

Exploring Variation in the d_N/d_S Ratio Among Sites and Lineages Using Mutational Mappings: Applications to the Influenza Virus

Weiwei Zhai · Montgomery Slatkin ·
Rasmus Nielsen

Received: 20 March 2007 / Accepted: 22 July 2007 / Published online: 11 September 2007
© Springer Science+Business Media, LLC 2007

Abstract We use a likelihood-based method for mapping mutations on a phylogeny in a way that allows for both site-specific and lineage-specific variation in selection intensity. The method accounts for many of the potential sources of bias encountered in mapping of mutations on trees while still being computationally efficient. We apply the method to a previously published influenza data set to investigate hypotheses about changes in selection intensity in influenza strains. Influenza virus is sometimes propagated in chicken cells for several generations before sequencing, a process that has been hypothesized to induce mutations adapting the virus to the lab medium. Our analysis suggests that there are approximately twice as many replacement substitutions in lineages propagated in chicken eggs as in lineages that are not. Previous studies have attempted to predict which viral strains future epidemics may arise from using inferences regarding positive selection. The assumption is that future epidemics are more likely to arise from the strains in which positive selection on the so-called “trunk lineages” of the evolutionary tree is most pervasive. However, we find no difference in the

strength of selection in the trunk lineages versus other evolutionary lineages. Our results suggest that it may be more difficult to use inferences regarding the strength of selection on mutations to make predictions regarding viral epidemics than previously thought.

Keywords Mutational mapping · Positive selection · Lineage variation · Influenza · Host-mediated mutations

Introduction

Codon-based likelihood models are commonly used in studies of molecular evolution (Yang and Bielawski 2000). By employing a likelihood ratio test, such methods can detect positive selection in the form of an elevated rate ratio of nonsynonymous-to-synonymous substitutions (d_N/d_S) in specific genes or specific lineages (Nielsen and Yang 1998; Yang and Nielsen 2002; Yang et al. 2000). Once positive selection is detected, Bayesian methods can be used to identify specific sites that are under positive selection (Nielsen and Yang 1998; Yang et al. 2000, 2005).

Most of the available computational methods assume either that intensity of selection varies among lineages but is the same for different codon positions or that selection varies among amino acid sites but is the same on different lineages. However, selection may vary both among lineages (Alba et al. 2000; Guindon et al. 2004; Huttley et al. 2000; Stewart et al. 1987) and among sites (e.g., Nielsen and Yang 1998; Yang et al. 2000). Yang and Nielsen (2002) suggested a method for allowing variation both among sites and among lineages, but this method requires that the set of lineages potentially undergoing positive selection to be specified *a priori*, which restricts its utility for exploratory data analysis. A more versatile method built

Reviewing Editor: Dr. Willie Swanson

Electronic supplementary material The online version of this article (doi:10.1007/s00239-007-9019-7) contains supplementary material, which is available to authorized users.

W. Zhai (✉) · M. Slatkin
Department of Integrative Biology,
University of California Berkeley, 3060 Valley Life Science
Building, Berkeley, CA 94720-3140, USA
e-mail: weiweizhai@berkeley.edu

R. Nielsen
Centre for Comparative Genomics, University of Copenhagen,
Universitetsparken 15, 2100 Kbh Ø, Copenhagen, Denmark

on codon-based models is that of Guindon et al. (2004), which allows for variation in selection intensity among both lineages and sites. However, the method will be computationally demanding for large data sets and is difficult to adapt to exploratory analysis, especially in comparing and quantifying the amount of changes in different parts of the tree, as we demonstrate in this study.

Inferences on the character history, especially the locations of substitutions on a phylogeny, have a long history in molecular evolution. In most applications, parsimony-based mappings have been used to infer the history of character changes (e.g., Bush et al. 1999b; Dayhoff et al. 1978; Suzuki and Gojobori 1999). One of the challenges with the parsimony approach is that uncertainties in the mapping of mutations are not taken into account. Analyses based on a single inferred mutational history can lead to problems such as consistent underestimation of the number of mutations (Nielsen 2002; Whelan and Goldman 2001). Moreover, parsimony-based inferences can be affected by the specifics of the algorithm used (Swofford and Maddison 1987).

The approach presented here is based on obtaining a sample of mappings of mutations from the posterior distribution and basing inferences on this sample (Nielsen 2001, 2002). To illustrate the utility of our method, we apply it to a previous published data set of hemagglutinin gene sequences from influenza H3N2.

