-type cluster. The three-dimensional-protease structure reveals that the sites with high rates of mutation in cluster 5 (10, 93, 46, 54, and 82) tend to be in close proximity to the binding site. The sites that characterize cluster 4 (77 and 93) are just to the left of the first set of defining sites. Finally, site 63, which characterizes the most sensitive cluster, is even farther to the left.

These clusters capture information on general trends in the sequence data. For example, based on Figure 1, it is clear that many clones exhibit mutations only at sites 63 or 77 and no other sites associated with PI resistance. On the other hand, clones with a mutation at site 82 tend to have an array of mutations at other sites. Of the eight clones that are wild type at site 82 in cluster 5, all are mutant at site 46 and wild type at site 54. Seventy percent of the clones in this cluster that are mutant at site 82 are also mutant at sites 46 and/or 54. Clusters 1 and 2 differ in the prevalence of mutations at a few sites not indicated in this figure. Most notably, 62.2% of the clones in cluster 2 exhibit a mutation at site 13, whereas only 4.7% of the clones in cluster 1 are mutant at this site. Higher prevalences of mutations in cluster 2 versus cluster 1 are also observed at sites 37 (49.2% vs. 10.9%) and 72 (37.2% vs. 7.7%). On the other hand, only 4.1% of the clones in cluster 2 are mutant at site 64 versus 38.5% in cluster 1. At baseline, the percentages of patients with the majority of their clones in clusters 1–5 are 38, 17, 24, 21, and 0%, respectively.

3.1 Assuming State Membership Is Known

In this section, a patient’s state membership is defined as the cluster that is most prevalent among the observed clones at a given time point. In 14 cases, two clusters are observed in equal proportions and the patient is assigned the cluster with the fewest average number of mutations. The number of observed transitions is given by

$$N_T = \begin{pmatrix} 142 & 2 & 5 & 1 & 2 \\ 0 & 62 & 1 & 4 & 3 \\ 7 & 1 & 86 & 3 & 2 \\ 0 & 2 & 0 & 81 & 10 \\ 0 & 0 & 0 & 1 & 33 \end{pmatrix}, \quad (10)$$

where the (i, j) element of N_T is the number of transitions from S_i to S_j .

In general, patients tend to remain in the same state from one observed time point to the next over the observation period. The numbers of observed transitions from S_3 to S_1 ($n = 7$) and S_4 to S_5 ($n = 10$) are greater than the number of transitions between any other two states. The estimated transition rates for these changes are .0069 (S_3 to S_1) and .0080 (S_4 to S_5 .) The estimated probabilities of transitioning in 16 weeks is given by

$$p(16) = \begin{pmatrix} .923 & .015 & .036 & .009 & .017 \\ .001 & .879 & .014 & .058 & .049 \\ .096 & .015 & .817 & .041 & .032 \\ .000 & .022 & .000 & .860 & .117 \\ .000 & .000 & .000 & .026 & .973 \end{pmatrix}. \quad (11)$$

The estimated probabilities of remaining in the same state over a 16 week period are all greater than 80%. The probability of staying in S_5 , the most mutant state (97.3%), is 5–10% greater than the probabilities of staying in any one of the other states.

Patients who change states tend to go from more wild-type to more mutant states, with the highest transition probability being 11.7% for the transition from S_4 to S_5 . In other words, patients tend not to revert to more wild-type states. This suggests that patients with viral mutations predominantly at sites 63 and 93 who change states are more likely to move to a relatively sensitive state than to a resistant state characterized by many mutations. The relatively high prevalence of transitions from S_4 to S_5 suggests that patients who develop mutations at 93 and 77 are at high risk of developing mutations at sites 10, 82, and 46 or 54. The latter set of mutations is associated with high levels of multidrug resistance. Thus patients in S_4 are at highest risk for development of high-level resistance to all protease drugs.

