Modeling site-specific amino-acid preferences deepens phylogenetic estimates of the divergence of viral proteins.

Sarah K. Hilton[1,2] and Jesse D. Bloom[1,2*]

[1]Basic Sciences and Computational Biology Program, Fred Hutchinson Cancer Research Center

[2]Department of Genome Sciences, University of Washington

Seattle, WA, USA

[*]E-mail: jbloom@fredhutch.org.

Abstract

Molecular phylogenetics is often used to estimate the time since the divergence of modern gene sequences. For highly diverged sequences, such phylogenetic techniques can estimate surprisingly recent divergence times. In the case of viruses, there is independent evidence that the estimates of deep divergence times from molecular phylogenetics are too recent. This discrepancy is caused in part by inadequate models of purifying selection leading to branch-length underestimation. Here we examine the effect on branch-length estimation of using models that incorporate experimental measurements of purifying selection. We find that models informed by experimentally measured site-specific amino-acid preferences estimate longer deep branches on phylogenies of influenza virus hemagglutinin. This lengthening of branches is due to more realistic stationary states of the models, and is independent of the branch-length-extension from modeling site-to-site variation in amino-acid substitution rate. The branch-length extension from experimentally informed site-specific models is similar to that achieved by other approaches that allow the stationary state to vary across sites. However, the improvements from these site-specific but time-homogeneous and site-independent models are limited by the fact that a protein's amino-acid preferences gradually shift as it evolves. Overall, our work underscores the importance of modeling site-specific amino-acid preferences when estimating deep divergence times—but also shows the inherent limitations of approaches that fail to account for how these preferences shift over time.

Introduction

[I have commented on the general structure of the Introduction, which I think is mostly OK. Add in references next, as this is crucial. The Intro should frame the general problem and describe specifically what is already known. Right now it does a bit too much of repeating your first two sections of Results. In particular, there is some balance between how much goes in Intro and Results. Right now we have a lot of good background in Results, so don't get too repetitive in Intro.]

When studying a protein's evolutionary history, it is important to understand the timescale at which the evolutionary events have occurred. For example, understanding how long a particular viral lineage has been circulating informs hypotheses about population size, evolutionary rate, and possible previous hosts. Molecular phylogenetic is commonly used to make sure estimates of time since the divergence of two lineages. Under a molecular clock assumption, the branch lengths on a phylogenetic tree can be transformed into estimates of absolute time. However, it has been observed that these phylogenetic estimates often appear to estimate times since deep divergences that appear to be too recent. That is, when two nodes are separated by a very long branch, the phylogenetic estimate of this branch length often appears to be too short, which leads to an estimate of time that is too recent. [I would make the first sentence more general than just proteins. Second sentence: good idea to use viruses as examples, but try to make the sentence a bit more specific and less general—also, population size may not be relevant to estimates of deep divergence times. For the last part, can we in fact cite general non-viral references about times being too recent? If you can't find several good ones, then just drop that and make first paragraph solely about divergence estimates and have next paragraph go into underestimation.]

In the case of viruses, there is independent evidence that the estimates of these deep divergences are indeed too recent. The times since the divergence of lineages within lentiviruses, paramyxoviruses and filoviruses have been estimated by methods which rely on information outside of or in addition to the viral phylogenetic tree. These independent estimates are often orders of magnitude older than the estimates using the viral phylogenetic tree only. For example, there are multiple examples of filovirus integrations into the genomes of rodents. Using these viral integrations and the fossil dates of the rodent species tree, the divergence of two main filovirus groups has been estimated to be > 7 millions years ago. This estimate stands in stark contrast to the phylogenetic estimate of filovirus divergence which estimates this event happened $\sim 10,000$ years ago. All together, this indicates that phylogenetic branch length estimation does truly have a systematic bias towards underestimation. [Good content, although obviously needs references. Also, I think it is best to have at most one or two general sentences and then get right into specific examples. So I think paragraph would be stronger if you just have one (or two) general sentences

and then bring up at least two specific examples, and finally return to general with "similar things have been seen in viruses X, Y, and Z (citing each).]

Branch length underestimation is due in part to inadequate modeling of purifying selection. During evolution, protein sites do not sample all 20 available amino-acids equally. Instead, proteins have site-specific amino-acid preferences which are necessary to maintain the structure and function of the protein. Failure to account for these constraints will lead to branch length underestimation because the model will assume that sequences which have been evolving for a very long period of time should be more diverged than is actually the case. Furthermore, the site-specific amino-acid preferences at one site are often dependent on the amino-acid identity at another site. As a result, the preferences "shift" over time time due to the accumulation of substitutions in other parts of the protein. [Start with: "The under-estimation of deep branches is known to occur in part due to... Also, we need to make this paragraph about what has *already* been shown. It is not clear how much of what you argue here is things you can clearly cite to previous papers versus your new results. So overall content is good, but make this paragraph talk about prior literature with specific citations to specific points.]

Most phylogenetic substitution models are time-homogenous and site-independent: they do not take into account how sites may interact with each other or how preferences may shift over time. However, they can model purifying selection with varying degrees of complexity. One method is to allow the parameter controlling the rate of nonsynonymous to synonymous substitution, the dN/dS parameter, to vary among the protein sites. This variation accounts for the fact that constrained sites will have a lower rate of amino-acid substitution than mutationally constrained sites. However, the stationary state of this model, which describes the expected sequence composition after a very long evolutionary time, will be exactly the same as a model without any variation in the dN/dS parameter. Alternatively, models can have site-specific stationary states which explicitly acknowledge the protein's site-specific amino-acid preferences. Such models should estimate deep divergence times because model still expects relatively low sequence divergence even after a very long evolutionary time. [Generally good (except for lack of citations), but again is last point the one you are making or something that you can clearly cite from literature? See comment after next paragraph.]

Here, we tested the effect of modeling site-specific amino-acid preferences though a model with a site-specific stationary state on an influenza hemagglutinin (HA) phylogenetic tree. Specifically, we used Experimentally Informed Codon Models (ExpCM's) defined by site-specific amino-acid preferences measured by a high-throughput functional assay called deep mutational scanning. We found that these ExpCM's estimated longer branches. We found that the extension in branch length due modeling purifying selection via a site-specific stationary state is largely independent of the the effect of modeling purifying selection through rate variation. However, our results

make it clear that the site-specific amino-preferences of HA are constant across the entire tree. In particular, the branch lengthening is most pronounced on branches leading to the HA sequence that was used in the deep mutational scanning experiment that parameterized the ExpCM. These results underscore the importance of modeling how site-specific purifying selection affects the stationary state when estimating deep divergence times—but also shows the inherent limitations of approaches that fail to model how this selection shifts over time. [Looking at this final paragraph, I would suggest structuring it and the two above it (last three) like this:

- Under-estimation can result from failure to model purifying selection.
- Most models are homogenous / independent, and model by rate variation.
- Some models use site-specific stationary states.
- Here we use ExpCMs and find...

So overall, pretty good structure that you already have but don't put too much of your results into the Introduction.

The one thing that you need to add is a better description of deep mutational scanning. So I would say spend less time repeating your results and more time providing that background in the last paragraph (or an additional added paragraph).]

Results

Different ways substitution models account for purifying selection

Proteins evolve under purifying selection to maintain their structure and function. This purifying selection is not homogenous across sites in a protein. It is also not homogenous among the different amino acids at a given site. For instance, some protein sites strongly prefer hydrophobic amino acids, others are constrained to just one or a few amino acids, and yet others tolerate many amino acids. In general, these constraints are highly idiosyncratic among sites, and so pose a challenge for phylogenetic substitution models.

Here we consider how purifying selection is handled by codon models, which are the most accurate of the three classes (nucleotide, codon, and amino acid) of phylogenetic substitution models in widespread use (Arenas, 2015). Standard codon models distinguish between two types of substitutions: synonymous and nonsynonymous. The relative rate of these substitutions is referred to as dN/dS or ω. In their simplest form, codon substitution models fit a single ω that represents the gene-wide average fixation rate of nonsynonymous mutations relative to synonymous ones. Here we will use such substitution models in the form proposed by Goldman and Yang (1994).

