Learning to Count: Robust Estimates for Labeled Distances between Molecular Sequences

John D. O'Brien,[1] *Vladimir N. Minin,*†[1] *and Marc A. Suchard**‡§

*Department of Biomathematics, University of California, Los Angeles; †Department of Statistics, University of Washington; ‡Department of Biostatistics, University of California, Los Angeles; and §Department of Human Genetics, University of California, Los Angeles

Researchers routinely estimate distances between molecular sequences using continuous-time Markov chain models. We present a new method, robust counting, that protects against the possibly severe bias arising from model misspecification. We achieve this robustness by generalizing the conventional distance estimation to incorporate the empirical distribution of site patterns found in the observed pairwise sequence alignment. Our flexible framework allows for computing distances based only on a subset of possible substitutions. From this, we show how to estimate labeled codon distances, such as expected numbers of synonymous or nonsynonymous substitutions. We present two simulation studies. The first compares the relative bias and variance of conventional and robust labeled nucleotide estimators. In the second simulation, we demonstrate that robust counting furnishes accurate synonymous and nonsynonymous distance estimates based only on easy-to-fit models of nucleotide substitution, bypassing the need for computationally expensive codon models. We conclude with three empirical examples. In the first two examples, we investigate the evolutionary dynamics of the influenza A hemagglutinin gene using labeled codon distances. In the final example, we demonstrate the advantages of using robust synonymous distances to alleviate the effect of convergent evolution on phylogenetic analysis of an HIV transmission network.

Introduction

How to estimate the evolutionary distance between molecular sequences is a fundamental problem for comparative analyses (Lio and Goldman 1998; Gascuel 2005). In these analyses, distance often implies an estimate of the expected number of substitutions between aligned sequences under an assumed continuous-time Markov chain (CTMC) model of nucleotide or codon substitution. Reflecting distances' importance in biology, a large volume of literature of different methods has grown up for making this calculation (Jukes and Cantor 1969; Hasegawa et al. 1985; Nei and Gojobori 1986; Li and Gu 1995; Rzhetsky and Nei 1995). Researchers, for example, rely on these distances to generate phylogenies, estimate adherence to a molecular clock, and identify nucleotide or amino acid sites that experience distinct evolutionary pressure. Evolutionary distances can be generalized to "count" only certain subsets of transitions in the CTMC. We call these labeled distances. An important example of labeled distances is nonsynonymous and synonymous distances that mark codon transitions by whether or not they induce an amino acid change, respectively. Researchers eagerly exploit these distances, for example, to identify positive selection across the influenza genome (Earn et al. 2002, see citations therein), study convergent evolution in the HIV genome (Lemey et al. 2005), and establish orthologous and paralogous gene relationships (Goodstadt and Ponting 2006).

In this paper, we introduce a new framework for calculating labeled distances between sequences that we call robust counting. The framework builds on any reversible CTMC model of substitution and provides strong protection against bias originating from model misspecification. The method achieves robustness by first conditioning on pairwise site patterns to calculate the conditional mean numbers of labeled substitutions and then averaging these conditional expectations over the empirical distribution of site patterns found in the "observed" sequence data. This is in contrast to conventional distance estimation methods that implicitly average over the theoretical distribution of pairwise site patterns. The conditioning part of our method capitalizes on the recent progress in efficient calculations of the distributional moments of a Markov chain–induced counting process (Holmes and Rubin 2002; Hobolth and Jensen 2005; Minin and Suchard 2008).

Distance estimation methods for nucleotide sequences commonly rely on maximum likelihood estimates (MLEs) of CTMC models of nucleotide substitution. Usually, these CTMC models impose that nucleotide changes between sequences arise from a stationary, time reversible Markovian process that is independent across sites. An infinitesimal rate matrix specifies these models, determining the rates at which nucleotide states transition among themselves. Jukes and Cantor (1969) introduce the use of CTMC models for nucleotide substitution by assuming an equal frequency for all nucleotides at stationarity and no difference in rates among nucleotides states. Kimura (1980) refines this model by introducing a parameter κ to account for differences between transitional and transversional nucleotide substitution rates. Kishino and Hasegawa (1989) and Hasegawa et al. (1985) further expand these efforts to account for unequal nucleotide frequencies in the F84 and HKY models, respectively. More elaborate models follow these initial approaches, accounting for additional variability in substitution patterns (Tamura and Nei 1993), and lead to a general time reversible (GTR) model with maximal parametric freedom among reversible rates (Gu and Li 1996). Several of these authors also consider methods for determining labeled nucleotide distances, such as the expected numbers of transitions or transversions.

Methods for estimating labeled codon distances are noticeably fewer, primarily limited to the context of nonsynonymous or synonymous distance estimation, yielding

[1] These authors wish to be considered as joint first authors.

Key words: robust counting, labeled codon distance, empirical distribution, Markov chain substitution model.

E-mail: msuchard@ucla.edu.

Mol. Biol. Evol. 26(4):801–814. 2009
doi:10.1093/molbev/msp003
Advance Access publication January 8, 2009

two main approaches. The first approach parallels nucleotide distance estimation and suggests fitting an appropriate codon CTMC model to the observed data and using the fitted model to predict the expected number of labeled codon substitutions (Goldman and Yang 1994; Muse and Gaut 1994; Nielsen and Yang 1998). Two problems arise with this approach. First, the high complexity of the codon state space significantly increases one's chance of grossly misspecifying the codon evolutionary model. Second, even the simplest codon models require computationally costly numerical likelihood maximization, defeating the purpose of using fast distance-based phylogenetic reconstruction methods. Schneider et al. (2007) try to address these problems by using previously estimated empirical codon matrices to infer synonymous distances. However, the general applicability of preestimated empirical codon matrices is not yet clear. Doron-Faigenboim and Pupko (2007) find a middle ground between parametric and empirical approaches, but similarly to purely parametric approaches, their algorithm requires nontrivial numerical likelihood maximization. The second approach, taken by Nei and Gojobori (1986, denoted as NG86), Ina (1995), and Yang and Nielsen (2000, denoted as YN00), first uses a parsimony argument for counting synonymous and nonsynonymous mutations and then tries to correct the parsimony estimates using simple nucleotide evolutionary models. The chief advantage of these methods is their computational efficiency. The main drawback of the second approach lies in its heuristic nature, complicating theoretical analysis, and making generalizations nearly impossible. Robust counting, applied to synonymous and nonsynonymous distance estimation, rests somewhere between these two existing approaches for inferring these distances. Similarly to the heuristic approach, we use easy-to-fit nucleotide models to build a deliberately misspecified model of codon evolution. However, instead of the heuristic parsimony-based arguments, we profit from probabilistic conditioning and Markov chain theory to count the unobserved synonymous and nonsynonymous mutations.

In the case of labeled codon distances, researchers may intentionally oversimplify evolutionary models to reach computational tractability (Yang and Nielsen 2000). However, because the "true" model of nucleotide substitution is never known and is most likely not Markovian, model misspecification is inherent to "almost" any CTMC-based sequence analysis (Blount et al. 2008). Although the harmful consequences of CTMC model misspecification are well documented (Hasegawa et al. 1985; Yang 1997; Buckley et al. 2001), the only solution available to researchers is to choose among a small and possibly inadequate set of evolutionary models (Suchard et al. 2001; Sullivan and Joyce 2005).

We propose to approach the problem of model misspecification from the perspective of robust statistics, a set of statistical procedures that are relatively insensitive to deviations from assumed underlying distributions (Huber 1981). To arrive at our robust distance estimators, we first provide a general derivation for conventional distance calculations in terms of both Markov chains and Markov chain–induced counting processes. We then show how, in this framework, robust counting is a natural extension of conventional methods. As part of this discussion, we present a unified understanding of labeled distances, demonstrating how to make estimates of these distances with both conventional and robust methods.

To establish the efficiency of robust counting, we provide two simulation studies, the first describing the capacity of the method to calculate robust estimates of labeled nucleotide distances and the coverage properties of the estimators and the second demonstrating how the robust codon method can be used to estimate nonsynonymous and synonymous changes. This latter study finds that robust counting performance in estimating nonsynonymous distances is substantially better than existing methods. We conclude with three empirical examples. The first exhibits how robust counting estimates of nonsynonymous and synonymous distances can be used to detail the positive selection history of the influenza A hemagglutinin gene. The second employs the same data and considers the structure of volatility change distances (Plotkin and Dushoff 2003), illustrating the flexibility of robust counting in estimating novel labeled distances. The third example reveals how robust estimates of synonymous distances can be employed to reduce convergent evolution bias during a distance-based phylogenetic analysis of a known HIV transmission network (Lemey et al. 2005).