The one exception to this pattern is the transition from S_3 to S_1 , which has an estimated probability of 9.6%. We return to a discussion of this transition in the next section; the relatively high estimated probability found here may be an artifact of assigning a state based on the dominant clone rather than considering the role of minority species. Test statistics for testing equality of all transition rates from one state to another are given by

$$\hat{T} = \begin{pmatrix} - & -.80 & .58 & -1.26 & -.80 \\ -1.47 & - & -.69 & 1.66 & .88 \\ \mathbf{3.97} & -.74 & - & .83 & .05 \\ -1.52 & -.25 & -1.52 & - & \mathbf{4.84} \\ -1.21 & -1.21 & -1.21 & -.06 & - \end{pmatrix}. \quad (12)$$

and illustrated in Figure 2(a). Using both the Bonferroni and the FDR adjustment, the transitions from S_3 to S_1 and S_4 to S_5 are significantly different from the overall transition rate.

In general, a likelihood ratio test can be used to assess the appropriateness of the Markov model assumption as described by Kalbfleisch and Lawless (1985). However, in our data setting, the relatively large number of time points at which clones are sequenced and the irregularity of these times across patients results in small cell counts, which means that we cannot rely on asymptotic results. In addition, the power to detect departures from the Markov assumption is likely to be low because of the number of degrees of freedom for the test in this setting. To evaluate the reasonableness of the Markov assumption, we consider the ad hoc alternative of modeling the relationship between an indicator for any change in state from the first to second observed time points and an indicator for a change from the second to third time points. We also considered the relationship between a change in state from the third to fourth time points and from the fourth to fifth time points. Using logistic regression, in both cases we found no dependence between the two indicators, providing evidence against a major departure from the assumption of stationary transition probabilities.

3.2 Treating State Membership as Unobservable

This section treats state membership as an unobserved variable. The approach described in Section 3.1 is used to arrive at starting values for \hat{Q} . Convergence is met after five iterations,

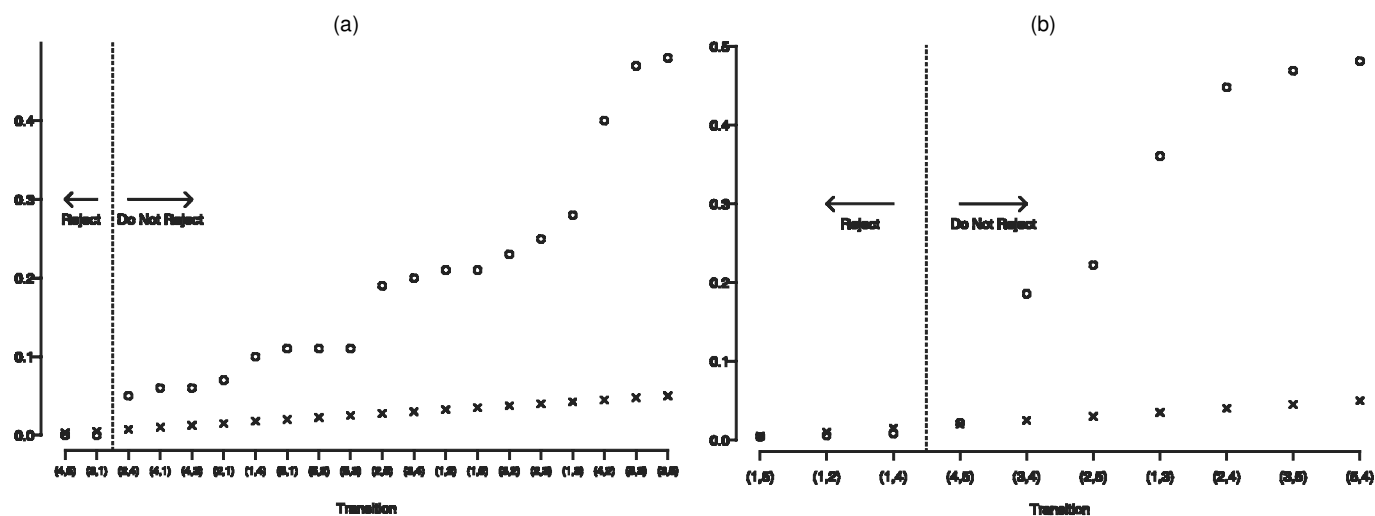


Figure 2. P-Values That Correspond to Tests of Departures From Equal Transition Rates: (a) Treating State as Known; (b) Treating State as Unobservable (o P-value; x FDR cutoff).