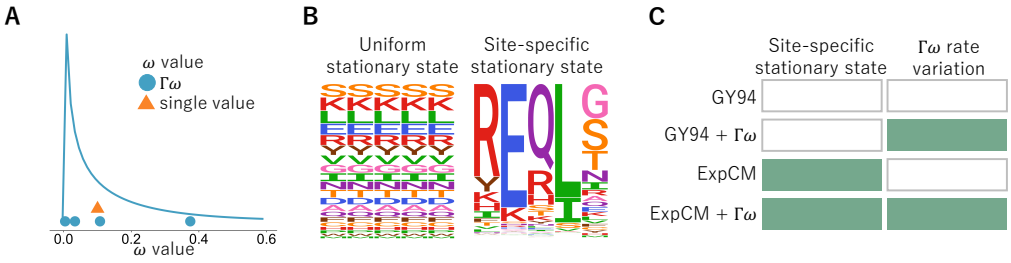


Figure 1: Different ways codon models account for purifying selection. (A) The dN/dS parameter, ω, can be defined as one gene-wide average (orange triangle) or allowed to vary according to some statistical distribution (blue circles). For computational tractability, the distribution is discretized into K bins and ω takes on the mean of each bin (Yang, 1994; Yang et al., 2000). A gamma distribution (denoted by Γ) with $K = 4$ bins is shown here. (B) A substitution model's stationary state defines the expected sequence composition after a very long evolutionary time. Most substitution models have stationary states that are uniform across sites. However, substitution models can have site-specific stationary states. In the logo plots, each column is a site in the protein and the height of each letter is the frequency of that amino acid at stationary state. (C) Substitution models can incorporate neither, one, or both of these features. Here we will use substitution models from the Goldman-Yang (GY94; Goldman and Yang, 1994; Yang et al., 2000) and experimentally informed codon model (ExpCM; Hilton et al., 2017) families with and without gamma-distributed ω to represent all possible combinations.

When these models have a single gene-wide ω they are classified as M0 by Yang et al. (2000). Here we will refer to M0 Goldman-Yang models simply as GY94 models (Equation 1). The gene-wide ω is usually < 1 (Murrell et al., 2015), and crudely represents the fact that many amino-acid substitutions are under purifying selection.

A single gene-wide ω ignores the fact that purifying selection is heterogeneous across sites. The most common strategy to ameliorate this defect is to allow ω to vary among sites according to some statistical distribution (Yang, 1994; Yang et al., 2000). For instance, in the M5 variant of the GY94 model (Yang et al., 2000), ω follows a gamma distribution as shown in Figure 1A. We will denote this model as GY94+$\Gamma\omega$. A GY94+$\Gamma\omega$ captures the fact that the rate of nonsynonymous substitution can vary across sites. However, this formulation treats all nonsynonymous substitutions equivalently, since the rate is agnostic to the amino-acid identity of the mutation.

A model formulation that accounts for the fact that purifying selection depends on the specific amino-acid mutation is provided by so-called "mutation-selection" models (Halpern and Bruno, 1998; Yang and Nielsen, 2008; Rodrigue et al., 2010; Tamuri et al., 2012; McCandlish and Stoltzfus, 2014). Here we will consider mutation-selection models where the site-specific selection is assumed to act solely at the protein level (different codons for the same protein are treated as selectively equivalent). Such models explicitly define a different set of amino-acid preferences at

each site in the protein. This more mechanistic formulation results in a site-specific stationary state (Figure 1B). These models capture the site-to-site variation in amino-acid composition that is an obvious features of real proteins, and usually better describe actual evolution than models with only rate variation as assessed by Bayesian or maximum-likelihood criteria (Lartillot and Philippe, 2004; Le et al., 2008; Quang et al., 2008; Wang et al., 2008; Rodrigue et al., 2010; Bloom, 2014a,b; Hilton et al., 2017).

However, the increased realism of mutation-selection models comes at the cost of an increased number of parameters. Codon substitution models with uniform stationary states have only a modest number of parameters that must be fit from the phylogenetic data. For instance, a GY94+$\Gamma\omega$ model with the commonly used F3X4 stationary state has 12 parameters: two describing the shape of the gamma distribution over ω, a transition-transversion rate, and nine parameters describing the nucleotide composition of the stationary state. However, mutation-selection models must additionally specify 19 parameters defining the amino-acid preferences for *each* site (there are 20 amino acids whose preferences are constrained to sum to one). This corresponds to $19 \times L$ parameters for a protein of length L, or 9,500 parameters for a 500-residue protein. It is challenging to obtain values for these amino-acid preference parameters in a maximum-likelihood framework without overfitting the data (Rodrigue, 2013). Here we will primarily use experimentally informed codon models (ExpCM's), which define the site-specific amino-acid preference parameters *a priori* from deep mutational scanning experiments so that they do not need to be fit from phylogenetic data (see Methods and Bloom, 2014a; Hilton et al., 2017; Bloom, 2017). Because the amino-acid preference parameters in an ExpCM are obtained from experiments, the number of ExpCM free parameters is similar to a non-site-specific substitution model. An alternative strategy of obtaining the amino-acid preference parameters via Bayesian inference (Lartillot and Philippe, 2004; Rodrigue and Lartillot, 2014) is discussed in the last section of the Results.

Importantly, these two strategies for modeling purifying selection are not mutually exclusive. Mutation-selection models such as an ExpCM can still incorporate an ω parameter, which now represents the relative rate of nonsynonymous to synonymous substitution *after* accounting for the constraints due to the site-specific amino-acid preferences (Bloom, 2017; Rodrigue and Lartillot, 2017). This ω parameter for an ExpCM can be drawn from a statistical distribution (e.g., a gamma distribution) just like for GY94-style models (Rodrigue and Lartillot, 2014; Haddox et al., 2017). We will denote such models as ExpCM+$\Gamma\omega$. Figure 1C shows the full spectrum of models that incorporate all combinations of gamma-distributed ω and site-specific stationary states.

Effect of stationary state and rate variation on branch-length estimation

Given a single branch, a substitution model transforms sequence identity into branch length. Under a molecular-clock assumption, this branch length is proportional to time. The transformation from

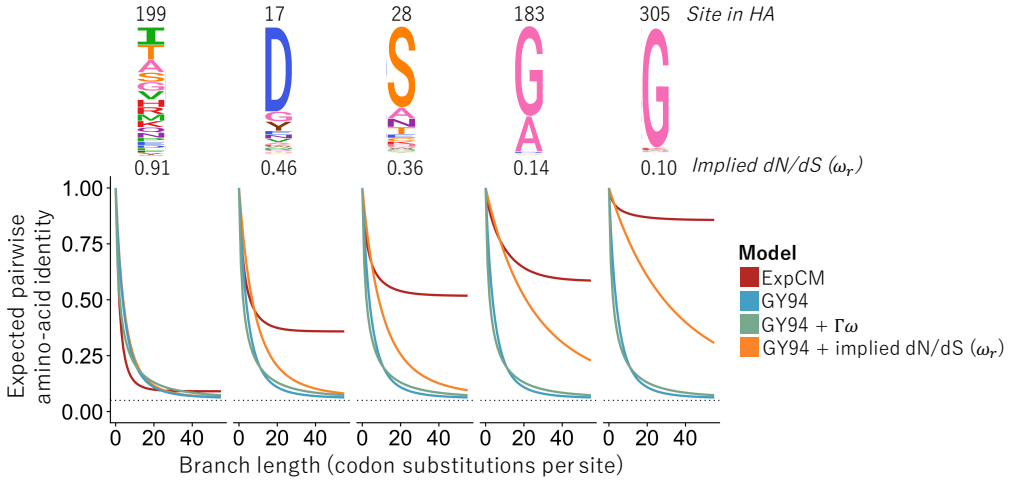


Figure 2: Effect of stationary state and $\Gamma\omega$ rate variation on predicted asymptotic sequence divergence. The logo plots at top show the amino-acid preferences for some sites in an H1 influenza hemagglutinin protein as experimentally measured by Doud and Bloom (2016). The graphs show the expected amino-acid identity at that site for two sequences separated by a branch of the indicated length (Equation 9). For the GY94 model, the graphs are identical for all sites since this model does not have site-specific parameters; the same is true for GY94+$\Gamma\omega$. The graphs do differ among sites if we calculate a different ω_r for each site r in the GY94 model using the amino-acid preferences (Equation 7; Spielman and Wilke, 2015b). However, all GY94 models, including the one with site-specific ω_r values, approach the same asymptote since they all have the same stationary state. The ExpCM has different asymptotes for different sites since it accounts for how amino-acid preferences lead to site-specific stationary states.

sequence identity to branch length is trivial when the sequence identity is high. For instance, when there has only been one substitution, then the sequence identity will simply be $\frac{L-1}{L}$ for a gene of L sites, and even a simple exponential model (Zuckerkandl and Pauling, 1965) will correctly infer the short branch length of $1/L$ substitutions per site. However, as substitutions accumulate it becomes progressively more likely for multiple changes to occur at the same site. In this regime, the accuracy of the substitution model becomes critical for transforming sequence identity into branch length. Any time-homogenous substitution model predicts that after a very large number of substitutions, two related sequences will approach some asymptotic amino-acid sequence identity. For instance, if all 20 amino acids are equally likely in the stationary state, then this asymptotic sequence identity will be $\frac{1}{20} = 0.05$. If the substitution model underestimates the asymptotic sequence identity then it will also underestimate long branch lengths, since it will predict that sequences that have evolved for a very long time should be more diverged than is actually the case.