Methods

Conventional Distances and Mutation Labeling

Let $\mathbf{Y}_1=(y_{11},\ldots,y_{1L})$ and $\mathbf{Y}_2=(y_{21},\ldots,y_{2L})$ be two aligned molecular sequences of length L. Each site (y_{1s}, y_{2s}) for $s=1,\ldots,L$ evolves independently under an irreducible, reversible M-state CTMC $\{X_t\}$ with infinitesimal generator $\boldsymbol{\Lambda} = \{\lambda_{ij}\}$. We define the rate of the leaving state i to be $\lambda_i=\sum_{j\neq i}^{M}\lambda_{ij}$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ be the stationary distribution of this process. The conventional method for finding the distance between two sequences consists of two steps. In the first step, the model parameters $\boldsymbol{\Lambda}$ are estimated between pairs of sequences by maximizing the likelihood of the data directly or by using a suitable approximation, as in Yang (1994). Using finite-time transition probabilities $\boldsymbol{P} = e^{\boldsymbol{\Lambda}t} = \{p_{ij}(t)\}$, we can write the likelihood of the data as

$$\mathbf{Pr}(\mathbf{Y}_1, \mathbf{Y}_2 \mid \boldsymbol{\Lambda}, t) = \prod_{s=1}^{L} \hat{\pi}_{Y_{1s}} p_{Y_{1s},Y_{2s}}(t), \qquad (1)$$

where $\hat{\boldsymbol{\pi}}$, for appropriate models, is usually estimated using empirical state frequencies, possibly from a larger set of sequences (Gu and Li 1996). For some models, the stationary distribution may be fixed and so $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}$. As sequences contain only information about the product $\boldsymbol{\Lambda}t$, we may either impose additional constraints on $\boldsymbol{\Lambda}$ or set $t = 1$. Throughout this paper, we will assume $t = 1$ and estimate $\hat{\boldsymbol{\Lambda}}$. In the second step of conventional estimation, $\hat{\boldsymbol{\Lambda}}$ is used to calculate the expected number of substitutions per site predicted under the Markov model

$$\hat{d} = \sum_{i=1}^{M} \hat{\boldsymbol{\pi}}_i \hat{\lambda}_i = \hat{\boldsymbol{\pi}}[\hat{\boldsymbol{\Lambda}} - \mathrm{diag}(\hat{\boldsymbol{\Lambda}})]\boldsymbol{I}, \qquad (2)$$

where $\mathrm{diag}(\hat{\mathbf{\Lambda}})$ is a matrix obtained by setting all off-diagonal elements of $\hat{\mathbf{\Lambda}}$ to 0 and $\mathbf{1}$ is an M-column vector of ones.

We generalize this definition of distance to any subset of labeled transitions by introducing the concept of a counting process induced by the underlying Markov chain. Define the set $\mathcal{L}$ to be a subset of the lattice, $\{1,\ldots,m\}^2$, of those specific $i\rightarrow j$ transitions we wish to count. Elements (i, i) are always excluded from $\mathcal{L}$. We now define $N_t^{\mathcal{L}}$ to be the counting process that tracks only transitions labeled by $\mathcal{L}$. Defining $g(t)=\mathbf{E}(N_t^{\mathcal{L}})$, we can write the Chapman–Kolmogorov equation for this process:

$$g(t+h)=\sum_{i=1}^{M}\pi_i\left[g(t)\cdot(1-\lambda_i h)+g(t)\sum_{j\neq i}^{M}\lambda_{ij}h1_{\{(i,j)\notin\mathcal{L}\}}+(g(t)+1)\sum_{j\neq i}^{M}\lambda_{ij}h1_{\{(i,j)\in\mathcal{L}\}}\right]+o(h), \tag{3}$$

where h is a small increment in time and $1_{\{\cdot\}}$ is the indicator function. Noting that $\sum_{i=1}^{M}\pi_i=1$, rearranging terms and dividing through by h, we arrive at

$$\frac{g(t+h)-g(t)}{h}=\sum_{i=1}^{M}\pi_i\left[-g(t)\lambda_i+g(t)\cdot\sum_{j\neq i}^{M}\lambda_{ij}+\sum_{j\neq i}^{M}\lambda_{ij}1_{\{(i,j)\in\mathcal{L}\}}\right]=\sum_{i=1}^{M}\pi_i\sum_{j\neq i}^{M}\lambda_{ij}1_{\{(i,j)\in\mathcal{L}\}}. \tag{4}$$

Taking the limit as $h\rightarrow 0$ yields

$$g'(t)=\sum_{i=1}^{M}\pi_i\sum_{j\neq i}^{M}\lambda_{ij}1_{\{(i,j)\in\mathcal{L}\}}. \tag{5}$$

Imposing the boundary condition that $g(0)=0$, equation (5) can be solved explicitly, giving

$$g(t)=t\cdot\sum_{i=1}^{M}\pi_i\sum_{j\neq i}^{M}\lambda_{ij}1_{\{(i,j)\in\mathcal{L}\}}. \tag{6}$$

Setting $t=1$ as before, we now construct distances for the labeling set $\mathcal{L}$

$$d_{\mathcal{L}}=\mathbf{E}(N_1^{\mathcal{L}})=g(1)=\sum_{i=1}^{M}\pi_i\sum_{j\neq i}^{M}\lambda_{ij}1_{\{(i,j)\in\mathcal{L}\}}=\boldsymbol{\pi}\mathbf{\Lambda}_{\mathcal{L}}\mathbf{1}, \tag{7}$$

where $\mathbf{\Lambda}_{\mathcal{L}}=\left\{\lambda_{ij}\times 1_{\{(i,j)\in\mathcal{L}\}}\right\}$ is the restricted generator matrix. As with unlabeled counts, estimates of $\hat{\boldsymbol{\pi}}$ and $\hat{\mathbf{\Lambda}}_{\mathcal{L}}$ are inferred from the data to give $\hat{\boldsymbol{d}}_{\mathcal{L}}$. In previous work involving conventional distances, several authors have calculated $d_{\mathcal{L}}$ for specific labeling sets of small Markov chains, usually of different types of nucleotide changes such as transitions or transversions (Felsenstein 2004, Chapter 13). We believe that this general formulation of conventional labeled distances in terms of any arbitrary Markov chain–induced counting process is novel.

Robust Distances

Using the law of total expectation, we can write $d_{\mathcal{L}}$ as

$$d_{\mathcal{L}}=\sum_{i,j=1}^{M}\mathbf{E}(N_1^{\mathcal{L}}|X_0=i,X_1=j)\Pr(X_0=i,X_1=j). \tag{8}$$

Notice that the conventional distance estimation implicitly assumes

$$\Pr\left(X_0=i,X_1=j\right)=\pi_i p_{ij}\left(1\right), \tag{9}$$

to arrive at the estimate $\hat{\boldsymbol{d}}_{\mathcal{L}}$. The central idea of robust counting is to replace the predicted probabilities of site patterns (9) with the empirical frequencies of (i, j) found in the sequence data, that is

$$\Pr(X_0=i,X_1=j)=\frac{1}{L}\sum_{s=1}^{L}1_{\{y_{1s}=i,\,y_{2s}=j\}}. \tag{10}$$

By taking this empirical approach to estimation of $\Pr(X_0=i,\,X_t=j)$, we partially free ourselves from the CTMC's parametric assumptions and so make our estimation more robust to model misspecification. Plugging empirical frequencies into equation (8), we arrive at a new definition of labeled distances

$$d_{\mathcal{L}}^{r}=\frac{1}{L}\sum_{s=1}^{L}r_{y_{1s},y_{2s}}(1),$$

where

$$r_{ij}\left(t\right)=\mathbf{E}\left(N_t^{\mathcal{L}}\middle|X_0=i,X_t=j\right). \tag{11}$$

In contrast to conventional distances, in order to estimate $d_{\mathcal{L}}^{r}$, we first need to obtain $\mathbf{E}\left(N_1^{\mathcal{L}}\,\middle|\,X_0=i,X_1=j\right)$ for all $i,j=1,\ldots,m$. Because these quantities are not directly observed in the data, we cannot use a completely empirical approach and return briefly to our Markov chain model. We follow the first step of conventional distance estimation and obtain $\hat{\boldsymbol{\pi}}$ empirically and $\hat{\mathbf{\Lambda}}$ by maximizing the likelihood (1) or employing a suitable approximation. Given $(\hat{\boldsymbol{\pi}},\hat{\mathbf{\Lambda}})$, it is now possible to efficiently compute the conditional expectations (11).

Define the restricted first moment of the counting process $N_t^{\mathcal{L}}$ to be

$$c_{ij}\left(t\right)=\mathbf{E}\left(N_t^{\mathcal{L}}\cdot 1_{\{X_t=j\}}\middle|X_0=i\right). \tag{12}$$

Assembling these terms into a matrix $\mathbf{C}(t)=\{c_{ij}(t)\}$ and following Ball and Milne (2005), we have that

$$\mathbf{C}(t)=\int_0^t \mathrm{e}^{\mathbf{\Lambda}u}\mathbf{\Lambda}_{\mathcal{L}}\mathrm{e}^{\mathbf{\Lambda}(t-u)}du. \tag{13}$$

This integral can be calculated efficiently for reversible Markov chains (Minin and Suchard 2008) and so we plug $\hat{\mathbf{\Lambda}}$ into equation (13) to obtain $\hat{c}_{ij}\left(1\right)$. We use these

values to calculate $\hat{\boldsymbol{r}}_{ij}(1)$ through the definition of conditional expectation:

$$r_{ij}(t) = \frac{c_{ij}(t)}{p_{ij}(t)}. \tag{14}$$

We use the estimate $\hat{\boldsymbol{\Lambda}}$ to calculate $\hat{\mathbf{P}}(1) = e^{\hat{\boldsymbol{\Lambda}}}$. Setting $\hat{\boldsymbol{r}}_{ij}(1) = \hat{c}_{ij}(1)/\hat{p}_{ij}(1)$, we use equation (15) to furnish our robust estimator

$$\hat{\boldsymbol{d}}^{r}_{\mathcal{L}} = \frac{1}{L}\sum_{s=1}^{L} \hat{\boldsymbol{r}}_{y_{1s},y_{2s}}(1). \tag{15}$$

It is important to note that if the Markov model is known without error, then the strong law of large numbers and asymptotic consistency of MLEs yield

$$\lim_{L\to\infty} \frac{1}{L}\sum_{s=1}^{L} 1_{\{y_{1s}=i,y_{2s}=j\}} = \pi_i p_{ij}(1) \text{ almost surely}, \tag{16}$$

and

$$\lim_{L\to\infty} \hat{\pi}_i \hat{p}_{ij}(1) = \pi_i p_{ij}(1) \text{ in probability}. \tag{17}$$

As the estimation of site pattern probabilities is the only difference between the robust and conventional distances, the asymptotic consistency of MLEs applied to equation (8), implies that

$$\lim_{L\to\infty} \left|\hat{\boldsymbol{d}}_{\mathcal{L}} - \hat{\boldsymbol{d}}^{r}_{\mathcal{L}}\right| = 0 \text{ in probability}. \tag{18}$$

However, these asymptotic results hold only if no model misspecification occurs. The hope is that using empirical site pattern frequencies will make robust distances less prone to biases due to model misspecification. Notice that in robust distance calculations we still use potentially misspecified $\hat{\boldsymbol{\Lambda}}$ to calculate site-specific expectations (11). For short evolutionary distances seen in real data, the error in these calculations should be relatively small. We do expect the bias of robust distances to increase with the number of substitutions between sequences. The behavior of this bias will be explored in our simulation studies.