The hemagglutinin protein is the major surface antigen on the viral lipid membrane. It is the critical protein involved in binding to sialic acid-containing receptors on the host cell surface. Once the virus enters the host cell through endocytosis, conformational changes in hemagglutinin facilitate the fusion of the endosomal membrane with the viral membrane, releasing the viral genome into the cytoplasm. Selection may, therefore, act on the genes as it adapts to its host. However, since the hemagglutinin protein is exposed on the surface of the viral particle, selection may also arise from pressures to avoid immune recognition.

DNA sequences of the hemagglutinin gene have been sampled at different points of time and in different geographic locations (Macken et al. 2001). Because sequences have been sampled at different points in time, the evolutionary tree relating the sequences has a characteristic shape often described as “cactus-like” (Fitch et al. 1991; Nelson and Holmes 2007), with short side branches emerging from a main trunk lineage. Since the trunk lineages are representatives of the strains that survive between years, it has been hypothesized that substitutions occurring on the trunk lineages are functionally important and presumably more likely to be adaptive. In contrast, the side branches may represent isolates that are not sufficiently antigenically novel and die out due to herd (population) immunity. There have been considerable interests in comparing the

molecular evolution in trunk lineages and side branches, especially for the purpose of developing strategies for prediction of future epidemic strains (Bush et al. 1999a; Lee and Chen 2004; Nelson and Holmes 2007).

Another characteristic of the sequences is that viral specimens usually experience several rounds of propagation in laboratory culture before direct sequencing. For influenza viruses, chicken eggs and various types of cell lines, especially Madin Darby Canine Kidney (MDCK), are often used (Macken et al. 2001; Meguro et al. 1979). Host-mediated mutations, meaning point mutations that accumulate during cultivation and are not present in the original clinical specimens, are often found. Sequence comparisons indicate that egg-grown viruses often have additional amino acid substitutions not present in cell lines or the original specimens (Robertson 1993). These additional substitutions are thought to result from selection acting to increase the affinity of the viral hemagglutinin to the NeuAc(α 2-3)Gal receptor form commonly found in chicken eggs (allantoic cells) but absent in humans (Ito 2000; Ito et al. 1997). Since the viral sequences available are typically “contaminated” with host-mediated substitutions (Cao et al. 1995; Graff et al. 1994; Itoh et al. 1997; Sawyer et al. 1994), efforts to detect positive selection that do not take account of this effect may give misleading results (Bush 2004). In this study, we try to quantify this level of contamination and address some problems associated with viral phylogenetic analysis.

Methods

Inferences on Character History and Mutational Mapping

Let D be the observed DNA sequence data and M be a particular mapping of mutations (character history) on a phylogeny. The marginal posterior distribution of mutational mapping is then given by

$$p(M|D) = \int_{\theta \in \Omega} \Pr(M|D, \theta) \Pr(\theta|D) d\theta \quad (1)$$

where θ is a vector of nuisance parameters, which includes the branch lengths, topology of the phylogeny, and parameters associated with the mutational process, and Ω is the sample space of θ .

Nielsen (2002) described a two-step algorithm for obtaining a sample of mutational mapping from $p(M|D, \theta)$ given θ . First, the nucleotide states of all internal nodes are simulated recursively from the root of the tree according to their joint probabilities. Second, conditional on the character states at all nodes, the mutational history is simulated according to the mutational process described by θ .

The general method according to Eq. (1) can be implemented as an algorithm that simulates θ (e.g., phylogenies) and mutational mappings from their joint posterior distribution using Markov chain Monte Carlo (MCMC) (Nielsen 2001). Alternatively, MrBayes (Huelsenbeck and Ronquist 2001) or a similar program can be used to sample θ from the marginal posterior density, after which mutational mappings can be sampled using the simulated values of θ (Nielsen 2002; Nielsen and Huelsenbeck 2002). Although MCMC methods are attractive because they take uncertainty in the parameter estimates into account, they can be computationally very slow because of the need for an MCMC algorithm, and it can at times be difficult to determine whether convergence criteria are satisfactorily met.

In this study, the tree topology is instead inferred using the neighbor-joining algorithm (Saitou and Nei 1987) and rooted using an outgroup sequence from an early year. Conditional on the topology, maximum likelihood estimates of the associated branch lengths, mutational matrices under a General Time Reversible Model (Lanave et al. 1984), base frequencies, and rate-variation parameters (the α parameter of the gamma distribution assumed for rate variation) are obtained using baseml from PAML 3.15 with partitions on the first/second/third codon positions (Yang 1997). Although this approach ignores the uncertainty of the tree topology, it is much faster computationally than previous approaches and can be applied to very large data sets. Moreover, alternative tree topologies obtained using other optimality criteria produced almost-identical results (data not shown).