Figure 2 shows how different substitution models predict amino-acid sequence identity to

decrease as a function of branch length using model parameters fit to a phylogeny of H1 influenza hemagglutinin (HA) genes. The GY94 model predicts the same behavior for all sites, since it does not have any site-specific parameters, with an asymptotic sequence identity of 0.059. While this predicted sequence identity is higher than $\frac{1}{20} = 0.05$, due to redundant codon and nucleotide biases favoring certain amino acids, it is much lower than the pairwise identity of even the most diverged HAs in nature. While it is of course possible that the identity of HAs in nature would become even lower given more time, it seems biochemically improbable that it would ever become as low as 0.059. The reason is that like many proteins HA has a highly conserved structure and function that imposes constraints that cause many sites to sample only a small subset of the 20 amino acids among all known HA homologs (Nobusawa et al., 1991).

Accounting for site-to-site dN/dS rate variation in GY94 models affects the rate at which the asymptotic sequence identity is approached, but not the actual value of this asymptote. For instance, Figure 2 shows that the GY94+$\Gamma\omega$ model takes longer to reach the asymptote than GY94, but that the asymptote is identical for both models. This fact holds true even if we use experimental measurements of HA's site-specific amino-acid preferences (Doud and Bloom, 2016) to calculate a different ω_r value for each site using the method of Spielman and Wilke (2015b) (see Equation 7). Specifically, this GY94+ω_r model predicts that different sites will approach the asymptote at different rates, but the asymptote is always the same (Figure 2). The invariance of the asymptotic sequence identity under different schemes for modeling ω is a fundamental feature of the mathematics of reversible substitution models. These models are reversible stochastic matrices, which can be decomposed into stationary states and symmetric exchangeability matrices (Nielsen, 2006). The stationary state is invariant with respect to multiplication of the symmetric exchangeability matrix by any non-zero number. Different schemes for modeling ω only multiply elements of the symmetric exchangeability matrix. Therefore, no matter how "well" a model accounts for site-to-site variation in ω, it will always have the same stationary state as a simple GY94 model.

However, mutation-selection models such as ExpCM's have site-specific stationary states. They predict that different sites will have different asymptotic sequence identities (Figure 2)—a prediction that accords with the empirical observation that some sites are much more variable than others in alignments of highly diverged sequences. For instance, Figure 2 shows that at sites such as 183 and 305 in the H1 HA, an ExpCM but not a GY94-style model predicts that the identity will always be relatively high. When sites with highly constrained amino-acid preferences such as these are common, an ExpCM can estimate a long branch length at modest sequence identities that a GY94 model might attribute to a shorter branch.

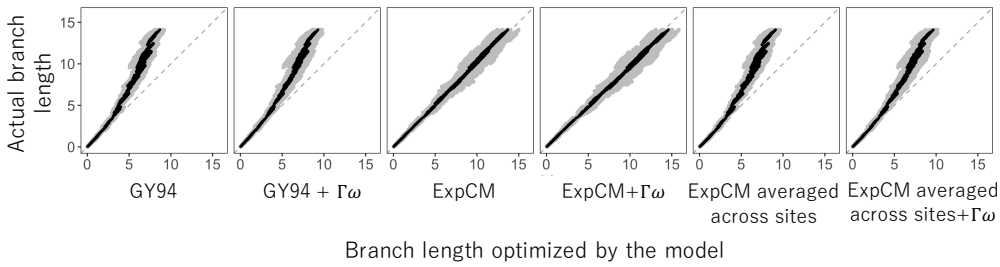


Figure 3: Branch lengths inferred on data simulated under a model with site-specific amino-acid preferences. We simulated alignments along a HA phylogenetic tree (see Figure 4) using an ExpCM parameterized by the actual site-specific amino-acid preferences for an H1 HA (Doud and Bloom, 2016). We then inferred the branch lengths of this tree from the simulated alignments. The inferred branch lengths for various models are plotted on the x-axis, and the actual branch lengths used in the simulations are on the y-axis. We performed 10 simulations and inferences, and gray points show each inferred branch length from each simulation, and black points show the average of each branch length across simulations. The grey dashed line at $y = x$ represents the behavior of an unbiased estimator.

Simulations demonstrate how failure to model site-specific amino-acid preferences leads to branch-length underestimation.

To directly demonstrate the effect of stationary state and $\Gamma\omega$ rate variation on branch-length estimation, we tested the ability of a variety of models to accurately infer branch lengths on simulated data (Figure 3). Specifically, we simulated alignments of sequences along the HA phylogenetic tree in Figure 4 using an ExpCM parameterized by the amino-acid preferences of H1 HA as experimentally measured by deep mutational scanning (Doud and Bloom, 2016). We then estimated the branch lengths from the simulated sequences using all the substitution models in Figure 1C, and compared these estimates to the actual branch lengths used in the simulations.

The models with a uniform stationary state underestimated the lengths of long branches on the phylogenetic tree of the simulated sequences (Figure 3). The GY94 model estimated branch lengths that are $\sim$60% of the true values for the longest branches. Accounting for site-to-site variation in ω did not fix the fundamental problem: the GY94+$\Gamma\omega$ did slightly better, but still substantially underestimated the longest branches. However, there was no systematic underestimation of long branches by the ExpCM and ExpCM+$\Gamma\omega$ models. The improved performance of the ExpCM's is due to their modeling of the site-specific amino-acid preferences: if we parameterize ExpCM's by amino-acid preferences that have been averaged across HA sites (and so are no longer site-specific), then they perform no better than GY94 models (Figure 3). Therefore, models with uniform stationary states underestimate the length of long branches in phylogenies of

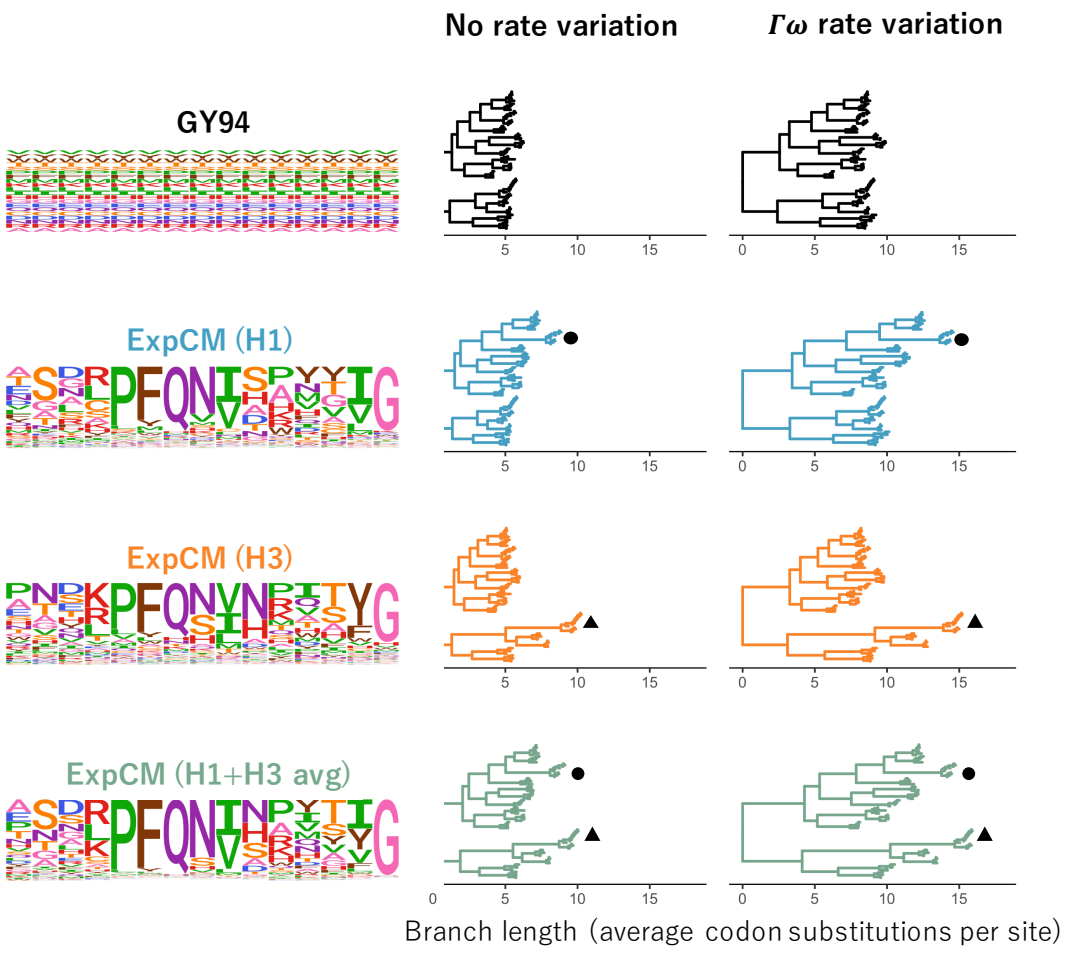


Figure 4: Effect of site-specific amino-acid preferences and $\Gamma\omega$ rate variation on HA branch length estimation. The branch lengths of the HA tree were optimized using the indicated ExpCM or GY94 model. The amino-acid preferences defining the model (ExpCM) or implied by the model (GY94) are shown as logo plots for 15 sites in HA; the full set of experimentally measured amino-acid preferences are in Supplementary figure 1, Supplementary figure 2, and Supplementary figure 3. The ExpCM's use amino-acid preferences measured in deep mutational scanning of an H1 HA (Doud and Bloom, 2016), an H3 HA (Lee et al., 2018), or the average of the measurements for these two HAs. Circle denotes the H1 clade and triangle denotes the H3 clade. The trees are midpoint rooted based on the tree inferred by `RAxML` using the GTRCAT model.

sequences that have evolved under strong site-specific amino-acid preferences.