Handling Missing Data

In the previous subsection, we assume that our pairwise sequence alignment $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ has no missing data. Now, suppose that $\mathbf{Y}$ may contain sites of the form $(i, -)$, $(-, i)$, and $(-, -)$ that require calculation of conditional expectations

$$\begin{aligned} r_{i-}(t) &= \mathbf{E}\left(N^{\mathcal{L}}_t \middle| X_0 = i\right),\ r_{-i}(t) \\ &= \mathbf{E}\left(N^{\mathcal{L}}_t \middle| X_1 = i\right) \text{ and } r_{--}(t) = \mathbf{E}\left(N^{\mathcal{L}}_t\right). \end{aligned} \tag{19}$$

As before, we accomplish these calculations with the help of estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\Lambda}}$ and setting $t = 1$.

First, we notice that reversibility of $\{X_t\}$ insures that $r_{i-}(t) = r_{-i}(t)$. Collecting these expectations into a column vector $\mathbf{r}(t) = (r_{1-}(t), \ldots, r_{m-}(t))^T$ and recalling that $e^{\boldsymbol{\Lambda}(t)}\mathbf{1} = \mathbf{1}$, we arrive at

$$\mathbf{r}(t) = \mathbf{C}(t)\mathbf{1} = \int_0^t e^{\boldsymbol{\Lambda}u} du \times \boldsymbol{\Lambda}_{\mathcal{L}}\mathbf{1}. \tag{20}$$

The integral on the right handside of equation (20) can be computed with the help of the fundamental matrix of $\{X_t\}$ (Ball and Milne 2005). However, because $\mathbf{C}(t)$ is available to us, we simply multiply this matrix by the column vector of ones. The last expectation in (19) reduces to $r_{--}(t) = \boldsymbol{\pi}\mathbf{C}(t)\mathbf{1} = \boldsymbol{\pi}\boldsymbol{\Lambda}_{\mathcal{L}}\mathbf{1}t$ because $\boldsymbol{\pi}e^{\boldsymbol{\Lambda}(t)} = \boldsymbol{\pi}$. Notice that when both nucleotides are missing at a site, we use the conventional labeled distance to impute the missing number of labeled mutations at this site.

Simulations

To demonstrate the utility of robust counting and to compare its operation with conventional methods, we perform two simulation studies. The first study compares conventional and robust methods in estimating the number of labeled nucleotide transitions when the underlying CTMC model of substitution is correctly and incorrectly specified. The second study contrasts a robust counting-based approach against two established methods for estimating the number of nonsynonymous or synonymous codon substitutions between sequences. We implement the robust counting methods in a new R package called markovjumps available at http://www.stat.washington.edu/vminin/markovjumps. R is an open source statistical software program available at http://www.r-project.org. Sequence data simulation employs the markovjumps package and PAML (Yang 2007), where the latter also provides the established nonsynonymous distance estimators.

Nucleotide Distances under Model Misspecification

In this first study, we compare the conventional approach against robust counting for estimating labeled nucleotide distances. As we can specify the model of nucleotide substitution from which the data derive, we can test the performance of the two methods under both the true and misspecified models. We first generate data under the GTR model for nucleotide substitution. For a simple misspecified model, we consider F84 as detailed in Yang (1994). We then draw inference via conventional and robust counting methods building on both GTR and F84 as underlying CTMCs.

We generate 1,000 pairwise sequence data sets of $L = 2{,}000$ nt for a set of six different pairwise distances under GTR with

$$\boldsymbol{\Lambda}_{\text{GTR}} = \begin{pmatrix} \cdot & r_1\pi_G & r_2\pi_C & r_3\pi_T \\ r_1\pi_A & \cdot & r_4\pi_C & r_5\pi_T \\ r_2\pi_A & r_4\pi_G & \cdot & r_6\pi_T \\ r_3\pi_A & r_5\pi_G & r_6\pi_C & \cdot \end{pmatrix}, \tag{21}$$

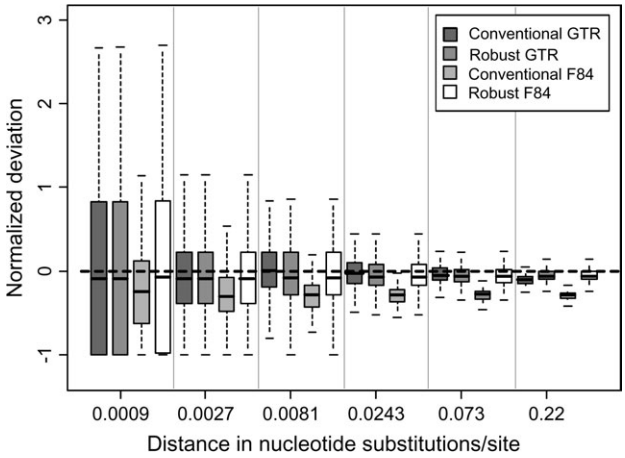

FIG. 1.—Performance box-plots for conventional GTR (dark gray), robust GTR (medium gray), conventional F84 (light gray), and robust F84 (white) estimators over six pairwise sequence distances. We standardize performance across distances via normalized deviations, $(\hat{d}_A - d_A)/d_A$. The true distances for each of the six simulation conditions are listed on the x axis.

specified by $\mathbf{r} = (0.5, 0.3, 0.6, 0.2, 0.3, 0.2)$ and $\boldsymbol{\pi} = (0.2, 0.2, 0.3, 0.3)$. We note that these parameter values ensure that the GTR model is not a degenerate form of the F84 model. The true pairwise distances d between sequences are given on the x axis in figure 1. For labeling, we consider counting transitions only into and out of the state "A" such that our choice of labeling corresponds to the generator matrix $\boldsymbol{\Lambda}_{\mathcal{L}} = \{\lambda_{ij} \times 1_{\{(i,j)\in\mathcal{L}\}} : \mathcal{L} = \{(i,A)\} \cup \{(A,j)\}\}$. We let d_A denote this labeled distance. Using equation (7), we can calculate the theoretical number d_A of such transitions for each value of d.

For each pair of simulated sequences $(\mathbf{Y}_1, \mathbf{Y}_2)$, we calculate the model parameters in the following way. The estimate of the stationary distribution, $\hat{\boldsymbol{\pi}}$, is the nucleotide frequencies observed across both sequences. We find the remaining parameters of $\hat{\boldsymbol{\Lambda}}_{\mathrm{GTR}}$ by maximizing equation (1) with respect to each of the model parameters using the Nelder–Mead algorithm. To find the remaining parameters for $\hat{\boldsymbol{\Lambda}}_{\mathrm{F84}}$, we employ a closed-form, approximate solution (Yang 1994; Yang and Nielsen 2000). We then use equation (7) to generate conventional distance estimates. Similarly, we use equation (15) to generate robust distance estimates.

Figure 1 presents the results from this experiment. For each simulated distance, we summarize the performance of four different estimators (conventional GTR, robust GTR, conventional F84, and robust F84) via box-plots of the estimators' normalized deviation from the true value. The bias of conventional estimates under the true model and robust estimates under both the true and misspecified model are virtually identical. Unsurprisingly, conventional estimates under F84 return substantially greater bias. Whereas the variance of the robust F84 estimator is considerably larger than for the conventional F84 estimator, the performance of the robust F84 estimator is nearly identical to both the robust and conventional estimators under the true model. Because we estimate estimator biases via Monte Carlo simulations, we need to examine the precision of these estimates. For example, in the second simulation regime with $d = 0.0027$ and $d_A = 0.0016$, we estimate F84-based conventional and robust estimator expectations as 0.0016 and 0.0011, respectively. The corresponding Monte Carlo standard errors of these expectations are 2.8×10^{-5} and 1.8×10^{-5}, two orders of magnitude smaller. Hence, 1,000 Monte Carlo iterations provide more than sufficient information to compare estimator biases.

To further explore the relationship between robust counting estimates under model misspecification relative to conventional estimates under the true model, we perform a simple simulation to examine how the variance and mean squared error (MSE) of these estimators change as the number of nucleotide sites L increases. We generate 500 additional pairwise alignments under GTR with $L = \{1{,}200, 2{,}400, 4{,}800, 9{,}600, 19{,}200, 38{,}400\}$ at $d_A = 0.061$ and apply the conventional GTR and robust F84 estimators to measure this distance. Figure 2 demonstrates that the variance of the robust F84 estimator is consistently smaller than the conventional GTR estimator variance. However, this advantage comes at the cost of larger asymptotic bias as revealed by the MSE plot in figure 2. Interestingly, the MSE of both estimators essentially coincide when the alignment length is equal to 1200 and 2400, showing that the robust F84 and conventional GTR estimators perform comparably at realistic sample sizes.

Nonsynonymous Distances

In this second study, we compare an application of robust counting with the NG86 and YN00 methods of estimating the number of nonsynonymous substitutions between sequences. To estimate the expected number of synonymous and nonsynonymous mutations with robust counting, we first require a Markov model on the state space of codons. We would like to avoid standard codon models (Goldman and Yang 1994; Muse and Gaut 1994) because their estimation requires computationally costly and occasionally unstable numerical optimization. Therefore, we settle for an easy-to-fit product composition of nucleotide models. This model does not appropriately account for differences in the rates of synonymous and nonsynonymous mutations in protein coding regions, leaving this task to robust counting.