using the criterion that the maximum of the absolute difference between parameter estimates was less than 10^{-5} . We also used two other sets of starting values. The first set was based on estimated transition rate when all transitions from one state to another are assumed to occur at the same rate; this assumption led to an estimate of .0019. A second set of starting values, using only patients with known pathways, is discussed at the end of this section. Convergence is reached at the same point for all sets of starting values. Several elements of \mathbf{Q} converged to values that were near 0 ($< 10^{-6}$), which is the boundary of the parameter space. These elements include the (2, 3) element of \mathbf{Q} and all lower diagonal elements except the (4, 5) element. Constraining these elements to be identically 0 and continuing the same iterative procedure until convergence led to a small increase in the likelihood compared to that evaluated at the estimates of \mathbf{Q} that are close but not equal to 0. Therefore, the results based on constraining elements of \mathbf{Q} to be 0 are reported. None of the parameter estimates changed by more than 10^{-6} , so the differences between estimates would be lost in rounding. The information matrix is negative definite at this point, indicating that it is a maximum. In this example, it is possible to reduce the number of parameters by eliminating those that characterize transitions that were never observed in the database, and therefore must be estimated to be 0. Eliminating such parameters is not necessary, however; the final estimates are the same whether or not they are included.

The estimated rate of transitioning from S_4 to S_5 is again approximately .007. However, now the estimated transition rate from S_3 to S_1 is 0. The weighted number of transitions at the final iteration is given by

$$\tilde{\mathbf{N}} = \begin{pmatrix} 145.40 & 1.07 & 3.13 & 1.08 & 1.05 \\ .00 & 63.80 & .00 & 2.02 & 1.04 \\ .00 & .00 & 94.54 & 1.05 & 2.06 \\ .00 & .00 & .00 & 88.46 & 8.48 \\ .00 & .00 & .00 & 1.02 & 33.80 \end{pmatrix}. \quad (13)$$

In almost all cases, the expected number of changes in state is smaller than what is observed [Eq. (10)]. The greatest difference is observed for the transition from S_3 to S_1 . The probabilities associated with patterns that involve this transition

approach 0; hence the weighted number of transitions is 0. This suggests that what appeared to be reversions to more fully sensitive states may have been artifacts of ignoring minority species. The estimated probabilities of transitioning in 16 weeks is given by

$$\mathbf{p}_u(16) = \begin{pmatrix} .950 & .008 & .024 & .009 & .009 \\ .000 & .950 & .000 & .031 & .018 \\ .000 & .000 & .953 & .015 & .031 \\ .000 & .000 & .000 & .902 & .098 \\ .000 & .000 & .000 & .031 & .969 \end{pmatrix}. \quad (14)$$

The test statistics that correspond to tests of departures from the null of equal transition rates are given by

$$\hat{\mathbf{T}}_{EM} = \begin{pmatrix} - & -2.52 & -.36 & -2.40 & -2.62 \\ - & - & - & .13 & -.76 \\ - & - & - & -.89 & .05 \\ - & - & - & - & 2.01 \\ - & - & - & .08 & - \end{pmatrix} \quad (15)$$

and are illustrated in Figure 2(b). Note that test statistics for transitions for which the rate is estimated to be 0 are not defined, because the corresponding information is undefined. Although there is a trend that implies that the rate of transition from S_4 to S_5 is greater than the overall rate, it is not significant at the .05 level after adjusting for multiple testing. The numbers in bold in (15) represent tests that reached the .05 level of significance using the FDR adjustment. In all three cases, the transitions are less likely than expected under the null hypothesis. The test for transition from S_1 to S_5 is also significant using a Bonferroni correction.

