

Statistical Applications in Genetics and Molecular Biology

Volume 3, Issue 1

2004

Article 38

Validation and Discovery in Markov Models of Genetics Data

Victor De Gruttola*

Andrea S. Foulkes†

*Harvard School of Public Health, victor@sdac.harvard.edu

†University of MA, foulkes@schoolph.umass.edu

Copyright ©2004 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress, which has been given certain exclusive rights by the author. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). <http://www.bepress.com/sagmb>

Validation and Discovery in Markov Models of Genetics Data*

Victor De Gruttola and Andrea S. Foulkes

Abstract

Markov models provide a natural framework for modeling cellular and molecular level changes over time. Kalbfleisch and Lawless propose using a Chi-squared statistic for assessing the appropriateness of assuming a first-order, homogeneous Markov process. While this statistic provides a global test of the Markov assumption, it does not permit identification of individual departures. We consider two approaches for discovering specific departures from the Markov assumption. First, we propose a diagnostic that tests whether the number of observed transitions out of a given state at a given time point is different than expected. Second, we construct statistics based on the number of observations in each state at each time point. In both cases, we construct multiple correlated statistics and testing is achieved through simulations. These approaches are applied to HIV genetics sequences measured over time.

KEYWORDS: Markov process, stationarity, first-order, genetics, HIV-1, biomarkers, repeated measures

*Authors contributed equally to this work. Support for this research was provided by NIH/NIAID grant R01AI51164.

1 Introduction

Markov models provide a natural framework for modeling cellular and molecular level changes over time. In a recent manuscript [Foulkes and DeGruttola, 2003], Foulkes and DeGruttola describe methods for characterizing viral genetic changes assuming a continuous time, first-order stationary Markov process as described by [Albert, 1962]. This method could be applied similarly to settings in which a large number of biomarkers are measured repeatedly and interest lies in characterizing the progression of these markers over time. Here we propose diagnostics for assessing the appropriateness of the Markov model; these diagnostics can be used to discover specific departures from model assumptions and are applicable in settings in which the observed time points are irregularly spaced.

Kalbfleisch and Lawless [Kalbfleisch and Lawless, 1985] suggest assessing the fit of a Markov model using Person's chi-squared statistic, given in Equation 1.1 where the n_{ijl} denote the observed transition frequencies, i.e. the number of observations that transition from state i at time t_{l-1} to state j at time t_l , the $e_{ijl} = n_{i.l}\hat{p}_{ijl}$ are the corresponding expected frequencies, m is the number of time intervals, k is the number of states, and b is the number of parameters. Note that $n_{i.l} = \sum_{j=1}^k n_{ijl}$ and \hat{p}_{ijl} is the estimated probability of transitioning from state i to state j in the time interval $t_l - t_{l-1}$. While this statistic gives us a global test of the first-order Markov assumption and stationarity (by summing over all time points and start and end states) it does not permit identification of individual departures from this assumption, i.e. specific transitions and time intervals in which these assumptions are violated.

$$\chi_{(mk(k-1)-b)}^2 = \sum_{l=1}^m \sum_{i,j=1}^k \frac{(n_{ijl} - e_{ijl})^2}{e_{ijl}} \quad (1.1)$$

In this manuscript, we consider the setting in which patients infected with the Human Immunodeficiency Virus (HIV) are treated with antiretroviral therapy and interest lies in characterizing the development of HIV mutations that confer resistance to such therapy. Successful treatment causes level of virus in blood to become undetectable, but treatment failure causes viral rebound. We consider data from studies in which virus is sequenced repeatedly over time during periods when there is sufficient virus in blood. Viruses are assumed to undergo transitions through underlying biological states characterized by the appearance (and possible disappearance) of specific genetic mutations.