Calculating Selection Intensity for Each Lineage

Given the maximum likelihood estimates of all parameters, a sample of mutational mappings is drawn independently for each codon position using position specific rate matrices. Mutational histories involving stop codons are discarded and resampled. For each replicate, the substitution rate for each nucleotide position is then drawn from the inferred gamma distribution of rates, independently for the first, second, and third codon positions. Samples of mutational mappings from $p(M|D, \theta)$ are then obtained as described by Nielsen (2002). For each mutational mapping, the ratio of nonsynonymous-to-synonymous substitution rates (d_N/d_S) for each lineage is then calculated by the nonsynonymous and synonymous sites determined from the ancestral sequence of that lineage. The d_N/d_S ratio is not estimated for lineages for which this ratio is not defined (i.e., no synonymous changes are inferred or the expected number of synonymous sites is zero). The final results are averaged over all replicates. While this procedure does not

take all the complexities of codon-based evolution into account, it provides a very fast computational framework that incorporates uncertainty in the mapping of mutations

Tests of Positive Selection at Individual Sites

Under neutrality, the expectation of the ratio of the d_N/d_S is 1. At each replicate, after using our method to infer the number of nonsynonymous and synonymous substitutions, we use a one-sided binomial test of whether the observed d_N/d_S ratio exceeds 1, thus indicating positive selection. To be more precise, the expected ratio of nonsynonymous-to-synonymous sites (binomial p) is given by the inferred ancestral sequences of that replication. The p -value for that replicate condition on the total number of changes is then calculated as the sum of the probabilities of those configurations that have at least as many nonsynonymous changes as the inferred mutational history. The tests can be done for a particular codon site combining all lineages in the phylogeny, for a particular lineage combining all sites, or for any combination of sites and lineages. The p -values are then averaged among simulated mutational mappings to form a posterior predictive p -value (Bollback 2005; Nielsen and Huelsenbeck 2002). Similarly, tests of homogeneity of hypotheses regarding the distribution of mutations among lineages and sites are performed using standard chi-square tests applied to the mutations inferred from each mapping in exactly the same fashion as the McDonald-Kreitman (1991) test and averaging p -values across replicates. This approach takes uncertainty in the mutational mappings into account, while being both computationally efficient and easy to adapt to new problems.

Sequence Alignments and Passage History

We applied our method to the influenza (H3N2) hemagglutinin gene HA1 domain collected from the Influenza Sequence Database (ISD) located at Los Alamos National Laboratory (Macken et al. 2001). The data set consists of 350 sequences which have been analyzed in several previous studies (Bush et al. 1999b; Fitch et al. 1997). We omitted 7 of the 357 sequences because of ambiguous characters or missing passage history. The specimens were collected during the years 1983–1997 and later sequenced at the Centers for Disease Control and Prevention and contain all 329 codons (987 nucleotides). Sequence alignment was done in ClustalW 1.83 (Thompson et al. 1994). Sequences for which there was any level of passage in eggs are labeled “egg-grown.” Strains for which records indicate that they were grown only in cellular medium are labeled “cell-grown.” Strains with an unknown history are

labeled “unknown.” The 350 sequences include 143 cell-grown strains and 142 egg-grown strains.

Simulations

In order to evaluate the performance of the new method, we simulate DNA sequences using the Branch Site model in Evolver from the PAML package on an eight-taxon tree (Fig. 1). Branch lengths are all 0.1 substitution per codon, and 300 amino acid positions are generated with equal codon frequencies. The transition/transversion rate ratio (κ) is set to 2.0. Highlighted lineages are experiencing potentially different selective intensities from the background lineages which have $d_N/d_S (\omega) = 1.0$. Four combinations of selection intensities ($d_N/d_S = 0.5, 1.0, 1.5, 3.0$) at the foreground lineages are generated. The Branch Site model from the PAML package and the mutational mapping method are used to estimate d_N/d_S for the highlighted lineages. More extensive simulation results are explored in the supplementary materials but are not presented here because they provide results that are very similar to those presented here.