In general, the algorithm places the most probability mass on sequences of state membership that involve staying in the same state over time. The posterior probabilities associated with each possible sequence of states can lend insight into which patterns of mutations are of most importance in determining state membership. Consider, for example, one patient whose clones are measured at five time points. The possible sequences of states and corresponding posterior probabilities are

	t1	t2	t3	t4	t5	Posterior Probability
(1)	1	2	4	4	4	.0009
(2)	1	4	4	4	4	.0407
(3)	2	2	4	4	4	.8019
(4)	2	4	4	4	4	.1565

In this example, the patient's clonal sequences are divided between clusters 1 and 2 at the first time point, and between 2 and 4 at the second point; all sequences are in cluster 4 at the remaining three time points. The sequence of state memberships labeled (2), (3), and (4) all involve one change of state. For this patient, the posterior probabilities associated with starting in S_2 are greater than the those associated with starting in S_1 . A change from S_2 to S_4 from time t_2 to t_3 is given more weight than the same change at time t_1 to t_2 . By providing information about the influence of each sequence of state assignments, the posterior probabilities can help identify whether, for example, patients are generally assigned to the more mutant state. In the example just described, more weight was given to the more mutant (S_2) of the two states (S_1 and S_2) at the first time point, but more weight was given to the more wild-type state (S_2) of the two states (S_2 and S_4) at the second time point. In the data example used in this paper, the only trend observed across all patients is that sequences that involve a single state over time, that is, with no transitions, have the greatest associated posterior probability.

Consistency of our estimates is assured when the likelihood has a unique maximum, but, as discussed in Section 2.2, this may not always be the case. One approach to investigation of the likelihood surface is simply to perform grid searches. Another ad hoc approach is to consider very different sets of starting values, as previously, and check that the algorithm converges to the same estimates. We note that if the starting values are consistent estimates, the final estimates will be as well, because in EM iterations, the likelihood is nondecreasing (Dempster et al. 1977). Therefore, it may be valuable to include among the sets of starting values those that might be consistent under certain assumptions. For example, in our setting some sequences are observed with variability (across multiple clones, for example), but others are not. Seventy-five patients have state sequences for which the order of transition is observed without variability, although for 10 of them, there was variability in the possible times of transition. As mentioned previously, data from these patients were also used to arrive at initial estimates of \mathbf{Q} . If inference based on the patients whose order of transitions are known with certainty provide consistent (although inefficient) estimates of \mathbf{Q} , then the final estimates are also consistent.

4. DISCUSSION

This paper proposes an approach to characterizing the pathways taken by the HIV-1 genome from the wild-type genotype to various resistant genotypes. As mentioned in the Introduction, these pathways may correspond to actual viral evolution, that is, development of new mutations under selective drug pressure, or simply detection of a preexisting resistant quasi-species that were not detectable until after the dominant drug-sensitive species were suppressed. Although these mechanisms are not distinguishable from the available data, the methods we develop apply equally well for both. Furthermore, the results

are clinically useful in either setting, because they permit investigation of genetic and other factors that predispose a patient to development of resistant strains as well as provide estimates of the rates of development of these strains. For example, although patients in state 4 may not have a highly resistant phenotype, such patients would be poor candidates for protease inhibitor therapy because they are at high risk of becoming highly resistant to these drugs. Even worse from the patient perspective is if they did so, they would be at high risk of developing resistance to the other therapies in their regimens, because effective treatment of HIV infection generally requires active drugs from at least two drug classes.