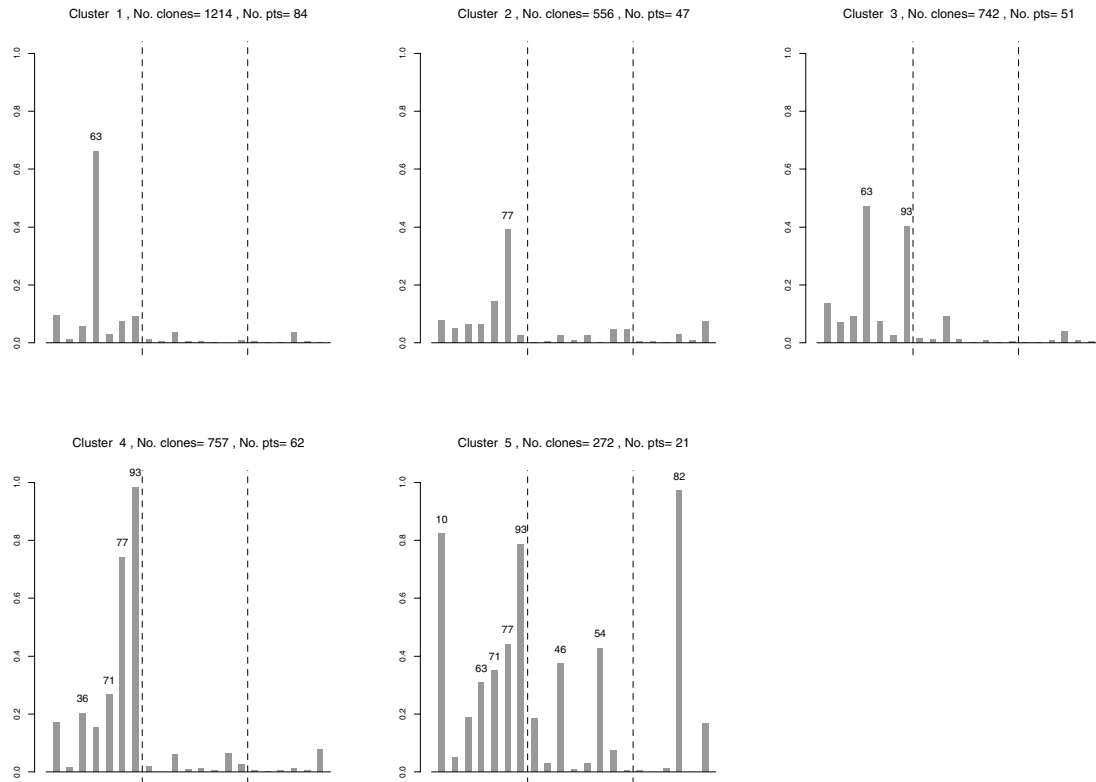
The data motivating our research were obtained during 3 clinical studies of Efavirenz (EFV) and are available in the Stanford HIV RT and Protease Sequence Database. We refer the reader to [Bachelier et al., 2001] and [Bachelier et al., 2000] for a complete description of the data under investigation. Briefly, these data include viral protease sequences from 170 patients, of whom 120 patients have sequences available at 2 or more time points. 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 at 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.

In [Foulkes and DeGruttola, 2003], K-means clustering is used on these data to cluster the sequences of multiple viral clones. Of the 581 unique patient/time combinations, 82% consist of clones which are all members of the same cluster. In 16%, clones are split between 2 clusters, in 1.7% clones are split between 3 clusters and in one case, clones are split between 4 clusters. A total of 7 clusters of clonal sequences are observed. Two of these consist only of sequences from a single patient and are disregarded in subsequent analysis. 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 of known resistance mutations observed. Since prior research indicates that a mutation at site 63 alone does not confer resistance [Foulkes and DeGruttola, 2002], the cluster with a high prevalence of this mutation alone (cluster 1 in Figure 1) is treated as the most nearly wildtype cluster. In this paper we assign patients to biological states according to the most prevalent cluster assignment of their clonal sequences at a given time point. We assume patients transition between these states over time and focus on identifying transitions between states that violate the assumptions of a first-order homogeneous Markov model.

First-order Markov models assume that specific transitions from one state to another do not depend on the history of the process. In our setting, the first-order assumption implies that the probability of moving from one genetic state to another depends only on the genetic make-up of an individual's current viral population, and not the viral genotypes at previous time points. The further assumption of stationarity implies that the time at which these transitions take place does not impact the rate of transition. Our methods investigate these assumptions to determine whether (1) specific transitions from one state to another do, in fact, depend on history and (2) specific time intervals (e.g. the time from the first to the second measurement or from the second to the third measurement) during which the rates of transitioning differ from the corresponding rates in other intervals. In other words, we are interested in identifying specific instances in which either the first-order Markov assumption or the stationarity of the process is violated. In the following section we propose a novel non-parametric approach that allows for both validation and discovery. We note that some of our diagnostics can be formulated as multiple tests of formal null hypotheses related to the Markov process, while others cannot. We return to this point in the discussion.

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 conferring high-level resistance to at least one PI (right) according to the Stanford PI Resistance Notes, April 2002.

2 Discovering departures from the Markov assumption

We discuss two approaches to discovering departures from the Markov assumption. In the first case we consider the use of a diagnostic that tests whether the number of observed transitions out of a given state at a given time point is different than expected. This approach is a straightforward extension of the method described by [Kalbfleisch and Lawless, 1985] in that it similarly involves constructing statistics based on the observed and expected transition counts; however, the statistics we generate are only asymptotically independent and χ^2 -distributed as the number of time intervals gets large. By summing over the end states in this ap-

proach, we reduce the sparseness of the data. The second approach we consider involves constructing multiple correlated statistics based on the number of observations in each state at each time point. Benjamini and Yekutieli demonstrate how to control for the false discovery rate in the context of positive dependency of the test statistics [Benjamini and Yekutieli, 2001]. Following the suggestion of Westfall and Young (Chapter 6), we propose tests that are based on simulated data sets generated from the estimated Markov model [Westfall and Young, 1993]. This approach does not require specification of the marginal and joint distributions of the statistics. The approach described in [Benjamini and Yekutieli, 2001], however may be appropriate for interpreting significance of multiple p-values.