Results

Test of Positive Selection at Individual Sites

We found evidence of positive selection at 12 sites in the 329 codons at the 5% significance level, without a correction for multiple tests (Table 1). The 12 sites identified here all reside in either antigenic or receptor binding domains, with a strong overrepresentation of sites located in binding domains. One site of particular interest is site 226. Structural

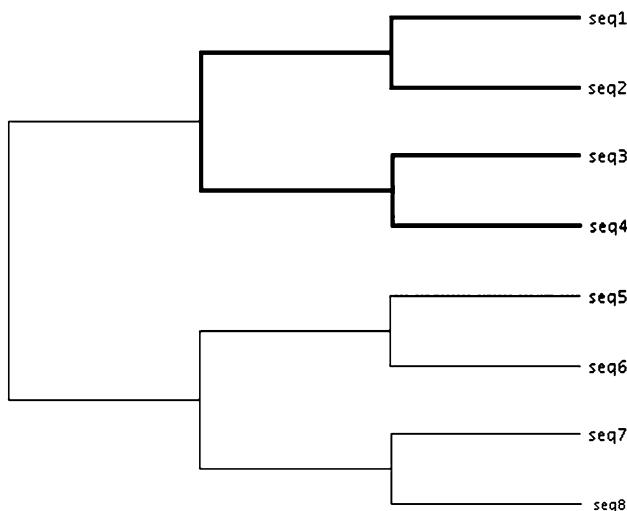


Fig. 1 The eight-taxon tree is used to check the performance of the mutational mapping method

studies have demonstrated the importance of this site in early adaptations as the H3N2 serotype switched from avian to human hosts (Rogers et al. 1983). Sialic acids are the major receptors on the surface of erythrocytes. Two forms of linkage between sialic acid and galactose residues exist in nature. Human erythrocytes carry mostly NeuAc(α 2-6)Gal, while NeuAc(α 2-3)Gal is commonly found in birds. A nonsynonymous mutation from glutamine to leucine causes a strong increase in affinity for NeuAc(α 2-6)Gal and decreased affinity for NeuAc(α 2-3)Gal. This substitution is thought to be one of the major contributors to the host switch of the H3N2 serotype from birds to humans in 1968 (Rogers et al. 1983). Structural studies indicate that site 226 does not come into direct contact with the sialic acid, but rather it alters the conformation of the receptor-binding pocket (Weis et al. 1988). The continuation of strong positive selection at this site even after the host switch might be due in part to selection for the success of membrane fusion and entry of the virus into the host cell, in addition to selection for increased efficiency of receptor binding (Robertson 1999; Skehel and Wiley 2000). Several other sites, such as 137, 138, 190, and 194, are also in the receptor-binding pocket, while sites such as 133, 156, and 193 are nearby (e.g., Skehel and Wiley 2000, Fig. 2).

Host-Mediated Mutations and Test of Contamination

Empirical studies reported that 22 codon positions can accumulate host-mediated substitutions (Bush et al. 2000;

Table 1 Test of positive selection at individual sites

Codon	Internal N/S	All N/S	<i>p</i> -value	Func
128	0/0	8/0	0.040*	B
133	9/0	14/0	0.017*	A
137	0/0	12/1	0.039*	A, R
138	6/0	22/0	0.0002**	A, R
145	11/0	22/0	0.004**	A
156	10/1	27/2	0.031*	B
190	4/0	14/0	0.027*	B, R
193	4/0	22/1	0.016*	B
194	1/0	13/2	0.007**	B, R
226	20/0	48/1	0.0000**	D, R
275	5/0	8/0	0.048*	C
276	6/0	13/0	0.018*	C

Note. “N/S” denotes average nonsynonymous-to-synonymous mutations. “Internal N/S” indicates that substitutions happened only in internal branches. “All N/S” includes all substitutions across the evolutionary tree. “*p*-value” is the average *p*-value over 10,000 replications. “Func” denotes whether it belongs to a certain functional domain(s): A–E represent the five antigenic domains, and R represents the binding domain

Gubareva et al. 1994; Hardy et al. 1995; Nakajima et al. 1983; Robertson 1993; Rocha et al. 1993). Our first goal is to examine the extent to which propagation of cells in chicken eggs has affected the estimated value of ω (d_N/d_S) in the 22 amino acid residues identified in previous studies to have accumulated host-mediated substitutions. We test whether estimates of ω for these 22 sites (which we will call host-mediated sites) are similar to estimates of ω for other sites on the terminal branches (i.e., edges in the tree connected to leaf nodes). In other words, we test the hypothesis

$$H_0 : \omega_{HM}^T = \omega_{HM^c}^T \text{ vs } H_A : \omega_{HM}^T \neq \omega_{HM^c}^T$$

where ω_{HM}^T and $\omega_{HM^c}^T$ are the ratios of nonsynonymous-to-synonymous substitution rates in terminal branches among the 22 host-mediated sites and the remaining sites, respectively. The test of homogeneity shows that on the terminal branches the 22 host-mediated sites show an excess of nonsynonymous mutations compared to the other sites (upper-left portion of Table 2).