Experimentally informed site-specific models estimate longer branches on real data.

The foregoing section shows the superiority of ExpCM's to GY94 models for estimating long branches on phylogenies simulated with ExpCM's. But how do these models perform on real data? Real genes do evolve under functional constraint, but these constraints are almost certainly more complex than what is modeled by an ExpCM. However, if ExpCM's do a substantially better job than GY94 models of capturing the true constraints, then we might still expect them to estimate more accurate branch lengths.

To test the models on real data, we used actual sequences of influenza HA. The topology of HA phylogenetic trees (Figure 4) makes these sequences an interesting test case for branch-length estimation. HA consists of a number of different subtypes. Sequences within a subtype have >68% amino-acid identity, but sequences in different subtypes have as little as 38% identity. However, HA proteins from all subtypes have a highly conserved structure that performs a highly conserved function (Ha et al., 2002; Russell et al., 2004). We used `RAxML` (Stamatakis, 2006) with a nucleotide substitution model (GTRCAT) to infer a phylogenetic tree for 92 HA sequences drawn from 15 of the 18 subtypes (we excluded bat influenza and one other rare subtype). For the rest of this paper, we fix the tree topology to this `RAxML`-inferred tree. Although the nucleotide model used with `RAxML` to infer this tree topology is probably less accurate than codon models, the modular subtype structure of the HA phylogeny means that most of the phylogenetic uncertainty lies in the length of the long branches separating the subtypes rather than in the tree topology itself.

Deep mutational scanning has been used to measure the amino-acid preferences of all sites in two different HAs. One scan measured the preferences of an H1 HA (Doud and Bloom, 2016) and the other measured the preferences of an H3 HA (Lee et al., 2018). The amino-acid preferences measured for these two HAs are shown in Supplementary figure 1 and Supplementary figure 2. The H1 and H3 HAs have only ∼42% amino-acid identity, and so are separated by a large distance on the phylogenetic tree (see triangle and circle on Figure 4). As described in Lee et al. (2018), the amino-acid preferences clearly differ between the H1 and H3 HA at a substantial number of sites (these differences are apparent in a simple visual comparison of Supplementary figure 1 and Supplementary figure 2; see site 33 as an example). Therefore, we also created a third set of amino-acid preferences by averaging the measurements for the H1 and H3 HAs, under the conjecture that these averaged preferences might better describe the "average" constraint on sites across the full HA tree [reference suppfig with the average logo plot here]. These three sets of HA amino-acid preferences define three different ExpCM's.

We fit the GY94 model and each of the three ExpCM's to the fixed HA tree topology estimated using `RAxML`, and also tested a version of each model with $\Gamma\omega$ rate variation. Table 1 shows that all ExpCM's fit the actual data much better than the GY94 models. The best fit was for the ExpCM informed by the average of the H1 and H3 deep mutational scans. For all models, incorporating

Table 1: Fitting of substitution models to the HA phylogenetic tree. The models fit here are the same ones in Figure 4. All ExpCMs describe the evolution of HA better than the GY94 models, as evaluated by the Akaike information criteria (ΔAIC, Posada and Buckley, 2004) The ω value for each of the $K = 4$ bins is shown for the models with $\Gamma\omega$ rate variation. All ExpCM's fit a stringency parameter > 1.

Model	ΔAIC	Log Likelihood	ω	Stringency parameter (β)
ExpCM (H1+H3 avg) + $\Gamma\omega$	0	-51083	0.19, 0.50, 0.91, 1.86	1.69
ExpCM (H1+H3 avg)	1063	-51616	0.14	1.77
ExpCM (H1) + $\Gamma\omega$	1321	-51744	0.12 , 0.42, 0.89, 2.13	1.11
ExpCM (H3) + $\Gamma\omega$	1777	-51972	0.10, 0.36, 0.76, 1.84	1.28
ExpCM (H1)	2670	-52419	0.12	1.21
ExpCM (H3)	3377	-52773	0.12	1.43
GY94 + $\Gamma\omega$	4817	-53487	0.00, 0.03, 0.08, 0.24	-
GY94	7892	-55025	0.07	-

$\Gamma\omega$ rate variation improved the fit, although even ExpCM's without $\Gamma\omega$ greatly outperformed the GY94+$\Gamma\omega$ model (Table 1). As mentioned in the previous section, ω is generally < 1 when a single value is fit to all sites in a gene (Murrell et al., 2015), and this is the case for all the models we tested (Table 1). However, the ExpCM's always fit an ω greater than the GY94 model, suggesting that the site-specific amino-acid preferences capture some of the purifying selection that the GY94 models can represent only via a small ω. Among the models with $\Gamma\omega$, the GY94+$\Gamma\omega$ model fits all four ω categories to values $\ll 1$, but the ExpCM+$\Gamma\omega$ models fit one of the ω categories to a value > 1. This increase in ω values makes sense given the different interpretation of ω for each family of models. The ExpCM ω is the relative rate of fixation of nonsynonymous to synonymous mutations *after* accounting for the functional constraints described by the amino-acid preferences. This more realistic null model gives ExpCM's enhanced power to detect diversifying selection for amino-acid change (Bloom, 2017; Rodrigue and Lartillot, 2017), which is known to occur at some sites in HA due to immune selection (Bedford et al., 2014).

Importantly, models that account for purifying selection via either $\Gamma\omega$ rate variation or a site-specific amino-acid preferences do not just exhibit better fit—they also estimate longer deep branches on the HA tree. Figure 4 shows the branch lengths optimized by each model on a common scale. The tree's deepest branches are shortest when they are optimized by the GY94 model, which lacks both $\Gamma\omega$ and site-specific amino-acid preferences. Adding either $\Gamma\omega$ rate variation or site-specific amino-acid preferences increases the length of the deep branches. Specifically, the tree's diameter (the distance from the two most divergent tips) for the GY94+$\Gamma\omega$ model is 159% of the GY94 model tree diameter (Supplementary table 1). The tree diameter is 122% and 135% of the GY94 model tree diameter for ExpCM's informed by H1 or H3 amino-acid preferences,

respectively, and 160% of the GY94 model for the ExpCM informed by the average of the H1 and H3 preferences (Supplementary table 1).

The deepening of branch lengths that results from the $\Gamma\omega$ and site-specific amino-acid preference approaches to modeling purifying selection are largely independent. This can be seen by examining the ExpCM+$\Gamma\omega$ models, which combine $\Gamma\omega$ rate variation with site-specific amino-acid preferences. As shown in Figure 4, these ExpCM+$\Gamma\omega$ models estimate longer branches than models with just $\Gamma\omega$ rate variation (GY94+$\Gamma\omega$) or just site-specific amino-acid preferences (ExpCM's). The near independence of these effects is quantified in Supplementary table 1, which shows that 76% of the tree diameter extension of ExpCM(H1+H3 avg)+$\Gamma\omega$ versus can be explained by simply adding the extension from incorporating $\Gamma\omega$ (GY94+$\Gamma\omega$ versus GY94) to the extension from incorporating site-specific amino-acid preferences (ExpCM(H1+H3 avg) versus GY94).

However, while adding $\Gamma\omega$ rate variation increases the length of deep branches in a roughly uniform fashion across the tree, the branch lengthening from adding site-specific amino-acid preferences is not uniform across the tree (Figure 4). Instead, the increase in branch length is most pronounced on branches leading to the HA sequence that was used in the deep mutational scanning experiment that informed the ExpCM. For instance, the ExpCM informed by the H1 data most dramatically lengthens branches near the H1 clade of the tree, while the ExpCM informed by the H3 data has the largest effect on branches near the H3 clade. The ExpCM informed by the average of the H1 and H3 data has a more uniform effect across the tree, but still most strongly extends branches leading to either the H1 or H3 clade. Therefore, Figure 4 shows that ExpCM's estimate longer branches, but that the effect is shaped by the set of amino-acid preferences used to inform the model.

Shifting amino-acid preferences limit the benefits of models with site-specific stationary states for estimating long branch lengths.