2.1 Testing based on transitions over time intervals

First, let the l th element of the sum in Equation 1.1 be given by $V_{ijl} = (n_{ijl} - e_{ijl})^2/e_{ijl}$. It is straightforward to show that the sum of the V_{ijl} over all j where j is the state *to* which a person is transitioning is asymptotically chi-squared with $C - 1$ degrees of freedom where C is the number of states. We denote this sum by $V_{i.l}$. Furthermore, $V_{i.l}$ is independent of $V_{i'.l}$ for $i \neq i'$. That is, at a given time point, the number of transitions from state i to any other state is independent of the number of transitions from state i' to any other state for $i \neq i'$. However, $V_{i.l}$ is not independent of $V_{i.l'}$ for $l \neq l'$. In general, if the number of time intervals is M , then we have $C * M$ correlated statistics. These are asymptotically independent and χ^2 as the number of time points, M gets large. Let $V_{(1)}, \dots, V_{(C * M)}$ represent the corresponding order statistics.

We begin by calculating the expected number of transitions between biological states under the assumption of a first-order homogeneous Markov process. Here we use the estimated transition rates between states based on the methods described in [Foulkes and DeGruttola, 2003], the observed starting states at each time point and the observed time intervals for each individual. Using these expected counts (denoted e_{ijl}) and the observed counts, we then calculate $V_{i.l}$ for all starting states i and time intervals l .

In order to test formally hypotheses that the first-order, stationary Markov model holds for all transitions and all time points, without relying on large sample theory we propose the following procedure. Specifically, we aim to assess whether (1) $p_{ij}(t, s) = Pr[Z(s) = j | Z(t) = i]$ where $Z(s)$ and $Z(t)$ are the states at times s and t respectively and $p_{ij}(t, s)$ is the probability of moving from state i to j from time t to time s and (2) $P(t, s) = P(s - t)$, that is the probabilities of moving between states only depends on the length of time passed, not the absolute time. First, we generate B datasets under the assumption of a first-order homogeneous Markov process. Here we again rely on the estimated transition rates, the observed starting states *at each time point* and the observed lengths of time between observations. For example, if an observation starts in state 1 at week 0 and is next observed at week 16, we draw from a multinomial distribution with probabilities (0.923, 0.015, 0.036, 0.009, 0.017)

to select the next state for this individual. Note that these estimated probabilities were reported in [Foulkes and DeGruttola, 2003] and that the length of time between measurements may be different across individuals. In selecting the state for the third time point, we set the starting state equal to the observed state at the second time point. For a simulation b , the observed number of transitions between states i and j in time interval l is denoted $\tilde{n}_{ijl}^{(b)}$.

For each $b = 1, \dots, B$, we then calculate $\tilde{V}_{i.l}^{(b)}$ given in Equation 2.1. The corresponding order statistics for each simulation, b are given by $\tilde{V}_{(1)}^{(b)}, \dots, \tilde{V}_{(C*M)}^{(b)}$ and the r th expected order statistic is the average $E_{(r)} = \frac{1}{B} \sum_{b=1}^B \tilde{V}_{(r)}^{(b)}$. The difference between the r th order statistic for each simulation and the corresponding expected order statistic is denoted $\tilde{D}_r^{(b)} = \tilde{V}_{(r)}^{(b)} - E_{(r)}$. Finally, the ordered differences are denoted $\tilde{D}_{(1)}^{(b)}, \dots, \tilde{D}_{(C*M)}^{(b)}$. Similarly the observed differences are given by $D_r = V_{(r)} - E_{(r)}$ (based on the observed data, $V_{(r)}$) and the corresponding order statistics are denoted $D_{(1)}, \dots, D_{(C*M)}$. Significance is assessed by comparing the $D_{(r)}$ to the 95% of the distribution of $\tilde{D}_{(r)}^{(b)}$ over all $b = 1, \dots, B$. For example, if $r = C * M$ is the largest order statistic, we compare the greatest observed difference $D_{(C*M)}$ to the distribution, $\tilde{D}_{(C*M)}^{(1)}, \dots, \tilde{D}_{(C*M)}^{(B)}$

$$\tilde{V}_{i.l}^{(b)} = \sum_{j=1}^C \frac{(\tilde{n}_{ijl}^{(b)} - e_{ijl})^2}{e_{ijl}} \quad (2.1)$$

This algorithm is based on the method described in [DiRienzo and DeGruttola, 2003] and formalized in the following step-by-step procedure:

1. Using the estimated transition rates, the observed starting states at each time point and the observed time intervals, generate a data set assuming a first-order, homogeneous Markov process and record the number of transitions, \tilde{n}_{ijl}
2. Repeat Step 1, B times where B is a large number to obtain $\tilde{n}_{ijl}^{(1)}, \dots, \tilde{n}_{ijl}^{(B)}$.
3. Calculate $\tilde{V}_{i.l}^{(b)} = \sum_{j=1}^C \frac{(\tilde{n}_{ijl}^{(b)} - e_{ijl})^2}{e_{ijl}}$ for $b = 1, \dots, B$.
4. For each simulation, b order the $\tilde{V}_{i.l}^{(b)}$ and call these ordered statistics $\tilde{V}_{(1)}^{(b)}, \dots, \tilde{V}_{(C*M)}^{(b)}$
5. Calculate the r th expected order statistic, $E_{(r)} = \frac{1}{B} \sum_{b=1}^B \tilde{V}_{(r)}^{(b)}$ for $r = 1, \dots, C * M$ and let $\tilde{D}_r^{(b)} = \tilde{V}_{(r)}^{(b)} - E_{(r)}$. Let the order statistics corresponding to these differences be denoted $\tilde{D}_{(1)}^{(b)}, \dots, \tilde{D}_{(C*M)}^{(b)}$ and record the 95th percentile of the distribution of $\tilde{D}_{(r)}^{(b)}$ over all $b = 1, \dots, B$.
6. Let $D_r = V_{(r)} - E_{(r)}$ and $D_{(1)}, \dots, D_{(C*M)}$ be the corresponding order statistics. Compare $D_{(r)}$ to the percentile just calculated and reject the global null

hypothesis (i.e. data are generated from a stationary Markov process) if any $D_{(r)}$ exceed this percentile.

In this section we describe an approach that sums over end states. We take this approach for two reasons: First, it results in test statistics that are more nearly uncorrelated and χ^2 -distributed; and second, aggregating over end states improves power to detect specific departures from the null. Alternatively, we could have considered all elements of the sum in Equation 1.1 and performed analysis on these $M \times C \times C$ correlated statistics. In fact the procedure we describe does not require that test statistics be uncorrelated, but results may be more easily interpreted when they are nearly so.

In addition to providing a valid test of the global null hypothesis that the data are generated from a stationary Markov process, our procedure can be used to generate plots (shown in Section 3) that provide a basis for identifying outliers, e.g. time-points and transitions that likely contribute in important ways to departure from the Markov model. Our procedure can also be used to test formally, multiple null hypotheses that transitions out of each state during each time interval occur with a rate predicted by the Markov model. We note that our approach does not have the property of "subset pivotality" [Westfall and Young, 1993]. In our example, the limiting distribution of the vector of test statistics is the distribution of the vector of the appropriate order statistic; this distribution is not the same under the complete null hypothesis as under any subset of null hypotheses regarding individual time points and sets of transitions (partial null hypotheses). This condition is required for subset pivotality. As a result, a single application of our procedure does not provide a valid p-value associated with each individual test statistic. In the terminology of Westfall and Young, our resampling method controls only the FWEC (family-wise Type I error rates for the complete null hypothesis) and but not the FWE (family-wise Type I error rates) in the strong sense [Westfall and Young, 1993, Pollard and van der Laan, 2002].

Despite this limitation, our approach still is useful for identifying outliers that merit further investigation. For settings in which precise inference regarding outliers is desirable, we can use a step-down version of our procedure. If the complete null hypothesis is rejected at a given alpha level, the most extreme outlier is identified, and the method is reapplied to data in which transitions corresponding to this outlier are removed. For example, if the extreme outlier corresponds to transitions out of state 1 in the third time interval, then the statistic corresponding to the sum of transitions from state 1 to any other state during this time interval is removed from the next stage of analysis. This process can be repeated until all test statistics lie within the confidence bounds. The null hypotheses corresponding to test statistics that are not included within these bounds are rejected at the specified level.

2.2 Testing based on counts in each state at each time point

Another approach to constructing correlated statistics is based on the joint distribution of the numbers of observations in each state at each time point. The general form of these statistics is given by $V_{il}^* = (n_{il}^* - e_{il}^*)^2 / e_{il}^*$ where n_{il}^* and e_{il}^* are respectively the observed and expected numbers of observations in state i at time point l . Again we have $N = C * M$ dependent statistics.

The joint probability density function (pdf) of V_1, \dots, V_N is given by $f_V(\mathbf{v})$ and $F_V(\mathbf{v})$ is the corresponding cumulative density function (cdf). If we assume $F_V(vJ_k, \infty J_{N-k}) = F_V(\infty, vJ_k, \infty J_{N-k-1}) = \dots = F_V(\infty J_{N-k}, vJ_k)$, the distribution of the j th order statistic is given in Equation 2.2 [Casella and Berger, 1990]. Note that multiplying the integral by $\binom{N}{k}$ accounts for the number of ways of arranging the statistics. Rather than develop an analytical form for the distribution of order statistics in analytic form, which would be complex, we assess significance using a procedure similar to the one described in Section 2.1.