We also tested whether the host-mediated sites show evidence of an excess of nonsynonymous mutations on internal lineages (i.e., edges in the tree not connected to a leaf node), which do not reflect evolution during laboratory cultivation:

$$H_0 : \omega_{HM}^I = \omega_{HM^c}^I \text{ vs } H_A : \omega_{HM}^I \neq \omega_{HM^c}^I$$

This test is also strongly significant (Table 2; upper-middle portion), suggesting that the apparent excess of

nonsynonymous substitutions at the host-mediated sites is not explained solely by the effect of laboratory cultivation. In fact, a test of homogeneity shows that the ratio of nonsynonymous-to-synonymous mutations is not significantly different at host-mediated sites in internal versus external lineages (Table 2; upper-right portion). To further examine the pattern of adaptation in terminal lineages, we also test whether, within the host-mediated lines, there is a higher ratio of nonsynonymous-to-synonymous substitutions on terminal branches of strains cultured in chicken eggs than in those passed through cell lines. In other words, we test the hypothesis

$$H_0 : \omega_{HM}^{T,Chicken} = \omega_{HM}^{T,Cell} \text{ vs } H_A : \omega_{HM}^{T,Chicken} \neq \omega_{HM}^{T,Cell}$$

The test is significant at the 5% level, which indicates that the selective process associated with laboratory propagation is stronger than in the human host (Table 2; lower-left portion). The extent of synonymous change in the two types of lineages is approximately the same, while the level of replacement change in lineages undergoing propagation in eggs is twice that observed for lineages passed through cell lines (Table 2; lower-left portion).

Lineage Variation in Selection Intensity and Influenza Epidemics

Following Fitch et al. (1997), we partition the lineages of the tree into three major categories: trunk, twig, and

Fig. 2 Estimation of Ω using the mutational mapping method is plotted on the x-axis against estimates from the codon-based likelihood methods from PAML

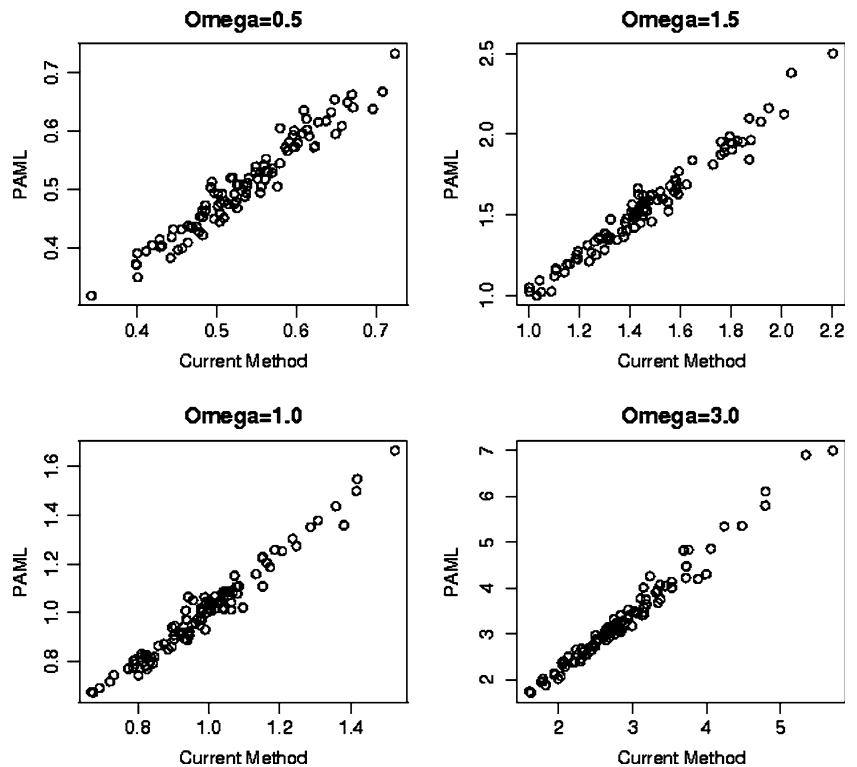


Table 2 Tests for lineage variations

Terminal	HM	Non-HM	<i>p</i> -value	Internal	HM	Non-HM	<i>p</i> -value	HM sites	Internal	Terminal	<i>p</i> -value
Nonsyn	183	341			95	151			95	183	
Syn	35	381	E-6**		17	188	E-6**		17	25	0.488
HM sites	Cell	Egg	<i>p</i> -value		All sites	Trunk	Twig	Terminal	<i>p</i> -value		
Nonsyn	48	103				54	192	524			
Syn	11	9	0.046*			45	159	407			0.786