Let $X_t^{(c)}$ be a CTMC model of nucleotide substitution for codon positions $c = 1, 2, 3$, specified by infinitesimal rate matrices $\boldsymbol{\Lambda}^{(c)}$. For each codon position, we use empirical nucleotide frequencies to estimate stationary distributions $\hat{\boldsymbol{\pi}}^{(1)}$, $\hat{\boldsymbol{\pi}}^{(2)}$, and $\hat{\boldsymbol{\pi}}^{(3)}$. Next, we maximize the likelihood (1) or employ a suitable approximation for each codon position and arrive at $\hat{\boldsymbol{\Lambda}}^{(c)}$. Depending on the informativeness of the nucleotide data, we may be able to estimate each $\hat{\boldsymbol{\Lambda}}^{(c)}$ independently or allow all codon sites to be identically distributed by letting $\hat{\boldsymbol{\Lambda}}^{(1)} = \hat{\boldsymbol{\Lambda}}^{(2)} = \hat{\boldsymbol{\Lambda}}^{(3)}$. In either case, we establish three models for nucleotide substitution at each of the codon positions. From these, we generate a new CTMC Z_t that is the product space composition of the three $X_t^{(c)}$'s

$$Z_t = \left(X_t^{(1)}, X_t^{(2)}, X_t^{(3)}\right). \tag{22}$$

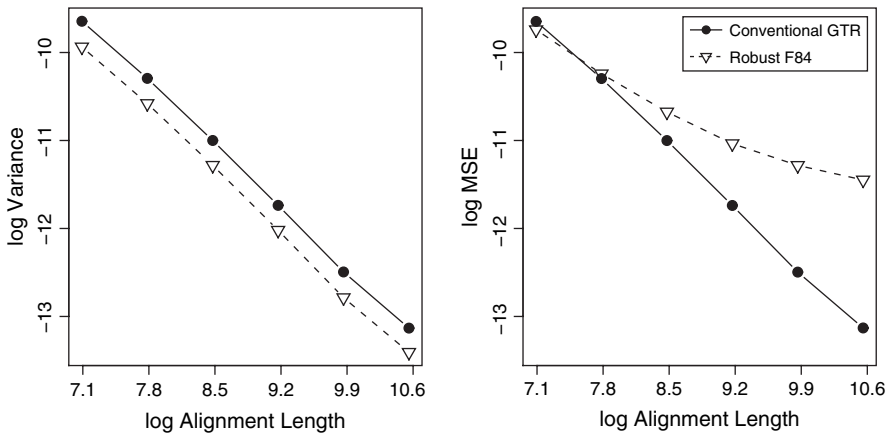

FIG. 2.—Log-log plot of estimator variance and MSE by the alignment length. We report variance and MSE over 500 simulated data sets for the conventional GTR (solid lines) and robust F84 (dotted line) estimators.

Assuming independence among codon positions, the 64×64 infinitesimal generator of Z_t, $\mathbf{W}$, is obtained by the Kronecker sum of the infinitesimal rate matrices of the nucleotide models (Neuts 1995)

$$\mathbf{W} = \mathbf{\Lambda}^{(1)} \oplus \mathbf{\Lambda}^{(2)} \oplus \mathbf{\Lambda}^{(3)}. \tag{23}$$

In this simulation study, we use three different F84 models to model codon evolution. Approximate expressions for the F84 infinitesimal generator MLE are available in closed form, making the estimation part of robust counting procedure extremely fast. We define nonsynonymous labeling set $\mathcal{L}_N$ by letting $(i,j) \in \mathcal{L}_N$ if there is a single nucleotide change in any one of the three codon positions and the change does alter the translated amino acid, otherwise not. Although start and stop codons are permissible states in our product composition of F84 models, we exclude them from the nonsynonymous labeling set.

The NG86 method compares the estimated numbers of synonymous and nonsynonymous mutations with expectations under neutrality. The method begins by calculating "potential" numbers of synonymous and nonsynonymous sites, $\hat{S}$ and $\hat{N}$, that represent the numbers of alignment sites expected to experience synonymous and nonsynonymous mutations under a neutral model of evolution. The second step estimates the numbers of observed synonymous and nonsynonymous mutations using a parsimony argument with some heuristic modifications to account for situations where two codons differ at two or three positions. Finally, the estimator divides these parsimony-derived counts by the potential quantities and applies a Jukes–Cantor correction to these normalized counts. One can interpret the resulting estimates $\hat{r}_S^{\mathrm{NG86}}$ and $\hat{r}_N^{\mathrm{NG86}}$, more commonly denoted by dS and dN, as expected numbers of synonymous and nonsynonymous substitutions per synonymous or nonsynonymous site, respectively. Ina (1995) modifies the NG86 method to account for different transitional and tranversional rates in the neutral evolutionary model. Finally, Yang and Nielsen (2000) substantially refine the method by exploiting the F84 nucleotide model and empirical codon frequencies to make the neutral model more realistic, providing estimates $\hat{r}_S^{\mathrm{YN00}}$ and $\hat{r}_N^{\mathrm{YN00}}$.

Because robust counting is designed to estimate the absolute expected number of labeled substitutions, we convert the NG86 and YN00 estimates to the absolute labeled distance scale via the following transformations:

$$\hat{d}_S = \hat{r}_S\left(\frac{3\hat{S}}{\hat{N}+\hat{S}}\right) \text{ and } \hat{d}_N = \hat{r}_N\left(\frac{3\hat{N}}{\hat{N}+\hat{S}}\right). \tag{24}$$

We simulate sequence data under the M_0 model of codon substitution (Goldman and Yang 1994) parameterized in terms of the transition/transversion rate ratio κ and nonsynonymous/synonymous rate ratio ω, employing a range of parameter values for κ, ω, the number of codon sites L_C and the pairwise sequence distances (table 1), as well as considering three different codon frequency distributions. These are the uniform distribution (equal), primate mitochondrial RNA frequencies (mt), and HIV *env* frequencies (env) provided in PAML (Yang 2007). Our parameter space covers that used in Yang and Nielsen (2000) for a similar estimator comparison. For each simulation, we calculate $\hat{d}_N$, the estimated nonsynonymous distance, under robust counting, NG86 and YN00.

In finding estimates from YN00, we employ the codon weighting option provided in PAML, although we observe only a slight improvement relative to unweighted runs.

Across this broad range of parameter values, distances, and codon distributions, we observe that robust counting generally performs better than either YN00 or NG86 in estimating d_N. Figure 3 presents estimator comparisons under four prototypical regimes.

We see the largest qualitative difference in performance across estimators when considering different codon frequencies. Under the uniform distribution (fig. 3, bottom-left), YN00 estimates demonstrate the smallest bias; however, advantage fades for increasing values of κ and ω (data not shown). Variances here also remain grossly equal across estimators. Under the two empirically derived distributions, robust counting returns the smallest bias at the cost of increased variance. This trade-off becomes more pronounced as distance increases. However, the exchange is beneficial as the 95% coverage of robust counting much more likely covers the true value than the other estimators (e.g., see fig. 3, top-left).

Table 1
Parameter Values Used for Nonsynonymous Distance Simulations

Parameter	Values
κ	1, 2, 5, 10, 20
ω	0.33, 1, 2, 3, 5, 10
L_C	100, 300, 500
Distance	0.017, 0.033, 0.083, 0.17, 0.33, 0.5

NOTE.—We provide distances in number of nucleotide substitutions per codon site.

The number of codon sites in the data also affects the estimators' performances. Estimator variances naturally decrease with an increasing number of sites. The rate of shrinkage varies by method, with NG86 demonstrating only slight improvement and robust counting and YN00 decreasing comparably, consistent with our findings for nucleotide distances. However, the relative performance of the estimators varies in different regimes of κ and ω; some regimes proffering one estimator over another and some not. Although the examples in figure 3 vary in their sequence length and codon frequency, the four are also prototypical of estimator behavior with increasing κ and ω separately and together. Despite that this robust counting implementation only relies on a product composition of three nucleotide models instead of a full codon model, the method generally provides the best coverage properties.

Empirical Examples

Evolution of Influenza H3N2 Hemagglutinin

Studies of influenza evolutionary dynamics focus heavily on patterns of change in the hemagglutinin protein (HA0), the primary determinant of viral cell surface binding and membrane fusion. Over the past 20 years, research supports the existence of a phylogenetic backbone for hemagglutinin, where temporally sequential clusters of taxa originate successively from a single trunk lineage (Fitch