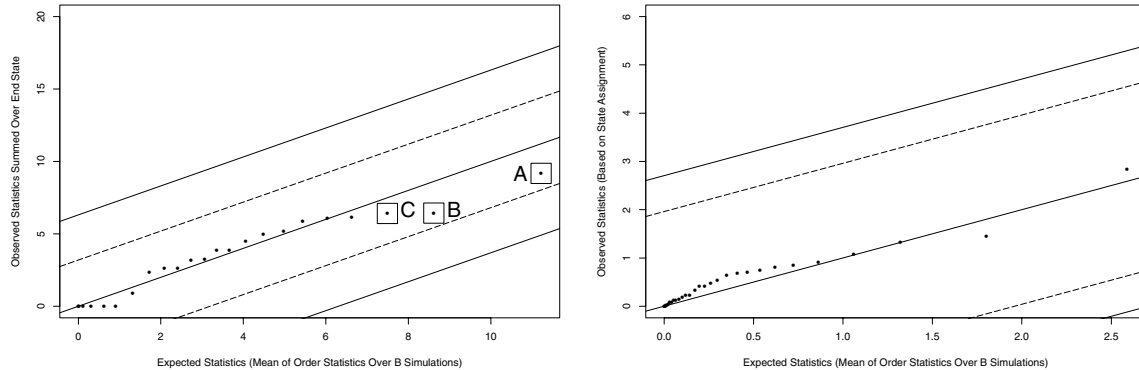
$$F_{V_j}(v) = \sum_{k=j}^N \binom{N}{k} \int_{(0,0,\dots,0)}^{(vJ_k, \infty J_{N-k})} f_V(u_1, u_2, \dots, u_N) du_1 du_2, \dots, du_N \quad (2.2)$$

The expected counts, given by e_{ij}^* are calculated as follows. Using the observed starting states at the the first time point, we simulate the Markov process using the estimated transition rates and the observed lengths of time between measurements. Note that this procedure is different from the one described above since in the latter, the expected counts are calculated using the observed starting states at each time point. The previous approach is consistent with the method described by [Kalbfleisch and Lawless, 1985], and has an interpretation related to the testing of individual null hypotheses. Here we only use the observed state at the first time point and simulate the entire process; we consider only tests of the global null hypothesis. The expected number of observations in each state at each time point is defined as the average over B simulated datasets. Using the same simulated datasets, we can test the significance of the order statistics corresponding to the V_{il}^* . This is achieved using the step-by-step procedure given in Section 2.1. Here the $\tilde{n}_{ijl}^{(b)}$ are replaced by $\tilde{n}_{il}^{*(b)}$ and $\tilde{V}_{(i,l)}^{(b)}$ are replaced by $\tilde{V}_{(il)}^{*(b)} = (n_{il}^{*(b)} - e_{il}^*)^2 / e_{il}^*$.

3 Example: Characterizing viral genetic changes over time

As described above, the data motivating are research include viral protease sequences on $N = 120$ patients measured on at least 2 occasions. We restrict our application to the first 6 observed time intervals (7 time points) for each individual

Figure 2: Observed and Expected Statistics



(a) Statistics based on transitions summed over end states. Point in box labeled **A** corresponds to transitions from state 1 during the 4th time interval, point in box labeled **B** corresponds to transitions from state 3 during the 1st time interval and point in box labeled **C** corresponds to transitions from state 1 during the 3rd time interval.

(b) Statistics based on counts in each state at each time point.

to eliminate intervals with sparse data. $B = 100$ datasets are simulated under the assumption of a first-order homogeneous Markov model, given by $P(t) = \exp(Qt)$ where Q is the infinitesimal generator described in [Albert, 1962]. Since the expected cell counts are generally under 5, we calculate exact p-values using Proc StatXact and invert these to arrive at the corresponding chi-squared statistics. We choose to invert the p-values based on a χ^2 distribution since the statistics are asymptotically χ^2 . Since inference is based on simulations, correctly specifying this distribution is not required.

Figures 2(a) and 2(b) illustrate the results of our analysis. Each dot represents a specific time interval and starting state (in Figure 2(a)) or time point and state (in Figure 2(b)). Since there are $M = 6$ time intervals and $C = 5$ starting states, a total of $C * M = 30$ dots are plotted in Figure 2(a). Similarly, there are $M = 6$ time points (since the observed and expected counts are identical at the first time point) and $C = 5$ states of interest, yielding 30 dots in Figure 2(b). The lines in these Figure represent the 90% confidence bands for the 30th (solid) and 29th (dotted) order statistics. A single dot outside of the solid lines or two dots outside of the dotted line suggest a departure from the Markov assumption. In assessing significance based on the $(C * M - 1) = 29$ th order statistic, dividing alpha by 2 to adjust for testing 2 hypotheses may be appropriate.

Using the first approach, there is a suggestion of underdispersion (represented by points **A**, **B** and **C**.) In these instances, the number of observed transitions are slightly

smaller than expected; however, this difference is not significantly different from 0 at the 0.10 level. On the other hand, the second approach suggests the observed statistics are slightly greater than expected. Again this finding is not significant. These results suggest that it is reasonable to assume the data from the first 6 time intervals arise from a first-order, stationary Markov process.