Were covariates such as baseline viral load available in the dataset we consider, our methods could be modified to permit analysis of the effect of such covariates. Such analyses would also be helpful for selecting appropriate treatment regimens. Consider, for example, the case in which we are interested in determining whether viral load above or below 400 copies influences the transition rates between states. If we assume a multiplicative effect, we have $q_{ij}(Z=1) = \alpha[q_{ij}(Z=0)]$, where Z is an indicator for viral load greater than 400 copies. The likelihood can now be described by replacing q_{ij} by $q_{ij}[I_z + (1 - I_z)\alpha]$, where I_k is equal to Z for patient k . Estimation of q_{ij} and α can be arrived at by setting the score equations equal to 0 and using a Gauss-Seidel iterative procedure (Thisted 1998).

We note that even in population sequencing where mixtures of amino acids may be observed at specific sites, our approach may prove to be useful. Suppose there were mixtures of two amino acids at two sites in a given sequence; we could infer the existence of four possible sequences (without mixtures) from this sequence. Of course, not all of these four sequences are necessarily present in the patient's plasma or even biologically possible, but this concern does not prevent the application of our approach. Because our methods down-weight pathways that are not supported by the totality of the data, they may perform better than methods that either ignore mixtures or treat them in an ad hoc way.

The estimation procedure described in Section 2 assumes that all transitions occur at the time the new state is observed. This allows us to treat the time in S_i , given by $A_T(i)$, as observed data. We consider that this assumption provides reasonable approximations to reality because the intervals were fairly short, in the sense that rates of transition within intervals were fairly low. Were the intervals longer, however, alternative approaches might be required. One alternative is to assume a parametric distribution for time in state and then allow the transition to occur anywhere in the interval. The most straightforward distribution is exponential, which corresponds to the likelihood described in Albert (1962). Assumptions other than constant hazard require a different formulation for the likelihood. In our example, we do not expect results to change much using this approach because of the shortness of the intervals between observed sequences. To investigate the potential impact of ignoring interval censoring of the sequence data, we performed an additional analysis using the midpoint between time intervals as the time at which transitions occur. The same transitions were identified as being significantly different from the others as in Section 3.1. The estimated transition rate from S_5 to S_4

decreased by 25%, while all other nonzero estimates increased, with an average increase of 4%.

Although it is possible that additional, undetected, transitions may have occurred between times of observations, we do not think that the existence of such transitions would have a major impact on our findings. The effect of such transitions within intervals would tend to reduce the observed number of transitions directly from the most nearly wild type to the most resistant state. The reason for this is that such transitions would likely be to intermediate states. One major finding, that transitions that “jump” from sensitive to highly resistant states are relatively rare, would only be strengthened by the presence of additional unobserved transitions. One additional feature of the data requires consideration. As mentioned in the Section 1, it is not possible to sequence virus from patients whose viral burden is below a certain threshold. Even when interval-censored approaches are used, it may be preferable to assume, as we do in this paper, that viruses that remain below the level of detection do not change state.

As mentioned in Section 2.2, the likelihood surface is not constrained to be convex, so multiple maxima may arise. It is easy to check that the convergence point of the algorithm is at least a local maximum by checking that the information matrix is negative definite at that point. In practice, different starting values should be considered to ensure that they all lead to convergence at the same point. The starting values considered in the example all led to identical estimates.

The high dimension of viral genetics data led us to consider dimension-reduction techniques. In our example, we used cluster analysis to define states between which patients transition over time. Note that states could also be defined by specific patterns of mutations, rather than by clusters. These patterns can be defined based on preexisting knowledge about resistant variants. Approaches such as those described in Sevin et al. (2000) may be useful for such purposes. To assess the sensitivity of our findings to the initial clustering, we considered two alternative approaches to creating clusters. First, we based our clusters on only the 21 sites known to be associated with PI resistance, instead of using all 99 sites as in our primary analyses. This approach is equivalent to weighting known sites by a factor of 1 and all other sites by a factor of 0 in the distance calculations. Second, we grouped patients according to the following rule based on a classification system for mutations described on the Stanford website (Shafer 2001): (1) no mutations known to be associated with PI resistance (with the exception of 63); (2) one or more accessory mutations and no mutations associated with intermediate or high-level resistance; (3) one or more mutations associated with intermediate or high-level resistance. Although these approaches to grouping patients are very different from each other and from the methods described, the central findings of our analyses were qualitatively similar across them. These findings include low rates of reversion to wild type, a tendency for transitions to occur only among adjacent states, and high risk of transition from an intermediate to a fully resistant state.