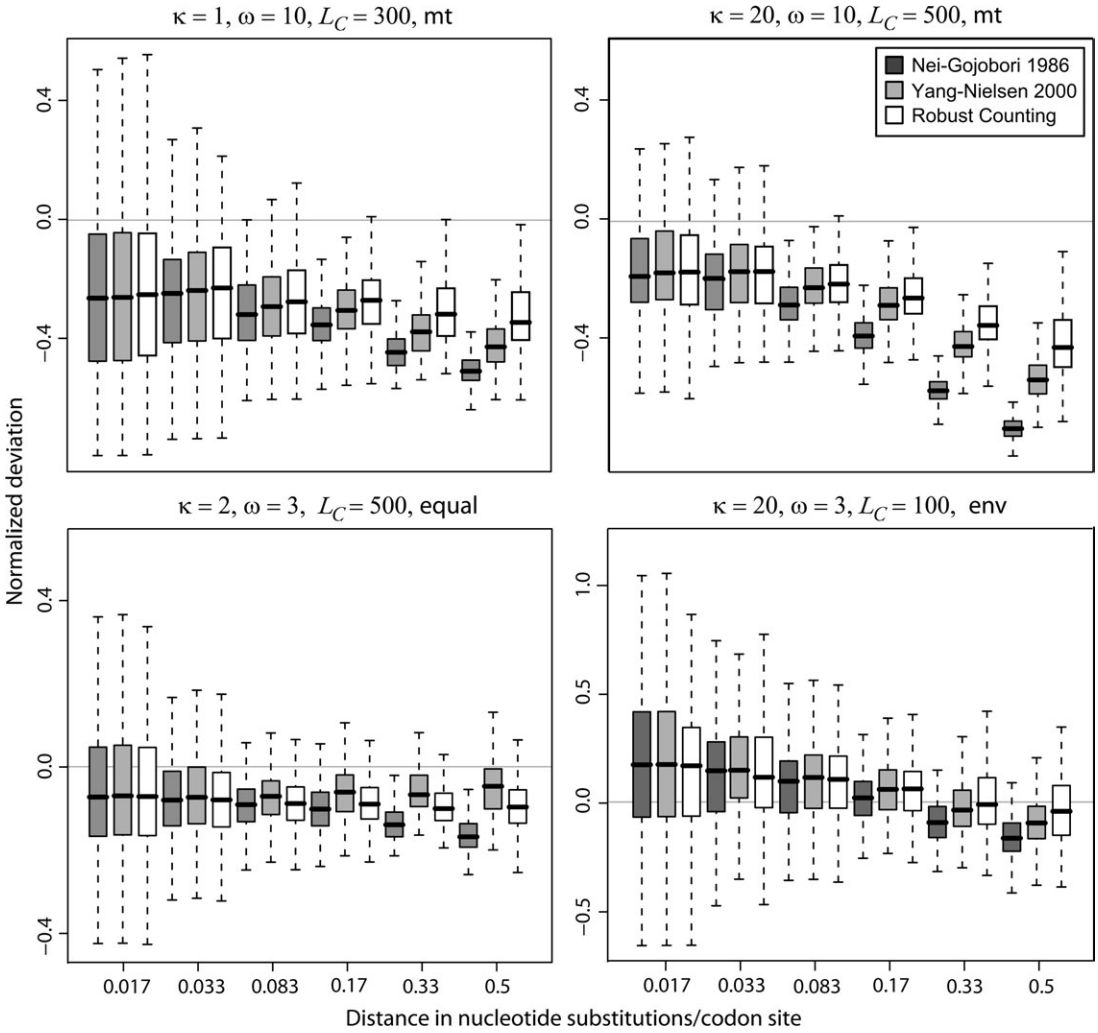

FIG. 3.—Typical performance box-plots for three methods of nonsynonymous distance estimation under different parameter values across distances. We report performance in terms of normalized deviations, $(\hat{d}_N - d_N)/d_N$, and plot headings list specific values of κ, ω, the number of codon sites L_C, and the codon distribution. "Equal" denotes equal codon frequencies, "mt" denotes the empirical distribution of primate mitochondrial RNA, and "env" denotes the empirical distribution of the HIV *env* gene.

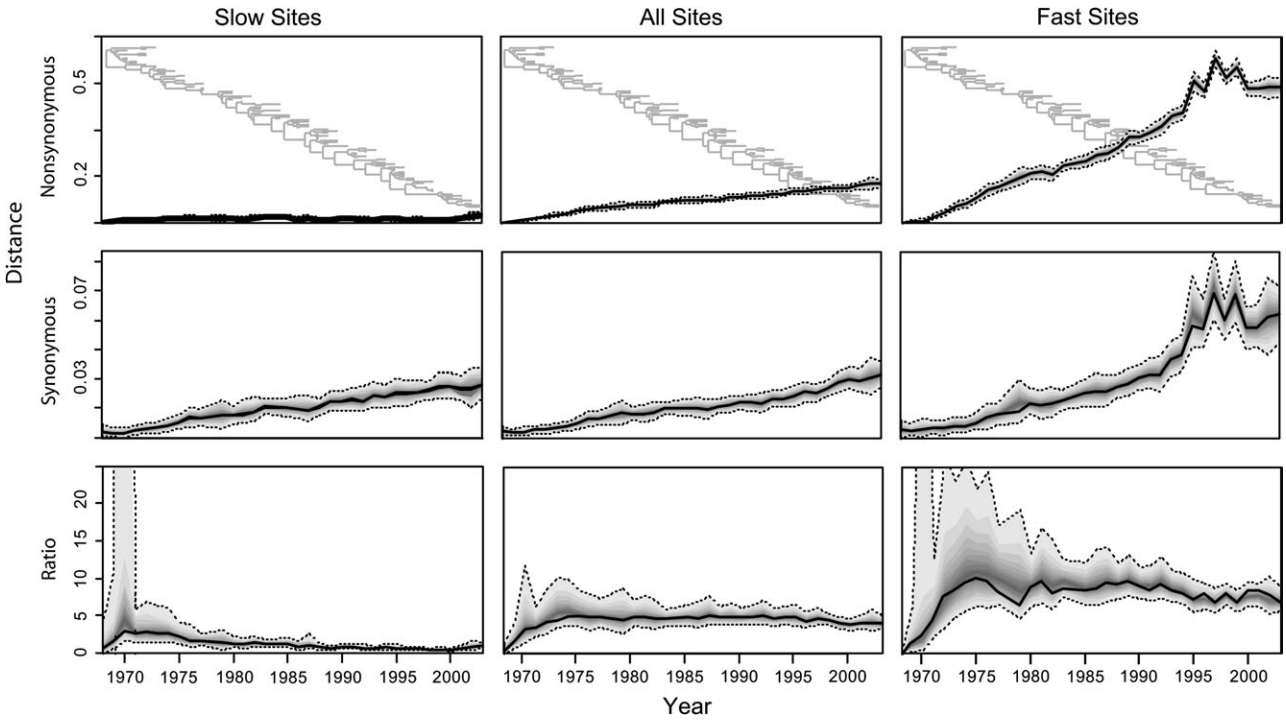

FIG. 4.—Time course of nonsynonymous distances (top), synonymous distances (middle), and their ratio (bottom) of influenza A H3N2 hemagglutinin sequences from 1969 to 2003. We report confidence intervals computed through 2,000 bootstrap iterations. The dark lines denote observed values. The lower and upper dashed lines trace 5% and 95% quantiles. Each gradation in gray reflects a 10% change in confidence. We report distances in substitutions per site and embed a time-scaled phylogenetic tree, constructed via TREBLE (Yang et al. 2007), in the top row of plots.

et al. 1997). The number of nonsynonymous substitutions along the backbone is in excess relative to the number expected by a neutral model of codon substitution, indicating strong positive selection (Bush et al. 1999; Plotkin et al. 2002). A phylogenetic reconstruction of HA0 influenza subtype H3N2 depicts this backbone clearly (fig. 4, embedded in the first row of plots). In this example, we illustrate how robust counting can aid in studying sequence evolution along phylogenetic backbones. Our data comprise 96 nt sequences of the influenza A subtype H3N2 HA0 gene, sampled over 36 years from 1968 until 2003, downloaded from the National Center for Biotechnology Information Influenza Database (Bao et al. 2007). GenBank accession numbers are available upon request. To generate a codon alignment, we first convert the nucleotide sequences into their corresponding amino acid sequences using Jalview (Clamp et al. 2004). We then align the amino acid sequences using Clustal (Li 2003). RevTrans (Wernersson and Pedersen 2003) finally imposes the amino acid alignment on the nucleotide sequences. To ensure consistency across columns in the codon alignment, we remove the final columns for which some sequences have no nucleotide characters, resulting in a final alignment of 972 nt (324 codons).

To count the synonymous and nonsynonymous substitutions occurring along the phylogenetic backbone, we take the earliest sequence–Hong Kong, 1968–to represent the origin and examine pairwise distances from the origin to each of the subsequent sequences, ordered by ascending sampling time. This is an approximation as sequence samples do not lie along the backbone. However, all samples are proximal to the backbone, making the approximation reasonable. We employ nonsynonymous labeling set $\mathcal{L}_N$ as in the simulation study and synonymous labeling set $\mathcal{L}_S$ similarly defined. We pool the pairwise labeled distance estimates for all sequences within a given year. The number of sequences within each year varies, from none for 1979 to five for 2003, with the median number at three sequences per year. To quantify uncertainty in all estimates, we perform a bootstrap on the alignment with 2,000 iterations (Efron 1979).

To capture potential variation within the alignment in the number and ratio of synonymous and nonsynonymous mutations between rapidly evolving sites and slowly evolving sites, we generate two sub-alignments using consensus scores: 1) a "fast" evolving set of 113 codons and 2) a "slow" evolving set of 211 codons. In inferring the number of codon substitutions for the full and sub-alignments, we employ a product composition of F84 nucleotide substitution models fit to the full alignment but modify equation (11) to only count sites within a particular (sub-)alignment. We compare the full, fast, and slow alignments in terms of nonsynonymous and synonymous distances and their ratio as functions of time. The ratio furnishes an informative portrait of the patterns and tempo of selection in HA0 evolution. In considering this ratio, we note that it is distinct from the quantity ω often employed in the literature, although the quantities should be highly correlated.

In the full alignment, we observe clear trends that indicate that certain codon sites experience strong positive selection, whereas other sites simultaneously experience strong negative selection (fig. 4). We see a strong linear trend in time, both for nonsynonymous and synonymous

distances, with a small variance distributed symmetrically about the observed values. This indicates a regular fixation of amino acid changes in time along the hemagglutinin backbone. Considering the ratio of the distances, we see a sharp initial increase terminating in 1975, followed by a long plateau. In the flat regime, we note that the mean value of the ratio is slightly less than 5, suggesting a strong excess of nonsynonymous as compared with synonymous substitutions. The flatness of the curve in time suggests that the rates of fixation of synonymous and nonsynonymous substitutions are highly correlated in time.