The fact that an ExpCM leads to the most profound increase in branch length leading to the sequence used in the experiment can be rationalized in terms of existing knowledge about epistasis during protein evolution. Each ExpCM is informed by a single set of experimentally measured amino-acid preferences. But in reality, the effect of a mutation at one site in a protein can depend on the amino-acid identities of other sites in the protein (Ortlund et al., 2007; Gong et al., 2013; Harms and Thornton, 2014; Tufts et al., 2014; Starr et al., 2018). This epistasis can lead to shifts in a protein's amino-acid preferences over evolutionary time (Pollock et al., 2012; Doud et al., 2015; Shah et al., 2015; Bazykin, 2015; Haddox et al., 2017). Because the deep mutational scanning experiments that inform our ExpCM's were each performed in the context of a single HA genetic background, their measurements do not account for the accumulation of epistatic shifts in amino-

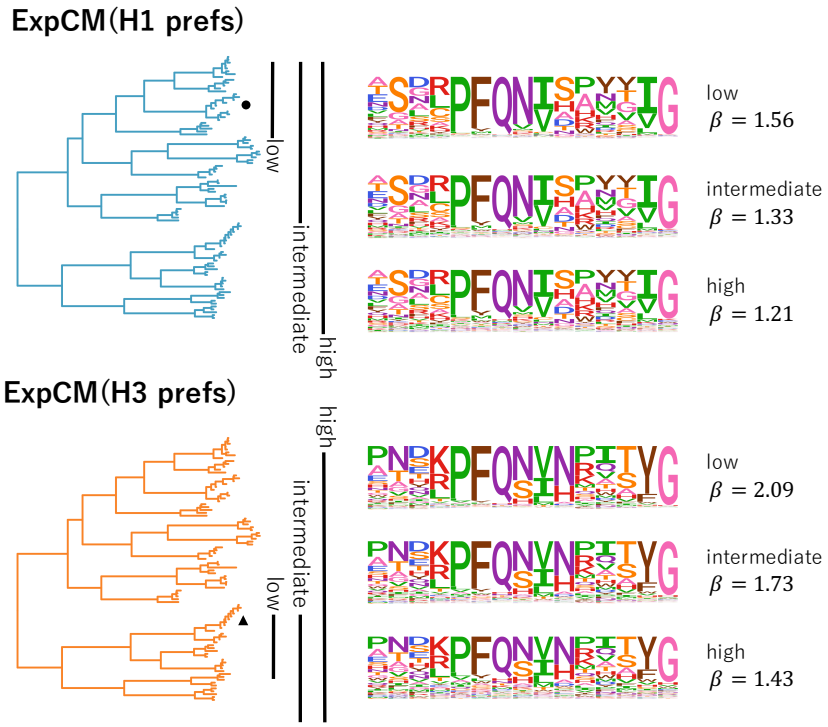


Figure 5: The congruence between natural selection and the deep mutational scanning measurements decreases with sequence divergence. We fit an ExpCM informed by the H1 or H3 deep mutational scanning experiments to trees spanning sequences with low, intermediate, and high divergence from the sequence used in the experiment. The ExpCM stringency parameter (β) is a measure of the congruence between natural selection and the experimental measurements (Bloom, 2014b; Hilton et al., 2017). Larger values of β indicate that natural selection prefers the same amino acids as the experiments but with greater stringency. As divergence increases between the HA used in the experiment and the other sequences in the tree, the β value decreases and the amino-acid preference "flatten." Therefore, the preferences measured in each experiment are progressively less congruent with natural selection as we include increasingly diverged sequences.

acid preferences as HA evolves. Therefore, an ExpCM is expected to most accurately describe the evolution of sequences closely related to the one used in the experiment.

We can observe how shifting amino-acid preferences degrade the accuracy of an ExpCM by fitting the model to trees containing increasingly diverged sequences. For both H1 and H3 HAs, we created three phylogenetic trees (Supplementary figure 5): a "low" divergence tree that contains sequences with $\geq$59% amino-acid identity to the HA used in the experiment, an "intermediate" divergence tree that contains sequences with $\geq$46% amino-acid identity to the HA in the experiment, and a "high" divergence tree that contains all HAs (which have as little as 38% identity to the HA in the experiment). Figure 5 shows the subtrees containing each of these sets of HA

sequences. For each subtree, we examined the congruence between site-specific natural selection and the amino-acid preferences measured in the deep mutational scanning experiment using the ExpCM stringency parameter β (Bloom, 2014b; Hilton et al., 2017). Values of β that are >1 indicate that natural selection prefers the same amino acids as the experiments but with a greater stringency, suggesting strong congruence between natural selection and the experimental preferences. In contrast, values of β that are <1 flatten the preferences, suggesting that they provide a relatively poor description of natural selection on the protein.

Figure 5 shows that as the divergence from the sequence used in the deep mutational scan increases, the value of β decreases. This inverse relationship between β and overall divergence is seen for the ExpCM's informed by both the H1 and H3 experiments. As β value decreases, the preferences "flatten" and so the ExpCM draws less information from the experiment. At the most extreme value of $\beta = 0$, the preferences would be perfectly uniform and look similar to the GY94 preferences in Figure 4. In reality, β never reaches a value this low, indicating the deep mutational scanning experiments remain somewhat informative about real natural selection across the entire swath of HAs. However, Figure 5 shows that the amino-acid preferences clearly become less informative about natural selection as we move away from the experimental sequence on the tree. This shifting of amino-acid preferences helps explain why the ExpCM informed by the average of the H1 and H3 experiments performs best (Table 1 and Figure 4): averaging the measurements across these two HAs is a heuristic method of accounting for shifts in preferences during HA evolution.

The fact that amino-acid preferences shift as a protein evolves leaves us with an inherent tension: models with site-specific amino-acid preferences only become important for accurate branch-length estimation as sequences become increasingly diverged, but this same divergence degrades the accuracy of extrapolating the amino-acid preferences from any given experiment across the phylogenetic tree. Crucially, this problem is expected to be more fundamental than the inability of a single deep mutational scanning experiment to measure amino-acid preferences in more than one genetic background. If amino-acid preferences shift during evolution, there simply will not be any model with a single set of time-homogeneous site-specific stationary states that accurately describes evolution along the entirety of a phylogenetic tree that covers a wide span of sequences.

A model with amino-acid preferences estimated from natural sequences give similar results to an ExpCM

The previous sections used ExpCM's, which are mutation-selection models that use site-specific amino-acid preferences that have been measured by experiments. However, there are other mathematically similar implementations of mutation-selection models that infer the amino-acid pref-

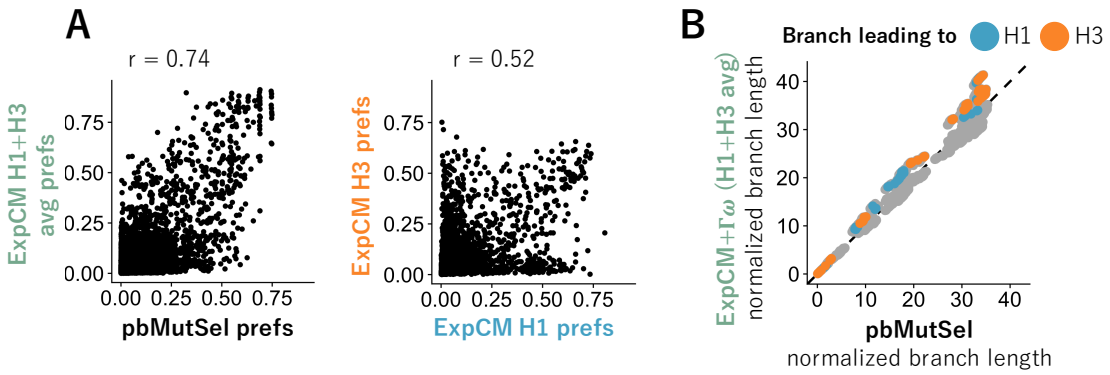


Figure 6: Models inferred from natural sequences have similar stationary states to models defined by experimental preferences and estimate similar branch lengths. [Move the general text that applies to the all the figure (e.g., the first two sentences below) to *before* the (A) label.] (A) We fit an ExpCM(H1+H3 avg)+$\Gamma\omega$ and a pbMutSel to the full HA tree in Figure 4. The pbMutSel amino-acid preferences are inferred from the natural HA sequences, while the ExpCM amino-acid preferences are experimentally measured and then rescaled by the stringency parameter in Table 1. The pbMutSel preferences are more correlated with the re-scaled average of the H1 and H3 deep mutational scanning preferences than the individual re-scaled H1 and H3 deep mutational scanning preferences are to each other (Pearson's r: 0.74 versus 0.52). (B) The ExpCM(H1+H3 avg)+$\Gamma\omega$ and pbMutSel models estimated similar branch lengths when fit to the entire HA tree. Points denote branch lengths between all pairs of tips on the tree. Blue and orange denote branches that lead to the H1 and H3 deep mutational scanning reference sequences respectively. The `phydms` program implementing ExpCM's and the `PhyloBayes-MPI` program implementing pbMutSel models give branch lengths in different units, so to facilitate direct comparison between the models, we have normalized all branch lengths returned by each program by the length of the branches separating the earliest (A/South Carolina/1918) and latest (A/Solomon Islands/2006) seasonal human H1 sequences on the tree.

erences directly from the natural sequence data. When these models are designed for use in phylogenetic inference, they are generally implemented in a Bayesian framework, which avoids the overfitting problems associated with trying to make maximum-likelihood estimates of the thousands of amino-acid preference parameters (Lartillot, 2014). (Note that the maximum-likelihood implementations of Tamuri et al. (2012, 2014) are designed for estimating the amino-acid preferences, *not* for phylogenetic inference.) The model most comparable to our ExpCM's is the codon mutation-selection model implemented in `PhyloBayes-MPI`, which we will refer to as pbMutSel (Rodrigue and Lartillot, 2014). In the pbMutSel model, the amino-acid preferences are modeled using Dirichlet processes rather than derived from experiments. However, like an ExpCM, a pbMutSel model still assumes a single set of time-homogeneous site-specific amino-acid preferences for the entire tree.