Finally, we implemented two additional approaches that incorporate information on all possible state paths given observed states: (1) equal weight for each possible state path for each in-

dividual; (2) state paths weighted by the product across time points of the prevalences of each cluster. The rates of the transitions from one state to another have approximately the same rank order for these approaches and the one treating cluster as known.

Our approach for estimating and testing the rates of transition between states accommodates the detected presence of quasi-species and, therefore, may provide a more realistic picture of progression to resistance. The usefulness of this feature is demonstrated by the example where the transition rate from S_3 to S_1 appeared to be significantly greater than the overall rate until the minority strains were considered. Estimation of the frequency of viral genotypic reversion to more sensitive states is highly relevant in AIDS clinical research, because such reversion increases the treatment options for drug-experienced patients. Our findings, however, imply that the appearance of reversions based only on consideration of majority species may be misleading; valid investigation of reversion must take into account the existence of minority species.

[Received May 2003. Revised July 2003.]

REFERENCES

- Albert, A. (1962), “Estimating the Infinitesimal Generator as a Continuous Time, Finite State Markov Process,” *Annals of Mathematical Statistics*, 33, 727–753.
- Bacheler, L., Anton, E., Kudish, P., Baker, D., Bunville, J., Krakowski, K., Bolling, L., Aujay, M., Wang, X., Ellis, D., Becker, M., Lasut, A., George, H., Spalding, D., Hollis, G., and Abremski, K. (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., D’Aquila, R., Wallace, L., Logue, K., Cordova, B., Hertogs, K., Larder, B., Buckery, R., Baker, D., Gallagher, K., Scarnati, H., Tritch, R., and Rizzo, C. (2001), “Genotypic Correlates of Phenotypic Resistance to Efavirenz in Virus Isolates From Patients Failing Non-nucleoside Reverse Transcriptase Inhibitor Therapy,” *Journal of Virology*, 75, 4999–5008.
- Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate: A Practical, and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300.
- 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, Ser. B*, 39, 1–22.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1997), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, New York; Cambridge, U.K.: Cambridge University Press.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), “Cluster Analysis, and Display of Genome-Wide Expression Patterns,” *Proceedings of the National Academy of Sciences*, 95, 14,863–14,868.
- Foulkes, A. S., and De Gruttola, V. (2002), “Characterizing the Relationship Between HIV-1 Genotype and Phenotype: Prediction Based Classification,” *Biometrics*, 58, 145–156.
- Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.
- Kalbfleisch, J., and Lawless, J. (1985), “The Analysis of Panel Data Under a Markov Assumption,” *Journal of the American Statistical Association*, 80, 863–871.
- Segal, M. R., Cummings, M. P., and Hubbard, A. E. (2001), “Relating Amino Acid Sequence to Phenotype: Analysis of Peptide-Binding Data,” *Biometrics*, 57, 632–643.
- Sevin, A., De Gruttola, V., Nijhuis, M., Schapiro, J. M., Foulkes, A. S., Para, M. F., and Boucher, C. A. B. (2000), “Evaluating the Relationship Between Drug Susceptibility Phenotype and Genotype Among HIV From Patients Treated With Protease Inhibitors,” *The Journal of Infectious Diseases*, 182, 59–67.
- Shafer, R. (2001), The Stanford HIV RT and Protease Sequence Database. Available at <http://hivdb.stanford.edu/hiv/>.
- Thisted, R. A. (1998), *Elements of Statistical Computing*, London; New York: Chapman & Hall.