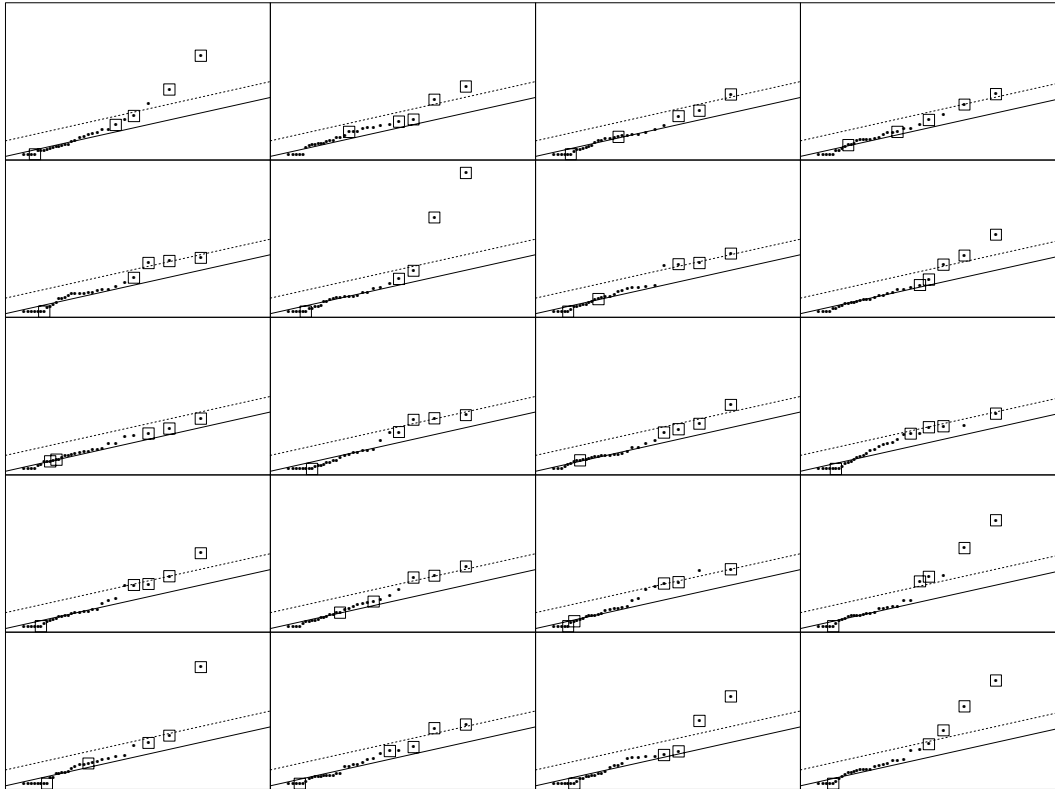
4 Sensitivity analysis

In order to assess the sensitivity of the procedure outlined above for detecting departures from the model assumptions, we generate 20 data sets under the assumption that viral populations progress more quickly to the most resistant state after reaching a threshold of drug exposure. Specifically, we let the transition rates into state 5 be 10 times greater during the last of 6 time intervals. The transition rates in the first 5 time intervals are set equal the estimated rates in the observed data. Each simulated dataset has 120 individuals and the distribution of starting states is equal to the observed distribution. Finally, each measurement is set to be 16 weeks apart. This process is referred to as a time-dependent process. Note that although our transition approach is based on summing over states to which a patient transitions (end states) for all subjects with a given starting state, we create outliers by changing the rates of transition to an end state at the final time point, not by changing rates of transition from a starting state. Thus we investigate the ability of our transition approach to detect outliers of a more general type than that which the transition approach would have the most power to detect. We return to this point below.

Plots of the expected (x-axis) versus observed (y-axis) statistics for each of the 20 simulated data sets are given in Figures 3 and 4. In general, the observed statistics are greater than expected. Figure 3 illustrates the transition approach; and Figure 4, the "counts at each time point" approach. In Figure 3, the squares denote statistics corresponding to the last time interval while in Figure 4 the squares represent statistics corresponding to the most resistant state and the diamonds indicate this state at the last time point. As we expect, the more extreme order statistics correspond to observations in the last time interval (in (a)) and the most resistant state (in (b)). The dotted lines in these figures indicate the 90% confidence band. Using the transitions approach, we reject in 18/20 simulated datasets. We reject in all (20/20) cases using the counts at each time point approach. The overall χ^2 statistic proposed by [Kalbfleisch and Lawless, 1985] has 100 degrees of freedom in this example and is significant at the 0.10 level in 19/20 of these simulations. Due to small expected cell counts, we calculated exact p-values and inverted them to arrive at this statistic.