Comparing ExpCM and pbMutSel models can help determine the ultimate limits of mutation-selection models that assign each site a single set of amino-acid preferences. If the limitations of ExpCM's described above arise simply because the deep mutational scanning experiments do not correctly measure the "true" amino-acid preferences of HA across the entirety of a highly diverged phylogenetic tree, then we would expect the pbMutSel models (which infer these preferences from the entire tree) to perform better. On the other hand, if the major limitation is that no single set of time-homogenous amino-acid preferences can fully describe HA evolution over the entire tree, then we would expect ExpCM and pbMutSel models to perform similarly.

We fit a pbMutSel model to the entire HA phylogenetic tree, and compared the results to those from analyzing the same tree with the best ExpCM, which is the ExpCM(H1+H3 avg)+$\Gamma\omega$ variant. This is a direct apple-to-apples comparison, since the pbMutSel model also draws ω from a gamma-distribution (Rodrigue and Lartillot, 2014). [It is not clear how many bins `Phylobayes MPI` uses. I am going to go back and re-run the analysis specifying $k = 4$ bins.] First, we compared the amino-acid preferences inferred by the pbMutSel model to the preferences measured in the experiments. Figure 6A shows that the preferences inferred by pbMutSel are quite similar to the (H1+ H3 avg) obtained by averaging the deep mutational scanning measurements for the H1 and H3 HAs. Notably, the amino-acid preferences from the pbMutSel model are more correlated with the (H1+ H3 avg) than the H1 and H3 measurements are with each other (Figure 6A). This strong correlation indicates that the ExpCM(H1+H3 avg)+$\Gamma\omega$ is unlikely to be much different than a pbMutSel model that is parameterized only using the natural sequence data.

We next compared the branch lengths estimated by using the ExpCM(H1+H3 avg)+$\Gamma\omega$ and pbMutSel models. As shown in Figure 6B, these two models estimated similar branch lengths across the entire HA phylogenetic tree. However, the estimates are not identical, and the tension between local and global accuracy of the amino-acid preferences is still apparent. Specifically, the long branches between the H1 or H3 sequences used in the experiments and all other sequences were estimated to be slightly longer by the ExpCM, while many other branches were estimated to be slightly longer by the pbMutSel model. The relatively longer branches leading to the experimental sequences when using the ExpCM(H1+H3 avg)+$\Gamma\omega$ suggests that the "tree average" amino-acid preferences inferred by the pbMutSel model are not as accurate as the preferences from the deep mutational scanning for sequences close to those used in the experiments. However, for sequences distant from those used in the experiments, the "tree average" preferences inferred by the pbMutSel model appear to be slightly better than the experimental values. Therefore, while the ExpCM and pbMutSel models differ slightly in the extent to which they lengthen different branches, neither model can avoid the tension between the local and global accuracy of amino-acid preferences.

Discussion

[Add references to this section. Also, I think we are mostly using "preferences" rather than "stationary state", so use the former except in cases where the latter clearly makes more sense (e.g., talking about math rather than model.] Here, we tested the effect of models with site-specific stationary states on branch length estimation for the phylogenetic tree of highly diverged influenza HA sequences. We used site-specific stationary state models called ExpCM's, which are defined by amino-acid preferences measured by deep mutational scanning. We found that the ExpCM's estimated longer branches and did so independently from substitution rate variation via $\Gamma\omega$. We also found that the ExpCM's estimated branch lengths of a similar length to models which infer the stationary state from the natural sequences in a Bayesian framework.

These results underscore the importance of modeling site-specific amino-acid preferences when estimating long branchs. As the simulations Figure 3 show, models with uniform stationary states will always underestimate the lengths of branches which have evolved under site-specific constraints. The results with the influenza HA sequences shows that models with site-specific stationary states estimate longer branches when the stationary state is accurate. However, it is also clear that the accuracy of the ExpCM's stationary state degrades as the tree becomes more diverged from the HA used in the deep mutational scan. The desired effect of estimating long, accurate branches can only be achieved when the stationary state is accurate across the entire tree.

However, it is clear from our results that a major limitation of site-independent, time-homogenous models is their inability to capture epistatic interactions. The site-specific amino-acid preferences of a protein shift across time and models with a single stationary state across the entire tree cannot capture these dynamics. The ExpCM's defined by preferences measured in one of two highly diverged HA's had the largest effect on branches near the HA from the experiment. The local effect of the ExpCM's defined by a single preference set indicate that different parts of the tree have different stationary states. This point is underscored by the comparison of the ExpCM defined by the average preference set and the pbMutSel model, which infers the stationary state from the natural sequences. Even though the stationary state of the pbMutSel model is more representative of the global stationary state [maybe "tree average" rather than "global"], the ExpCM with average preferences has a larger effect on the branches from the H1 and H3 sequences. Therefore, while it is clear that modeling site-specific amino-acid preferences is important, neither one of these models is able to capture how these preferences shift over time.

[In particular, you need to add literature about how preferences shift. These can mostly be citations worked into above. In particular, there are two types of citations: papers about how preferences shift over time. This would include Hugh's paper and ones cited therein. You don't need to cite all general papers on epistasis, but just the ones that focus on this in terms of preferences.

Second, papers about the idea of modeling non-independence in phylogenetics. I believe that Richard Goldstein and David Pollock have a recent *Nature Ecology & Evolution* paper on this—look it up and also pull in relevant citations that they make. Also, there may be some paper in this area by Rodrigue and Lartillot. There are also a few cited in the third-to-last paragraph of discussion here (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4626756/).]

Clearly, the site-independence and time-homogeneity assumptions of the ExpCM and pbMutSel models inhibit their ability to accurately describe the effect of epistasis. In order to fully address the bias towards underestimation of divergence times by phylogenetic models, these models need to be able to accommodate how site-specific amino-acid preferences shift over time. One way to achieve this goal would be to void the site-independence assumption and create a model which explicitly describes the interactions between sites in a protein. However, while even modeling simple pairwise interactions would be a daunting task, epistasis could have higher order interactions which be extremely difficult to model. Another strategy would be to allow the stationary state to shift over lineages in the tree. Such a model might be still be site-independent but, by breaking the time-homogeneity assumption, would capture some effect of the shifting preferences. Models which account for epistasis would be inherently more complex but it clear from our results that it is important to capture how preferences shift over time to accurately estimate long branch lengths.

Methods

Substitution models

All of the substitution models used in this paper have been described previously. However, here we briefly recap their exact mathematical implementations.

GY94 model

The GY94 model is M0 variant of the Goldman-Yang model described by Yang et al. (2000). Specifically, the substitution rate P_{xy} from codon x to codon y is

$$P_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ by more than one nucleotide,} \\ \Phi_y & \text{if } \mathcal{A}\left(x\right) = \mathcal{A}\left(y\right) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transversion,} \\ \omega\Phi_y & \text{if } \mathcal{A}\left(x\right) \neq \mathcal{A}\left(y\right) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transversion,} \\ \kappa\Phi_y & \text{if } \mathcal{A}\left(x\right) = \mathcal{A}\left(y\right) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transition,} \\ \omega\kappa\Phi_y & \text{if } \mathcal{A}\left(x\right) \neq \mathcal{A}\left(y\right) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transition,} \\ -\sum\limits_{z \neq x} P_{xz} & \text{if } x = y, \end{cases} \quad \text{(Equation 1)}$$

where $\mathcal{A}\left(x\right)$ is the amino-acid encoded by codon x, κ is the transition-transversion rate, Φ_y is the equilibrium frequency of codon y, and ω is the relative rate of nonsynonymous and synonymous substitutions. We define the codon frequency parameters, Φ_y, using the "corrected F3X4" method from Pond et al. (2010). This method calculates the Φ_y values from the empirical alignment frequencies but corrects for the exclusion of sequences with premature stop codons from the analysis. [Also add a sentence saying how under these F3X4 there are 9 parameters defining Φ_x.]