Note. HM, host-mediated sites are those 22 reported sites (see text and supplementary materials); non-HM, non-host-mediated sites; nonsyn, nonsynonymous changes; syn, synonymous changes. *p*-values are two-sided. HM sites are also listed in the supplementary material. Total of 143 sequences of the cell type and 142 of the egg type. For classification please see Methods

terminal lineages (Fig. 1). The trunk of the tree is the set of lineages defining the path from the root node to the most distal tip group (Fitch et al. 1997). Lineages other than terminal or trunk lineages are the twigs. Previous studies (Bush et al. 1999a, b; Fitch et al. 1997) argued that the trunk can be interpreted as the lineage from which new epidemic strains arise each year. Twigs and terminal lineages represent viral strains that have gone extinct. Consequently, there is considerable interest in understanding differences in the evolution of trunk lineages and other lineages, especially in terms of predicting future strains (Bush et al. 1999a). We first test if there is any difference in selection intensity among the three types of lineages (trunk, terminal and twig). Using a chi-square test, we find no evidence that selection intensities in trunks, twigs, and terminal branches are different (Table 2; lower-right portion).

This suggests that although trunk lineages are those that happen to survive to the following year, they might not be the ones experiencing the strongest selection pressure. Our results seem to indicate that there is no strong association between which lineages survive from year to year and the d_N/d_S ratio calculated for each lineage.

Simulation Results

Four different combinations of d_N/d_S values are simulated ($d_N/d_S = 0.5, 1.0, 1.5, 3.0$). The Branch Site model from the PAML package and the mutational mapping method are used to estimate d_N/d_S for the highlighted lineages. Estimates of the d_N/d_S ratio obtained for the simulated data sets using PAML and the methods presented here are plotted against each other in Fig. 2. As shown in the figure, the mutational mapping method provides a good approximation to codon-based likelihood models. While this result holds true for a range of parameter conditions, it is not true when branch lengths become very long (see supplementary material). In such cases, the codon-based likelihood method is preferable.

Discussion

In this study, we present an efficient computational method for exploring variation in the intensity of selection across both sites and lineages. Our method has an advantage over parsimony-based methods because it does not rely on a single mapping of mutations on a phylogeny but accommodates the statistical uncertainty in the mapping. In addition, factors such as unequal rates of transitions and transversions and unequal base frequencies are explicitly taken into account. Simulation studies show that it provides a good approximation to codon-based likelihood models when branch lengths are small or moderate. Combined with the ATV package (Zmasek and Eddy 2001), our approach allows users to visualize the change of selection intensity over large phylogenies with several hundred taxa quite flexibly (supplementary material).

Some of our conclusions differ from those of previous studies (Bush et al. 1999b; Fitch et al. 1997). First, some of the codons we identify as being under positive selection are different. Only 9 of 12 sites we identified as being selected are among the 18 sites identified by Bush et al. (1999b). One reason for this discrepancy is the difference between the two methods for mapping mutations on the tree. Another reason is that the average ratio of nonsynonymous-to-synonymous substitution rates over the entire gene were used to calculate critical values in the study by Bush et al. (1999b). As pointed out by Suzuki and Gojobori (1999), using the average value across all positions fails to take account of variation in codon composition across sites and can potentially lead to biases.

Bush et al. (1999b) also concluded that 40% of the nonsynonymous changes on egg-grown terminal branches are host-mediated substitutions. They based this conclusion on the assumption that the expected number of substitutions occurring on either egg-grown or cell-grown grown terminal branches should be proportional to the number of such lineages (Bush et al. 1999b). However, this assumption is valid only if the number of rounds of passages in the

laboratory medium and years they are isolated are similar for both egg-grown and cell-grown lines. However, we found that strains isolated early are mostly propagated in chicken eggs and that early parts of the tree are sparsely sampled. The longer terminal branches leading to egg-grown strains might simply be due to the fact that more early strains were sampled, and for them terminal branches would be longer.

Previous studies have attempted to identify sets of sites that best predict the evolution of the trunk lineages (Bush et al. 1999a; Lee and Chen 2004). The rationale is that isolates that experience a higher number of nonsynonymous substitutions in this “index set” along the trunk lineages are considered more likely to be the group of isolates from which future epidemic strains will emerge. However, the fact that we find no significant differences between d_N/d_S ratios on trunk, twig, and terminal branches makes interpretation of the influenza DNA data much harder than previously thought. Although there are examples where nonsynonymous changes seem to be associated with a new arising epidemic (e.g., in dengue virus [Bennett et al. 2003]), the absence of differences in the d_N/d_S ratios among various parts of the tree of influenza seems to contradict the general conception that future epidemics can be predicted using comparisons of d_N/d_S ratios among lineages. Our results can be reconciled with previous results by noticing that the index set of sites incorporates a very high proportion of the total number of substitutions in the gene. Isolates with a higher number of nonsynonymous changes within the index set tend also to show patterns of a high level of substitution in general. In other words, previous predictions may simply reflect the fact that, as the trunk lineages are the lineages with the longest branch lengths, they will also have more nonsynonymous mutations in the index set of sites.