Using results from the simulations for the transitions approach, we can also show that if the outlier arises from a change in rates of transitioning from a particular starting state to different end states at a particular time, the transition approach may be more likely to identify this departure from model assumptions than the Kalbfleisch Lawless approach. We can easily determine how large the contribution of this outlier

Figure 3: Data arising from a time-dependent process: Transition approach



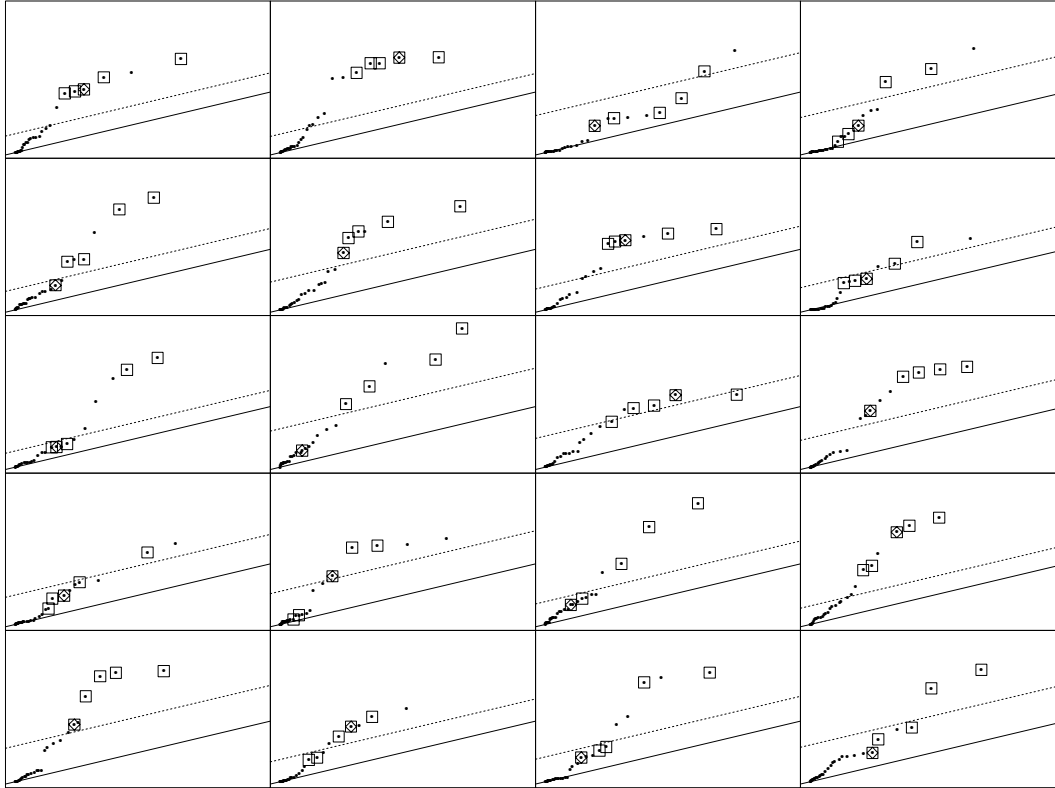
The x-axes represent the expected order statistics generated by repeatedly (100 times) sampling 30 statistics from a χ_4^2 distribution and averaging the resulting order statistics. The data represented on the y-axes are assumed to arise from a time-dependent process. The upper line in each figure corresponds to the 90% cutoff.

to the Kalbfleish and Lawless test statistic needs to be in order that this statistic will reject the null at the 0.1 level. Since this test statistic is distributed as a χ^2 statistic with 100 degrees of freedom, the magnitude of this contribution is just the difference between the 90th percentile of a χ_{100}^2 and the median of a χ_{96}^2 ; this difference equals 23.16. By contrast, the transition approach would lead to rejection of the null at the 0.1 level, if the most extreme of the 4-degrees-of-freedom test statistics were only 16.

5 Discussion

In this paper we describe two approaches to assessing the appropriateness of the Markov model. These approaches complement the method of [Kalbfleish and Lawless, 1985] by providing a means of visualizing and identifying specific departures from model assumptions and accommodating unevenly spaced data. The first feature is demon-

Figure 4: Data arising from a time-dependent process: Counts at each timepoint approach



The x-axes represent the expected order statistics (mean over 10 simulations) assuming a time-homogeneous Markov process. The data represented on the y-axes are assumed to arise from a time-dependent process. The upper line in each figure corresponds to the 90% cutoff.

strated in the simulations where the observed statistics that deviated from their expected values tended to correspond to later time points (in the transition approach) and the most resistant state (in the single time point approach.) These types of patterns can not be teased out by the overall χ^2 statistic of [Kalbfleisch and Lawless, 1985]. In searching for patterns that are suggestive of specific alternatives, we suggest tagging statistics corresponding to specific time intervals and states and visually inspecting the plots.