The frequency p_x of codon x in the stationary state of a GY94 model is simply

$$p_x = \Phi_x. \quad \text{(Equation 2)}$$

Overall, a GY94 model has 11 free parameters: κ, ω, and the 9 nucleotide frequency parameters used to define Φ_y.

Experimentally Informed Codon Model (ExpCM)

The ExpCM models used in this paper are the ones described in Bloom (2017). Briefly, the rate of substitution $P_{r,xy}$ of site r from codon x to y is

$$P_{r,xy} = Q_{xy} \times F_{r,xy} \quad \text{(Equation 3)}$$

where Q_{xy} is proportional to the rate of mutation from x to y, $F_{r,xy}$ is proportional to the probability that this mutation fixes, and the diagonal elements P_{xx} are set by $P_{xx} = -\sum_{z \neq x} P_{xz}$.

The rate of mutation Q_{xy} is assumed to be uniform across sites, and takes an HKY85-like (Hasegawa et al., 1985) form as

$$Q_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ by more than one nucleotide,} \\ \phi_w & \text{if } x \text{ can be converted to } y \text{ by a transversion of a nucleotide to } w, \\ \kappa \times \phi_w & \text{if } x \text{ can be converted to } y \text{ by a transition of a nucleotide to } w \end{cases} \quad \text{(Equation 4)}$$

where ϕ_w is the nucleotide frequency of nucleotide w and κ is the transition-transversion rate.

The deep mutational scanning amino-acid preferences are incorporated into the ExpCM via the $F_{r,xy}$ terms. The experiments measure the preference $\pi_{r,a}$ of every site r for every amino-acid a. $F_{r,xy}$ is defined in terms of these experimentally measured amino-acid preferences as

$$F_{r,xy} = \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y), \\ \omega \times \frac{\ln\left[\left(\pi_{r,\mathcal{A}(y)} / \pi_{r,\mathcal{A}(x)}\right)^{\beta}\right]}{1 - \left(\pi_{r,\mathcal{A}(x)} / \pi_{r,\mathcal{A}(y)}\right)^{\beta}} & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y), \end{cases} \quad \text{(Equation 5)}$$

where β is the stringency parameter (Bloom, 2014b; Hilton et al., 2017) and ω is the relative rate of nonsynonymous to synonymous substitutions after accounting for the amino-acid preferences.

The stationary state of an ExpCM is

$$p_{r,x} = \frac{\phi_{x_1} \phi_{x_2} \phi_{x_3} \left(\pi_{r,A(x)}\right)^{\beta}}{\sum_z \phi_{z_1} \phi_{z_2} \phi_{z_3} \left(\pi_{r,A(z)}\right)^{\beta}} \quad \text{(Equation 6)}$$

where ϕ_{x_1}, ϕ_{x_2}, and ϕ_{x_3} are the nucleotides at position 1, 2, and 3 of codon x.

An ExpCM has five free parameters: κ, ω, and the three independent ϕ_x values. The amino-acid preferences $\pi_{r,a}$ are *not* free parameters since they are determined *a priori* by an experiment independent of the sequence alignment being analyzed.

$\Gamma\omega$ rate variation

The GY94+$\Gamma\omega$ is equivalent to the M5 model in Yang et al. (2000) with ω drawn from $K = 4$ categories. The ExpCM+$\Gamma\omega$ similarly draws ω from a Γ distribution discretized into $K = 4$ bins. Each bin is equally weighted and ω takes on the mean value of the bin. Because the Γ distribution is defined by two parameters, adding $\Gamma\omega$ to a model with a single ω adds one free parameter. Therefore, the GY94+$\Gamma\omega$ model has 12 free parameters, and the ExpCM+$\Gamma\omega$ model has 6 free parameters.

GY94 with ω_r

In Figure 2, we describe GY94 models where each site r has its own ω_r value that is calculated from the amino-acid preferences using the relationship described by Spielman and Wilke (2015b). This relationship defines the expected rate of nonsynonymous to synonymous substitutions given the amino-acid preferences. We first fit an ExpCM to the "low divergence" H1 subtree (parameter values in Supplementary table 2), which allows us to calculate $P_{r,xy}$ (Equation 3), Q_{xy} (Equation 4), and $p_{r,x}$ (Equation 6). We then calculated ω_r using the equation of Spielman and Wilke (2015b), normalizing by the gene-wide ω fit by the ExpCM:

$$\omega_r = \frac{\sum_x \sum_{y \in N_x} p_{r,x} \times \frac{P_{r,xy}}{\omega}}{\sum_x \sum_{y \in N_x} p_{r,x} \times Q_{xy}}, \qquad \text{(Equation 7)}$$

where N_x is the set of codons that are nonsynonymous to codon x and differ from codon x by only one nucleotide.

HA amino-acid preferences from deep mutational scanning experiments

We used amino-acid preferences measured in deep mutational scans of the A/WSN/1933 H1 HA (Doud and Bloom, 2016) and the A/Perth/2009 H3 HA (Lee et al., 2018) to define the amino-acid preferences that inform the ExpCM's. We only used sites that can be unambiguously aligned in these H1 and H3 HAs. These alignable sites and their mapping to sequential numbering of the HA sequences used in the deep mutational scanning experiments are in Supplementary file 1. The experimentally measured amino-acid preferences masked to just include these alignable sites are in Supplementary file 2 and Supplementary file 3. For the average preference set, we took the pairwise average of the H1 and H3 preferences. The preference for every amino acid a at every site r in the average preference set is

$$\pi_{r,a,(\text{H1+H3 avg})} = \frac{\pi_{r,a,\text{H1}} + \pi_{r,a,\text{H3}}}{2} \qquad \text{(Equation 8)}$$

HA sequences and tree topology.

We downloaded all full-length, coding sequences for 15 of the 18 Influenza A virus HA subtypes from the Influenza Virus Resource Database (Bao et al., 2008) in June of 2017. We excluded rare subtypes 15, 17, and 18, which have limited sequences in the database. We filtered and aligned the sequences using `phydms_prepalignment` (Hilton et al., 2017). Specifically, we used `phydms_prepalignment` with the flag `--minidentity 0.3` to remove sequences with ambiguous nucleotides, premature stops, or frameshift mutations as well as redundant sequences. We also removed all codon sites which that are not alignable between the H1 HA and H3

HA used in the deep mutational scanning experiments (these alignable sites are listed in Supplementary file 1). We subsampled the remaining sequences to five per subtype with ≤ 1 sequence per year per subtype. We also included a small number of sequences from the major human and equine influenza lineages to ensure representation of these well-studied lineages. The resulting alignment contains 92 sequences, and is provided in Supplementary file 4.

We created four subalignments with "low" and "intermediate" divergence from either the H1 or the H3 deep mutational scanning reference sequence for the analysis in Figure 5. The "low divergence" alignments had $\geq 59\%$ amino-acid identity to the sequence used in the deep mutational scanning, and the "intermediate divergence" alignments had $\geq 46\%$ identity from the reference sequence (Supplementary figure 5).

We inferred the tree topology of each alignment using `RAxML` (Stamatakis, 2006) and the GTRCAT model. We estimated the branch lengths of this fixed topology using each ExpCM and GY94 models with `phydms_comprehensive` (Hilton et al., 2017).

Asymptotic amino-acid sequence identity

For the analysis in Figure 2, we fit models to the "low divergence" H1 subtree. This gave the parameter values in Supplementary table 2.

For each model, we calculated the expected amino-acid sequence identity for two sequences separated by a branch length of t as

$$\sum_a \sum_{x \in a} p_{r,x} \sum_{y \in a} \left[e^{t\mathbf{P_r}}\right]_{xy} \qquad \text{(Equation 9)}$$

where a ranges over all 20 amino acids, $x \in a$ indicates that x ranges over all codons that encode amino-acid a, $p_{r,x}$ is the stationary state of the model at site r and codon x (given by Equation 2 for GY94-family models, and Equation 6 for ExpCM-family models), and $\left[e^{t\mathbf{P_r}}\right]_{xy}$ is the value in row x and column y of the matrix obtained by exponentiating the product of t and the substitution matrix $\mathbf{P_r}$ for site r (defined by Equation 1 for GY94-family models and Equation 3 for ExpCM-family models).

Simulations

For Figure 3, we simulated sequences using `pyvolve` (Spielman and Wilke, 2015a) along the full HA tree using an ExpCM defined by parameters fit to the "low divergence" H1 subtree (Supplementary table 2). We performed 10 replicate simulations and estimated the branch lengths for each replicate using `phydms_comprehensive` (Hilton et al., 2017).

pbMutSel inference with `PhyloBayes-MPI`.