In the slow alignment, we find a strikingly different pattern across time. Whereas the synonymous distances' course is nearly identical to that observed in the full alignment, the nonsynonymous distances do not discernibly increase with time. Because we do indeed observe some nonsynonymous substitutions in the alignment, this indicates that these mutations are fixing at a negligible rate. Looking at the ratio of the nonsynonymous to synonymous distances, we find an initial spike due to a paucity of synonymous substitutions, trending downward to less than 1 as synonymous distances increase with time. This ratio is markedly lower than that observed for the full alignment, reflecting fewer nonsynonymous per synonymous substitutions at sites with more conserved amino acids.

Finally, considering the fast alignment, we observe a distinct pattern from the previous alignments. We again see a linear increase in the nonsynonymous distances with time, as in the full alignment. The magnitude of the increase is approximately three times that for the full alignment, suggesting that the fast alignment contains nearly all the fixed amino acid changes. Interestingly, years 1995, 1997, and 1999 exhibit three alternating peaks above the linear trend. Examining the relevant portions of the alignment, we find that these peaks are due to a high number of correlated amino acid substitutions (ca. 5) during those years that do not appear in later taxa. In part, the alternating pattern is likely due to binning of distances into years as we see many of the same amino acid changes within each peak. Looking at the synonymous distances, we observe a similar pattern to that for the nonsynonymous distances. As we do not observe any structural changes in the alignment, this close correlation suggests that the rate of synonymous substitution relates closely to that for nonsynonymous substitution at these sites. Considering the ratio of the distances, we find an initial peak followed by a relatively flat line at approximately 8, higher than that observed in both the full and slow alignments.

Viewed jointly, the full, fast, and slow alignments provide a coherent picture of the evolution along the phylogenetic backbone. In all alignments, we observe a strong linear time trend for synonymous distances, indicating the gradual accumulation of such substitutions across the gene. This suggest that the synonymous substitution process is approximately uniform across sites and independent of the nonsynonymous process. Independence advocates that changes in selective pressures, viewed here through the ratio of nonsynonymous to synonymous distances, occur primarily through alterations in the rate of fixation of nonsynonymous changes at certain codon sites (Cox and Bender 1995, see discussion therein).

Volatility Evolution of Influenza Hemagglutinin

Robust counting provides a convenient and versatile framework for estimating distances under arbitrary labeling sets. Whereas, by far, the most commonly used labelings identify nonsynonymous and synonymous changes, others of biological interest exist and are now conveniently approachable through robust counting. We construct a novel labeling set to explore changes in codon volatility (Plotkin and Dushoff 2003). The volatility of a given codon equals the number of potential single nucleotide changes that yield nonsynonymous amino acid changes to that codon. Plotkin and Dushoff (2003) conjecture that sites experiencing strong positive selection correlate with high volatility scores and, as such, volatility is an evolved property of a protein's nucleotide code.

To examine this codon volatility hypothesis in the context of HA0 evolution, we focus on a set of codons composing the hemagglutinin antibody interaction sites that previous research indicates experience strong positive selection (Wilson and Cox 1990; Bush et al. 1999). This set comprises 25 sites that we call the "epitope" codons, although they include only a fraction of all epitope sites. We ask if the distribution of volatility changes within epitope codons is the same as elsewhere; rejecting this limited hypothesis lends support to the more general volatility hypothesis.

First, we establish a measure of volatility change, d_V, by robust counting, through a labeling set that contains (i, j) if there is a change in volatility between those codons. Stop and start codons are excluded. Next, we consider the influenza alignment developed in our previous example, excluding the 140 codon sites that experience no amino acid changes over the 96 sequences. Taking the earliest sequence (Hong Kong 1968) to represent the origin, we estimate the pairwise labeled distances $\hat{d}_V$ between epitope codons from the origin to each of the remaining sequences ordered by ascending sampling time. To draw inference, we resample 25 codon sites without replacement from the reduced alignment 1,000 times and reestimate $\hat{d}_V^*$. We then compare over time the observed $\hat{d}_V$ to the distribution of $\hat{d}_V^*$ under the null hypothesis.

Figure 5 reports the average within each year of the proportion of resamples with distances more extreme than the observed epitope distance. The time series supports the volatility hypothesis. For the first several years (1968–1972), when little sequence change has occurred, we observe a high proportion of resamples that have higher volatility distance than those for the epitope codons. Looking directly at the codon alignment converted to volatility, we see only a small number of volatility changes, all at nonepitope codons. Subsequently, we see a rapid fall off in the fraction, indicating that the epitope codons experience significantly more volatility substitutions than other evolving sites. We observe a reversal of this trend in 1982, where the curve spikes dramatically. Examining the volatility alignment changes within this year we find two simultaneous progressions: a reversion to volatility states at the origin for some epitope codon sites, decreasing $\hat{d}_V$ for the set, and a large number of volatility changes at nonepitope sites, increasing the fraction of resamples with $\hat{d}_V^*$ higher than that

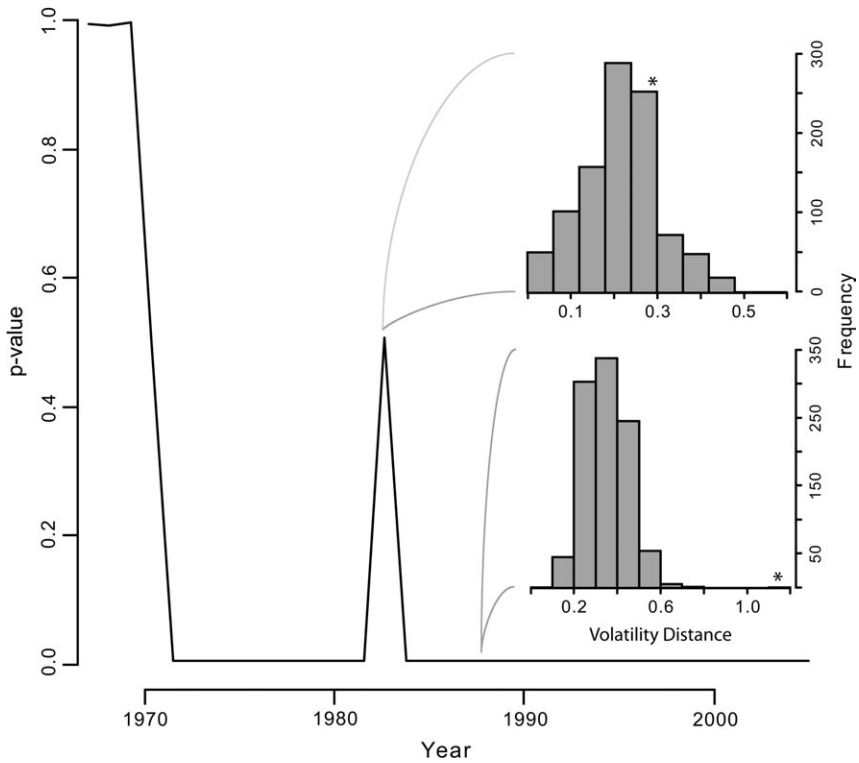

FIG. 5.—Time series plot of P_{value}, the fraction of random codon sets that have a more extreme estimated volatility change distance $\hat{d}_V$ than that for the epitope codon set. For each year, we average the fraction across all sequences within that year. Insets show $\hat{d}_V$ distributions of random sets for a special case within year (1982, top) and within a typical year (1988, bottom). Asterisks mark observed $\hat{d}_V$ value for epitope codons.

for the epitope sites. These two processes progress within the year but do not continue to subsequent years. As 1982 corresponds to a change in antigenic type, we conjecture that sequence changes within antigenic type are a scale of interest in understanding the structure of volatility.

The insets of figure 5 show histograms for a typical resample and for the special case within 1982. In the typical case, we see that the epitope codons give a volatility distance that departs substantially from the distribution of resampled distances. The volatility distance for epitope codons is close to one, indicating that every site in the antigenic epitope experiences change in volatility during this time interval. Given that the epitope codons were selected based on their positive selective pressure that they show such markedly higher volatility distance over the time course suggests that positive selection and volatility change are strongly related for hemagglutinin.

Convergent Evolution in an HIV Transmission Network

In this final example, we demonstrate how robust synonymous distances can be used to solve a convergent evolution problem, frequently faced during evolutionary analyses of molecular sequences under positive selection. Convergent evolution can be observed when selection drives molecular sequences to the same state, making some sequences appear more closely related than they really are (Doolittle 1994). The convergent evolution problem can be sometimes overcome by excluding the first two codon positions from phylogenetic inference. Mutations at the third codon position are almost always synonymous due to the genetic code redundancy. Therefore, using only the third codon positions, one diminishes the effect of nonsynonymous mutations on the phylogenetic reconstruction. Assuming that selection acts only at the amino acid level, phylogenetic inference based only on synonymous mutations should be free of the convergent evolution bias. Unfortunately, removing two-thirds of the data often leaves little information to reconstruct phylogenies. Similarly to the first two codon position removal approach, one can eliminate influence of nonsynonymous mutations on the distance-based phylogenetic reconstruction by considering only synonymous distances. Advantageously, this distance-based approach uses information more efficiently by not discarding informative synonymous mutations at the first two codon positions.