As described above, our approach provides a valid test of the global null hypothesis that the Markov process assumptions hold for all transitions at all time points. If this hypothesis is rejected, our approach can also identify time points and transitions suspected of violating these assumptions. A step-down version of the first procedure we describe can also provide a valid test of individual null hypotheses regarding transi-

tions from each starting state at each time point. An alternative resampling approach, described by [Troendle, 1995] steps down from the most extreme order statistic and rejects the null hypothesis at each step if the corresponding observed statistic is less than a defined threshold. This threshold is determined based on a prespecified alpha and the null distribution of the corresponding order statistic obtained from the resampled data. In this step down procedure, if a null hypothesis is not rejected, than none of the remaining null hypotheses (corresponding to less extreme test statistics) are assumed to be true. The primary difference between Troendle's approach and the one we describe is that our approach additionally evaluates hypotheses according to the distance between the observed order statistics and their expectations. As a result, we allow for rejection based on a single or multiple statistics that are not necessarily the most extreme (i.e. largest) observed statistics.

Our approach estimates state transition rates using maximum likelihood, and then estimates distributions of order statistics by generating data sets from the estimated Markov model. In a large sample setting, we would expect the variance of these estimates to be small and therefore would not expect the uncertainty in estimation to impact our inference greatly. Furthermore, as a diagnostic tool, the proposed approach helps to identify potential outliers even for smaller samples. For the latter setting, consideration might also be given to a different sampling approach that takes into account the uncertainty of estimation. This approach proceeds by repeatedly generating data from the Markov model using the estimated rates from the actual observed data. Each simulated dataset can itself be treated as "observed data," and the procedure described in this manuscript used to find the difference between the psuedo "observed" numbers of transitions and their expected values under the null. The procedure would involve estimating the parameters of the model for each pseudo data set, whose actual values would, of course, have been specified. The distribution of the maximum of these differences for each "observed" dataset" is estimated by repeated generation of such data sets. This distribution is then used for formal testing of outliers. This procedure requires two levels of simulation, but may more accurately reflect the process that generated our results because, under the null, the actual data were generated by a Markov process whose parameters can only be estimated with error. Thus, this approach provides a basis for testing the assumption of a time-homogeneous Markov process that takes into account the estimation of the Markov parameters.

In addition to serving as diagnostics, the approaches we describe can be used as tools for discovering uncharacterized biological and clinical relationships. For example, suppose we observed results similar to those in Section 4 in which data were simulated under a time-dependent process. This would suggest that after a certain amount of time on drug, the rate of transitioning from a sensitive to resistant viral population increases dramatically. Identifying departures with this method tells us, not only about overall lack of fit but about the specific time points and states that violate the Markov assumptions.

Finally, we note that our method is particularly useful for detecting departures

from stationarity when suspected departures occur fairly abruptly. This situation is likely to arise with investigations of HIV genomic data, where probabilities of transitions from one genetic pattern to another ("states" in our Markov model) may change over time because of events that occur during treatment. Such events could include the emergence of a virus archived in some latent cell population or even superinfection with another strain of virus. Further work is required to investigate the power of our method compared to standard methods when changes in transition rates over time are fairly smooth.

References

- [Albert, 1962] Albert, A. (1962). Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Annals of Mathematical Statistics*, 33(2):727–753.
- [Bacheler et al., 2000] Bacheler, L., Anton, E., Kudish, P., Baker, D., Bunville, J., Krakowski, K., Bolling, L., Aujay, M., X., W., 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(9):2475–84.
- [Bacheler et al., 2001] Bacheler, L., Jeffrey, S., Hanna, G., D’Aquila, R., Wallace, L., Logue, K., Cordova, B., Hertogs, K., Larder, B., Buckery, R. and 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 nonnucleoside reverse transcriptase inhibitor therapy. *Journal of Virology*, 75(11):4999–5008.
- [Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.
- [Casella and Berger, 1990] Casella, G. and Berger, R. (1990). *Statistical Inference*. Duxbury Press, Belmont, California.
- [DiRienzo and DeGruttola, 2003] DiRienzo, G. and DeGruttola, V. (2003). Nonparametric methods to predict HIV drug susceptibility phenotype from genotype. *Statistics in Medicine*, 22(17):2785–2798.
- [Foulkes and DeGruttola, 2002] Foulkes, A. and DeGruttola, V. (2002). Characterizing the relationship between HIV-1 genotype and phenotype: Prediction based classification. *Biometrics*, 58:145–156.
- [Foulkes and DeGruttola, 2003] Foulkes, A. and DeGruttola, V. (2003). Characterizing the progression of viral mutations over time. *JASA Applications and Case Studies*, 98(464):859–867.
- [Kalbfleisch and Lawless, 1985] Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a Markov assumption. *JASA*, 90(392):863–871.
- [Pollard and van der Laan, 2002] Pollard, K. and van der Laan, M. (2002). Resampling-based multiple testing: Asymptotic control of type 1 error and applications to gene expression data. *Journal of Statistical Planning and Inference*.
- [Troendle, 1995] Troendle, J. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association*, 90(429):370–378.

[Westfall and Young, 1993] Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: examples and methods for P-value adjustment*. John Wiley and Sons.