It is possible that the role of trunk lineages in the tree and the causes of selection acting on different lineages are more complex than previously assumed. The relationship between virulence and selection may not be simple. For example, the success of particular strains may depend on the functional complementation of hemagglutinin to several other genes, which may experience different evolutionary histories (Holmes et al. 2005). In addition, the amount of selection acting on the viral sequence may effectively depend on a number of epidemiological factors such as the duration of infection, the number of viral particles involved in an infection, and the efficacy with which new mutations lead to immune avoidance (e.g., the antigenic distances). Moreover, some empirical evidence shows that the actual epidemic strains do not necessarily lie on the trunk of the evolutionary tree (Robertson 1987). Our results suggest that selection acts much more homogeneously on the tree than previously thought. It is clear that

more work is needed to understand the complex relationship between epidemiology and DNA sequence variation.

Host-mediated substitutions have been documented in many viruses (Cao et al. 1995; Graff et al. 1994; Itoh et al. 1997; Sawyer et al. 1994). Our analysis confirms that inferences regarding adaptation in pathogens without distinguishing lineages propagated in egg cells from other lineages can be potentially misleading. Although many of the sequence data currently obtained have not been subject to propagation in egg cells, most of the early data deposited in databases have. As pointed out in previous studies (Bush et al. 2000), analyses based on these sequences must take the laboratory propagation history into account.

Similarly to the codon-based likelihood approaches (Nielsen and Yang 1998; Yang et al. 2000), our approximation method has some restrictions such as a lack of power when lineages are too short (Wong et al. 2004; Yang and Nielsen 2000). If we ignore mutations on terminal branches, only site 226 is identified as being positively selected, although several other positions have only nonsynonymous changes and no synonymous substitution changes (Table 1). The binomial-counting method is known to be too conservative and has less power than a likelihood-based method (Wong et al. 2004). This problem might also exist in estimating d_N/d_S ratios when there are not enough substitutions. However, our method provides a computationally efficient way to explore variation in selection intensity across lineages and sites in very large data sets.

Software Availability

A computer program implementing this method and associated influenza sequences will be distributed at the Slatkin group web site (<http://www.ib.berkeley.edu/labs/slatkin/software.html>).

Acknowledgments We thank Wendy Wong and three Slatkin group members—Philip Johnson, Anna-Sapfo Malaspinas, and Owen Solberg—for helpful discussions. We also thank two reviewers for helpful comments. W.Z. and M.S. are supported in part by National Institutes of Health Grant NIH-GM-40282 and R.N. is supported by Danmarks Grundforskningsfond and the Danish FSS and FNU.

References

- Alba R, Kelmenson PM, Cordonnier-Pratt MM, Pratt LH (2000) The phytochrome gene family in tomato and the rapid differential evolution of this family in angiosperms. *Mol Biol Evol* 17:362–373
- Bennett SN, Holmes EC, Chirivella M, Rodriguez DM, Beltran M, Vorndam V, Gubler DJ, McMillan WO (2003) Selection-driven evolution of emergent dengue virus. *Mol Biol Evol* 20:1650–1658