Lemey et al. (2005) encounter the convergent evolution problem in their analysis of a known HIV transmission network. The authors collect 16 *pol* and *env* gene sequences from 9 HIV+ individuals and find that their maximum likelihood and Bayesian analyses of the *pol* sequences produce phylogenetic tree estimates that disagree with the transmission history. In contrast, the phylogenies of the *env* coding region show no conflict with the transmission network information. Lemey et al. (2005) hypothesize that convergent evolution, driven by antiviral drug resistance-driven selective pressure, may bias phylogenetic estimation in the *pol* region. We revisit this problem to illustrate how synonymous distances can be used to reduce this bias. Starting with

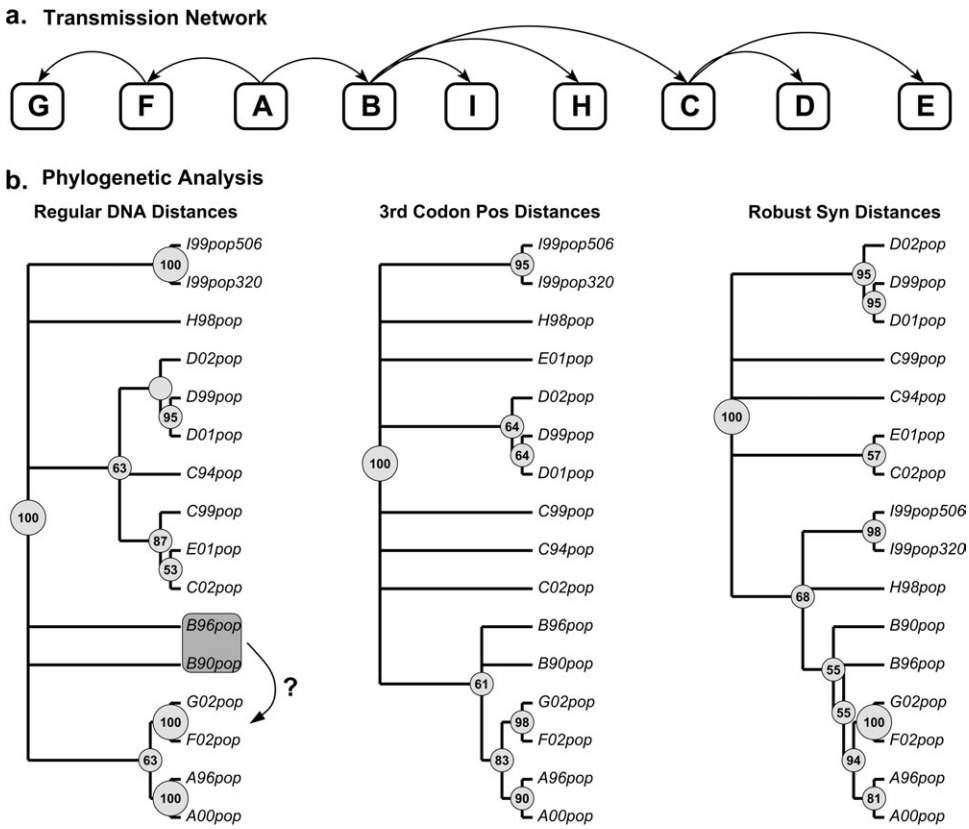

FIG. 6.—HIV transmission network. (*a*) Illustrates the HIV transmission network as a directed graph with nodes corresponding to the HIV+ patients. (*b*) Provides the results of our distance-based phylogenetic analysis. All three phylogenetic trees are majority consensus trees with bootstrap clade probabilities shown in light gray circles. The first tree (left) is a result of applying Neighbor-Joining to the conventional F84 DNA distances. The middle tree is constructed with the help of the conventional F84 DNA distances using only third codon positions. The last tree (right) is estimated using our robust synonymous distances. The dark gray box highlights the patient B clade that is erroneously grouped with the G and F sequences in the original study by Lemey et al. (2005).

a codon-based alignment of 16 *pol* sequences, we use a Neighbor-Joining algorithm (Saitou and Nei 1987) to reconstruct phylogenies of the *pol* sequences under three different distance measures. First, we estimate regular DNA distances using the F84 model. Next, we remove the first two codon positions from our alignment and reconstruct F84 pairwise distances based only on third codon positions. We obtain our third distance measure by calculating the robust synonymous distances between all pairs of the *pol* sequences using the same product composition of F84 models that we use in the simulations.

Figure 6*a* demonstrates the transmission network of the 9 HIV+ individuals, labeled as A, B, . . ., I. We show the majority rule consensus trees reconstructed under the three distance measures and over 1,000 bootstrap iterations. During the calculation of bootstrap support with the third codon position distances and robust synonymous distances, we resample codons rather than DNA sites of the alignment. The bootstrap support probabilities for each bipartition are shown in circles. Notice that we infer unrooted trees and arbitrarily root them for the sake of presentation. The branch lengths are purposefully not drawn to scale in order to align the tip labels. The first letter of each tip label corresponds to an individual's label.

In their likelihood-based phylogenetic analysis, Lemey et al. (2005) observe clustering of individual B's sequences with the G–F clade. This clustering suggests an HIV transmission between individual B and either individual F or G, creating an inconsistency with the transmission network. Interestingly, the consensus tree based on regular DNA distances correctly places the B sequences. However, the low bootstrap support of the node that separates individuals A, G, and F from everyone else indicates that many trees still support the incorrect placement of the individual B sequences. The support at this node significantly increases in the tree reconstructed from the third codon position distances. However, this reduction in bias comes at the expense of losing resolution in other parts of the tree. For example, the third codon position tree cannot resolve the clade of individual C sequences and has low support for the individual D's clade. The tree, reconstructed using robust synonymous distances, also successfully recovers the A–G–F clade. Similarly to the third codon position analysis, removing phylogenetic information provided by nonsynonymous mutations results in increased uncertainty in the inferred phylogeny. However, the loss of resolution is less severe in the robust synonymous tree than in the third codon position tree. For example, the individual D's sequence clade is well resolved using the robust synonymous distances.

Lemey et al. (2005) also resort to distance methods to study the effect of convergent evolution bias on the

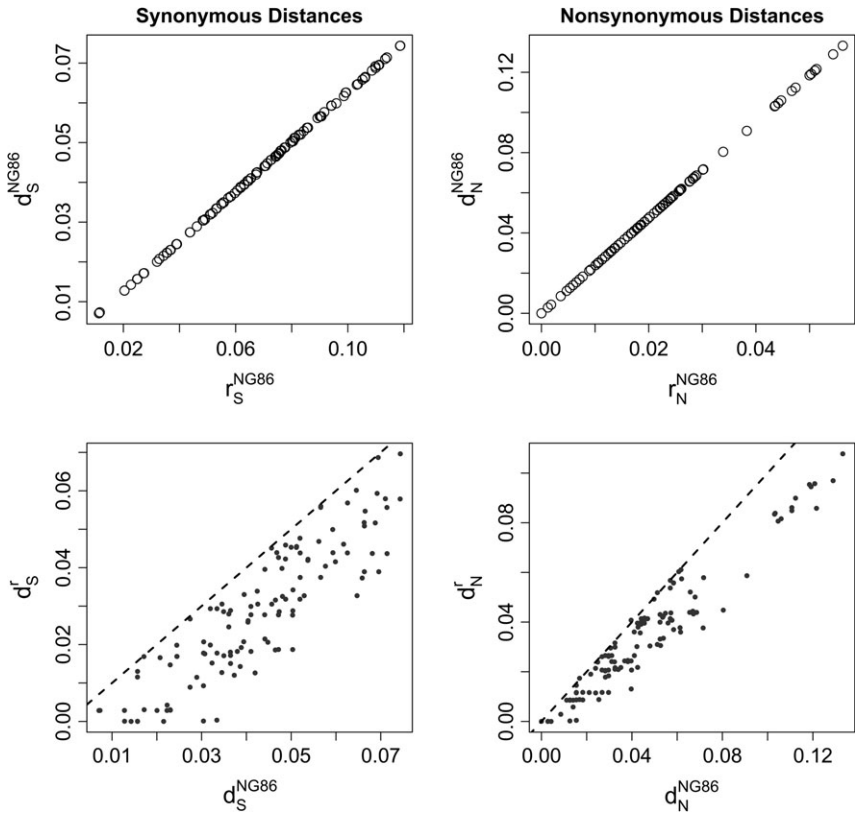

FIG. 7.—Analysis of Syn-SCAN distances. In the top row, we plot Syn-SCAN-estimated r_S^{NG86} and r_N^{NG86} values against expected number of synonymous and nonsynonymous mutations, respectively. The plots in the bottom row compare Syn-SCAN and robust synonymous and nonsynonymous distances. The dashed line in these plots corresponds to the $y = x$ line.

phylogenetic inference in the *pol* gene. They estimate $\hat{r}_N^{NG86}$ and $\hat{r}_S^{NG86}$ values using Syn-SCAN (Gonzales et al. 2002) and use these pairwise distance measures to reconstruct Neighbor-Joining trees of 16 *pol* sequences. The authors do not perform a bootstrap analysis to assess the uncertainty of their estimates. More importantly, it is not clear that $\hat{r}_N^{NG86}$ and $\hat{r}_S^{NG86}$ values are appropriate distance measures. As we mention earlier in the text, the expected number of synonymous and nonsynonymous mutations per codon site can be obtained from $\hat{r}_N^{NG86}$ and $\hat{r}_S^{NG86}$ via possibly nonlinear transformations (24). The nonlinearity of these transformations depends critically on the behavior of the potential synonymous and nonsynonymous counts $\hat{S}$ and $\hat{N}$ as functions of pairwise alignments.

Using the HIV transmission network sequences, we plot all possible pairwise $\hat{r}_N^{NG86}$ and $\hat{r}_S^{NG86}$ values against $\hat{d}_N^{NG86}$ and $\hat{d}_S^{NG86}$ in the first row of figure 7 and find a nearly perfect linear relationship between these two types of synonymous and nonsynonymous distance measures. This linear relationship is produced by an approximate invariance of $\hat{N}$ and $\hat{S}$ estimates among different sequence pairs. As a result, the $\hat{r}_N^{NG86}$ and $\hat{r}_S^{NG86}$ distance matrices can be approximated by multiplying $\hat{d}_N^{NG86}$ and $\hat{d}_S^{NG86}$ matrices by a scalar. This is good news for researchers who use $\hat{r}_N$ and $\hat{r}_S$ values for distance-based phylogenetic reconstruction because linear transformation of a distance matrix does not affect phylogenetic relationships encoded by the distances. However, phylogenetic branch lengths have much more meaningful interpretation if the phylogeny is estimated using absolute expectations, $\hat{d}_N$ and $\hat{d}_S$. Despite the strong empirical evidence in favor of the approximate invariance of $\hat{N}$ and $\hat{S}$ among all pairs of a sequence alignment, we are not aware of a general theoretical result that would guarantee this behavior. In addition, several authors point out serious problems in estimation of the potential quantities $\hat{N}$ and $\hat{S}$ that could compromise heuristic calculations of both $\hat{r}_N$, $\hat{r}_S$ and $\hat{d}_N$, $\hat{d}_S$ values (Ina 1995; Muse 1996; Bierne and Eyre-Walker 2003).

Finally, we compare performance of the robust counting and Syn-SCAN synonymous and nonsynonymous distance estimation in the HIV example by plotting these estimates against each other in the bottom row of figure 7. The robust counting consistently predicts smaller expected number of synonymous and nonsynonymous mutations between the HIV sequences. There is a substantial number of nucleotide ambiguities in the HIV sequences. Our current implementation of robust counting treats ambiguities as missing values, whereas Syn-SCAN uses this additional information more carefully. Therefore, we do not expect a perfect agreement between the robust counting and Syn-SCAN even if the robust counting and the NG86 method perform comparably on these data. We plan to include more efficient treatment of nucleotide ambiguities in the next release of markovjumps.

Discussion

The novel framework we present in this paper integrates model-based prediction of the number of labeled substitutions between two molecular sequences with a model-free estimate of the pairwise site pattern distribution. This combination yields robust estimates of labeled distances between sequences. Providing substantial protection against the bias arising from model misspecification of the underlying CTMC, robust counting is an important step in making consistent and reliable estimates of sequence change for biological analysis.

When the underlying CTMC is misspecified, robust counting exchanges a significant reduction in estimation bias for a moderate increase in variance relative to conventional estimates under the same CTMC. This trade-off is typically advantageous for the robust estimator as it leads to better coverage properties in a large majority of model misspecifications we consider here. As expected, comparing robust estimates under a misspecified model against conventional estimates under the true, more complex model, we observe smaller variance but larger bias. However, MSE of these two estimators suggest that for realistic sample sizes, robust estimators under simple models compete well with conventional estimators based on complex models.

Assessing the bias and variance of robust counting is only one of many interesting theoretical problems that our new method raises. Perhaps, the most important open question that we have not even attempted to answer is "How robust is robust counting? " or, in other words, how severely can we misspecify a CTMC model of substitution and still hope to recover true labeled distances via robust counting? We should be able to answer this question quantitatively by characterizing the dependence of conventional and robust distance estimator bias in terms of an as-yet-undefined metric between the misspecified and the true substitution model. With a solid theoretical understanding of the robust counting bias behavior in hand, we will be able to analyze many further model misspecifications without time-consuming simulations.

In pursuing these properties of robust counting, we stress that our procedure has very little in common with the standard robust estimation applied to generalized linear models (Huber 1981). Therefore, theoretical studies may find advantage in viewing robust counting as a semiparametric procedure, where nonparametric estimation of site pattern frequencies is combined with fully parametric estimates of conditional expectations of the number of labeled transitions. Clearly, the robust and semiparametric facets of robust counting do not contradict, but rather complement, each other.

Afforded by some simple models of evolution, computational efficiency is a very important property of any distance estimation procedure, constituting the primary advantage of distance-based phylogenetic reconstruction over full likelihood-based approaches. Fortunately, robust counting inherits its computational efficiency from conventional distance estimation, additionally requiring only the very rapid calculation of conditional expectations (Holmes and Rubin 2002; Minin and Suchard 2008). We emphasize that working with simple nucleotide models with approximate, closed-form expressions for the generator MLEs, such as the F84 model, keeps labeled distance estimation computationally inexpensive. For example, estimating the synonymous distance between two HIV sequences in our last example takes 0.08 s with an F84-based codon model coupled with our robust estimation. Conventional distance estimation requires fitting an M0 codon model and runs 83 s. Although our suboptimal implementation of M0 parameter estimation may partially explain the drastic difference between running times of conventional and robust procedures, this example clearly illustrates the advantage of avoiding numerical likelihood optimization. Understandably, we find that for labeled codon distances, robust counting compares fairly in speed with the established methods NG86 and YN00.

Although we concentrated on the nucleotide and codon state spaces in this paper, the generality of robust counting grants biological researchers freedom to fashion their own labeled distances for any discrete evolutionary trait, specific to their problem of interest. For example, we envision using robust counting for computing distances between molecular sequences measured in expected number of transitions between predefined amino acid classes, such as hydrophobic and hydrophilic amino acids. Clearly, protection against model misspecification will be of great advantage in these ambitious attempts to fill in the missing details of evolutionary history–relating sequences.

Acknowledgments

We thank Phillipe Lemey for his HIV alignments and Simon Whelan for his title suggestion. O'Brien was supported by a UCLA Dissertation Year Fellowship and an National Institute of General Medical Sciences Systems and Integrative Biology Training Grant. Suchard is supported by an Alfred P. Sloan Research Fellowship, a John Simon Guggenheim Memorial Fellowship and the National Institutes of Health R01 GM086887.

Literature Cited

Ball F, Milne R. 2005. Simple derivations of properties of counting processes associated with Markov renewal processes. J Appl Probab. 42(4):1031–1043.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Tatusova T. 2007. FLAN: a web server for influenza virus genome annotation. Nucleic Acids Res. 35(2 Suppl):280–284.

Bierne N, Eyre-Walker A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics. 165(3):1587–1597.

Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc Natl Acad Sci USA. 105(23):7899–7906.

Buckley TR, Simon C, Chambers GK. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst Biol. 50(1):67–86.

Bush R, Bender CA, Subbarao K, Cox NJ, Fitch WM. 1999. Predicting the evolution of human influenza A. Science. 286(5446):1921–1925.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. Bioinformatics. 20(3):426–427.
Cox NJ, Bender CA. 1995. The molecular epidemiology of influenza viruses. Semin Virol. 6(6):359–370.
Doolittle RF. 1994. Convergent evolution: the need to be explicit. Trends Biochem Sci. 19(1):15–18.
Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. Mol Biol Evol. 24(2):388–397.
Earn D, Dushoff J, Levin S. 2002. Ecology and evolution of the flu. Trends Ecol Evol. 17(7):334–340.
Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann Stat. 7(1):1–26.
Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.
Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA. 94(15):7712–7718.
Gascuel O. 2005. Mathematics of evolution and phylogeny. New York: Oxford University Press.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11(5):725–736.
Gonzales M, Dugan J, Shafer R. 2002. Synonymous-non-synonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN). Bioinformatics. 18(6): 886–887.
Goodstadt L, Ponting CP. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol. 2(3):e133.
Gu X, Li WH. 1996. A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc Natl Acad Sci USA. 93(10):4671–4676.
Hasegawa M, Kishino H, Yano TA. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2):160–174.
Hobolth A, Jensen JL. 2005. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Stat Appl Genet Mol Biol. 4(1):Article 18.
Holmes I, Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. J Mol Biol. 317(5):753–764.
Huber PJ. 1981. Robust statistics. New York: Wiley.
Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol. 40(2):190–226.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. New York: Academic Press. p. 21–32.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16(2):111–120.
Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol. 29(2):170–179.
Lemey P, Derdelinckx I, Rambaut A, Laethem KV, Dumont S, Vermeulen S, Wijngaerden EV, Vandamme AM. 2005. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol. 79(18): 11981–11989.
Li KB. 2003. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics. 19(12):1585–1586.
Li WH, Gu X. 1995. Statistical models for studying DNA sequence evolution. Physica A. 221(3):159–167.
Lio P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res. 8(12):1233–1244.
Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol. 56(3):391–412.
Muse S. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol Biol Evol. 13(1):105–114.
Muse S, Gaut B. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11(5):715–724.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(1):418–426.
Neuts M. 1995. Algorithmic probability: a collection of problems. London: Chapman and Hall.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148(3):929–936.
Plotkin J, Dushoff J, Levin SA. 2002. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA. 99(9):6263–6268.
Plotkin JB, Dushoff J. 2003. Codon-bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci USA. 100(12):7152–7157.
Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol. 12(1):131–151.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406–425.
Schneider A, Gonnet G, Cannarozzi G. 2007. SynPAM: a distance measure based on synonymous codon substitutions. IEEE/ACM Trans Comput Biol Bioinform. 4(4):553–560.
Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol. 18(6):1001–1013.
Sullivan J, Joyce P. 2005. Model selection in phylogenetics. Annual review of ecology. Evol Syst. 36(1):446–466.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10(3):512–526.
Wernersson R, Pedersen AG. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31(13):3537–3539.
Wilson I, Cox N. 1990. Structural basis of immune recognition of influenza virus hemagglutinin. Annu Rev Immunol. 8(1):737–787.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39(3):306–314.
Yang Z. 1997. How often do wrong models produce better phylogenies? Mol Biol Evol. 14(1):105–108.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Yang Z, Nielsen R. 2000. Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17(1):32–43.
Yang Z, O'Brien JD, Zheng XB, Zhu HQ, She ZS. 2007. Tree and rate estimation by local evaluation of heterochronous data. Bioinformatics. 23(2):169–176.

Asger Hobolth, Associate Editor

Accepted December 30, 2008