- Bollback JP (2005) Posterior mapping and predictive distributions. In: Nielsen R (ed) *Statistical methods in molecular evolution*. Springer Verlag, New York
- Bush RM (2004) Influenza as a model system for studying the cross-species transfer and evolution of the SARS coronavirus. *Philos Trans R Soc Lond B Biol Sci* 359:1067–1073
- Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999a) Predicting the evolution of human influenza A. *Science* 286:1921–1925
- Bush RM, Fitch WM, Bender CA, Cox NJ (1999b) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol* 16:1457–1465
- Bush RM, Smith CB, Cox NJ, Fitch WM (2000) Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc Natl Acad Sci USA* 97:6974–6980
- Cao JX, Ni H, Wills MR, Campbell GA, Sil BK, Ryman KD, Kitchen I, Barrett AD (1995) Passage of Japanese encephalitis virus in HeLa cells results in attenuation of virulence in mice. *J Gen Virol* 76(Pt 11):2757–2764
- Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Washington, DC, pp 345–352
- 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* 94:7712–7718
- Fitch WM, Leiter JM, Li XQ, Palese P (1991) Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci USA* 88:4270–4274
- Graff J, Normann A, Feinstone SM, Flehmig B (1994) Nucleotide sequence of wild-type hepatitis A virus GBM in comparison with two cell culture-adapted variants. *J Virol* 68:548–554
- Gubareva LV, Wood JM, Meyer WJ, Katz JM, Robertson JS, Major D, Webster RG (1994) Codominant mixtures of viruses in reference strains of influenza virus due to host cell variation. *Virology* 199:89–97
- Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci USA* 101:12957–12962
- Hardy CT, Young SA, Webster RG, Naeve CW, Owens RJ (1995) Egg fluids and cells of the chorioallantoic membrane of embryonated chicken eggs can select different variants of influenza A (H3N2) viruses. *Virology* 211:302–306
- Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol* 3:e300
- Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755
- Huttley GA, Eastaugh S, Southey MC, Tesoriero A, Giles GG, McCredie MR, Hopper JL, Venter DJ (2000) Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat Genet* 25:410–413
- Ito T (2000) Interspecies transmission and receptor recognition of influenza A viruses. *Microbiol Immunol* 44:423–430
- Ito T, Suzuki Y, Takada A, Kawamoto A, Otsuki K, Masuda H, Yamada M, Suzuki T, Kida H, Kawaoka Y (1997) Differences in sialic acid-galactose linkages in the chicken egg amnion and allantois influence human influenza virus receptor specificity and variant selection. *J Virol* 71:3357–3362
- Itoh M, Isegawa Y, Hotta H, Homma M (1997) Isolation of an avirulent mutant of Sendai virus with two amino acid mutations from a highly virulent field strain through adaptation to LLC-MK2 cells. *J Gen Virol* 78(Pt 12):3207–3215
- Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. *J Mol Evol* 20:86–93
- Lee MS, Chen JS (2004) Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg Infect Dis* 10:1385–1390
- Macken C, Lu H, Goodman J, Boykin L (2001) The value of a database in surveillance and vaccine selection. In: Osterhaus ADME, Cox N, Hampson AW (eds) *Options for the control of influenza IV*. Elsevier Science, Amsterdam, pp 103–106
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654
- Meguro H, Bryant JD, Torrence AE, Wright PF (1979) Canine kidney cell line for isolation of respiratory viruses. *J Clin Microbiol* 9:175–179
- Nakajima S, Nakajima K, Kendal AP (1983) Identification of the binding sites to monoclonal antibodies on A/USSR/90/77 (H1N1) hemagglutinin and their involvement in antigenic drift in H1N1 influenza viruses. *Virology* 131:116–127
- Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. *Nat Rev Genet* 8:196–205
- Nielsen R (2001) Mutations as missing data: inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics* 159:401–411
- Nielsen R (2002) Mapping mutations on phylogenies. *Syst Biol* 51:729–739
- Nielsen R, Huelsenbeck JP (2002) Detecting positively selected amino acid sites using posterior predictive P-values. *Pacif Symp Biocomput* 576–588
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- Robertson JS (1987) Sequence analysis of the haemagglutinin of A/Taiwan/1/86, a new variant of human influenza A(H1N1) virus. *J Gen Virol* 68(Pt 4):1205–1208
- Robertson JS (1993) Clinical influenza virus and the embryonated hen's egg. *Rev Med Virol* 3:97–106
- Robertson JS (1999) An overview of host cell selection. *Dev Biol Stand* 98:7–11, discussion 73–74
- Rocha EP, Xu X, Hall HE, Allen JR, Regnery HL, Cox NJ (1993) Comparison of 10 influenza A (H1N1 and H3N2) haemagglutinin sequences obtained directly from clinical specimens to those of MDCK cell- and egg-grown viruses. *J Gen Virol* 74(Pt 11):2513–2518
- Rogers GN, Paulson JC, Daniels RS, Skehel JJ, Wilson IA, Wiley DC (1983) Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. *Nature* 304:76–78
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Sawyer LS, Wrinn MT, Crawford-Miksza L, Potts B, Wu Y, Weber PA, Alfonso RD, Hanson CV (1994) Neutralization sensitivity of human immunodeficiency virus type 1 is determined in part by the cell in which the virus is propagated. *J Virol* 68:1342–1349
- Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu Rev Biochem* 69:531–569
- Stewart CB, Schilling JW, Wilson AC (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* 330:401–404
- Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16:1315–1328
- Swofford D, Maddison DR (1987) Reconstructing ancestral character states under Wagner parsimony. *Math Biosci* 87:199–229
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680

- Weis W, Brown JH, Cusack S, Paulson JC, Skehel JJ, Wiley DC (1988) Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* 333:426–431
- Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699
- Wong WS, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556
- Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496–503
- Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32–43
- Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917
- Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118
- Zmasek CM, Eddy SR (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17:383–384