For Figure 6, we fit a pbMutSel model to the full HA tree. We ran one chain for 5000 steps, saved every sample, and discarded the first 500 samples as a burnin. [re: measure of convergence. I need add some analysis to fully address this point. 1. I am going to run 2 chains and assess the congruency between these two chains. 2. I am going to look at the traces using Andrew Rambaut's program tracer. Both of these are in progress] We used `PhyloBayes-MPI` program `readpb_mpi` to compute the majority-rule consensus tree and the posterior average site-specific amino-acid preferences.

In order to make the branch lengths in Figure 6 comparable between the pbMutSel tree returned by `PhyloBayes-MPI` and the other trees returned by `phydms`, we we normalized the branch lengths on the pbMutSel consensus tree and the ExpCM(H1+H3 avg)+$\Gamma\omega$ by dividing each branch by the length from A/South Carolina/1/1918 and A/Solomon Islands/3/2006. These two H1 sequences are early and late representatives of the longest known human influenza lineage, and are of sufficiently high identity that different ExpCM and GY94 substitution models all estimate nearly identical branch lengths separating them.

Software versions and computer code

All code used for the analyses in this paper is available at `https://github.com/jbloomlab/divergence_timing_manuscript`. [make repo public] The external computer programs that we used were

- `phydms` (Hilton et al., 2017) version 2.2.2 (available at `github.com/jbloomlab/phydms`) to fit the ExpCM and GY94 models.
- `pyvolve` (Spielman and Wilke, 2015a) version 0.8.7 (available at `https://github.com/sjspielman/pyvolve`) to simulate the sequences.
- `PhyloBayes-MPI` (Rodrigue and Lartillot, 2014) version 1.8 (available at `https://github.com/bayesiancook/pbmpi`) to fit the pbMutSel model.
- `RAxML` (Stamatakis, 2006) version 8.2.11 (available at `https://github.com/stamatak/standard-RAxML` to infer tree topology.
- We used `ggplot2` (Wickham, 2016), `ggtree` (Yu et al., 2017), and `ggseqlogo` (Wagih, 2017) for visualization of the results.

References

Arenas M. 2015. Trends in substitution models of molecular evolution. *Frontiers in genetics*. 6:319.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. 2008. The influenza virus resource at the national center for biotechnology information. *Journal of virology*. 82:596–601.

Bazykin GA. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biology letters*. 11:20150315.

Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A. 2014. Integrating influenza antigenic dynamics with molecular evolution. *eLife*. 3:e01914.

Bloom JD. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution*. 31:1956–1978.

Bloom JD. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol. Biol. Evol.* 31:2753–2769.

Bloom JD. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*. 12:1.

Doud MB, Ashenberg O, Bloom JD. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.* 32:2944–2960.

Doud MB, Bloom JD. 2016. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses*. 8:155.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular Biology and Evolution*. 11:725–736.

Gong LI, Suchard MA, Bloom JD. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*. 2:e00631.

Ha Y, Stevens DJ, Skehel JJ, Wiley DC. 2002. H5 avian and h9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. *The EMBO journal*. 21:865–875.

Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. 2017. Mapping mutational effects along the evolutionary landscape of hiv envelope. *bioRxiv*. p. 235630.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*. 15:910–917.

Harms MJ, Thornton JW. 2014. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*. 512:203–207.

Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*. 22:160–174.

Hilton SK, Doud MB, Bloom JD. 2017. phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*. 5:e3657.

Lartillot N. 2014. The Bayesian Kitchen: overcoming the fear of over-paramerization. http://bayesiancook.blogspot.com/2014/01/the-myth-of-over-parameterization.html. Last accessed: March-12-2018.

Lartillot N, Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*. 21:1095–1109.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B*. 363:3965–3976.

Lee JM, Huddleston J, Doud MB, Hooper K, Wu NC, Bedford T, Bloom JD. 2018. Deep mutational scanning of hemagglutinin helps identify evolutionarily successful human h3n2 influenza viruses. *in prep*. .

McCandlish DM, Stoltzfus A. 2014. Modeling evolution using the probability of fixation: history and implications. *The Quarterly review of biology*. 89:225–252.

Murrell B, Weaver S, Smith MD, et al. (11 co-authors). 2015. Gene-wide identification of episodic selection. *Molecular Biology and Evolution*. 32:1365–1371.

Nielsen R. 2006. Statistical methods in molecular evolution. Springer.

Nobusawa E, Aoyama T, Kato H, Suzuki Y, Tateno Y, Nakajima K. 1991. Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza a viruses. *Virology*. 182:475–485.

Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*. 317:1544–1548.

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary stokes shift. *Proc. Natl. Acad. Sci. USA*. 109:E1352–E1359.

Pond SK, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*. 5:e11230.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*. 53:793–808.

Quang SL, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*. 24:2317–2323.

Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*. 193:557–564.

Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*. 30:1020–1021.

Rodrigue N, Lartillot N. 2017. Detecting adaptation in protein-coding genes using a bayesian site-heterogeneous mutation-selection codon substitution model. *Molecular Biology and Evolution*. 34:204–214.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*. 107:4629–4634.

Russell R, Gamblin S, Haire L, Stevens D, Xiao B, Ha Y, Skehel J. 2004. H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes. *Virology*. 325:287–296.

Shah P, McCandlish DM, Plotkin JB. 2015. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*. 112:E3226–E3235.

Spielman SJ, Wilke CO. 2015a. Pyvolve: a flexible Python module for simulating sequences along phylogenies. *PLoS One*. 10:e0139047.

Spielman SJ, Wilke CO. 2015b. The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*. 32:1097–1108.

Stamatakis A. 2006. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690.

Starr TN, Flynn JM, Mishra P, Bolon DN, Thornton JW. 2018. Pervasive contingency and entrenchment in a billion years of hsp90 evolution. *bioRxiv*. p. 189803.

Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*. 190:1101–1115.

Tamuri AU, Goldman N, dos Reis M. 2014. A penalized likelihood method for estimating the distribution of selection coefficients from phylogenetic data. *Genetics*. pp. genetics–114.

Tufts DM, Natarajan C, Revsbech IG, Projecto-Garcia J, Hoffmann FG, Weber RE, Fago A, Moriyama H, Storz JF. 2014. Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Molecular Biology and Evolution*. 32:287–298.

Wagih O. 2017. ggseqlogo: a versatile r package for drawing sequence logos. *Bioinformatics*. 33:3645–3647.

Wang HC, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC evolutionary biology*. 8:331.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–314.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*. 25:568–579.

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155:431–449.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 8:28–36.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins. New York, NY: Academic Press, pp. 97–166.

Supplemental Information

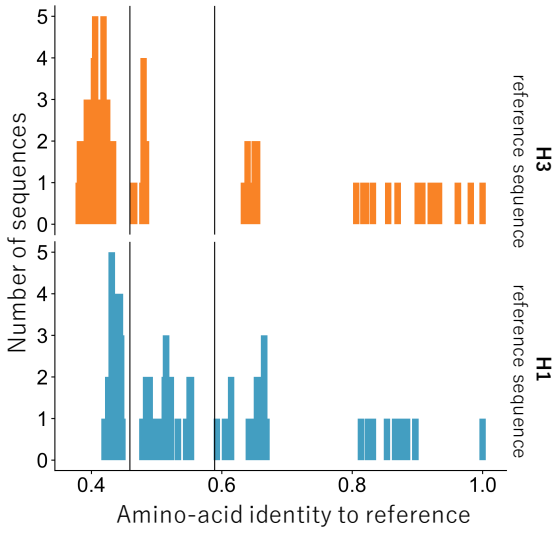
[Further increase the number of rows in the logo plots to make them larger. Since we aren't under length constraints for supplement, might as well make them readable. Maybe 8 or so rows instead?]

Supplementary figure 1: H1 HA amino-acid preferences measured by deep mutational scanning. Each column represents a site in the HA protein, and the height of each letter is proportional to the preference for the amino acid measured by Doud and Bloom (2016) and then re-scaled by the stringency parameter in Table 1. The plot only shows sites that are alignable between the H1 and H3 HAs, and these alignable sites are numbered sequentially starting from 1. The conversion between the numbering scheme in this figure and sequential numbering of the H1 HA reference sequence is in Supplementary file 1.

Supplementary figure 2: H3 HA amino-acid preferences measured by deep mutational scanning. Similar to Supplementary figure 1 but shows the re-scaled preferences for the H3 HA as measured by Lee et al. (2018).

Supplementary figure 3: Average of H1 HA and H3 HA amino-acid preferences measured by deep mutational scanning. Similar to Supplementary figure 1 but shows the re-scaled average of the preferences for the H1 and H3 HAs.

Supplementary figure 4: Amino-acid preferences inferred by the pbMutSel model. Similar to Supplementary figure 1, but shows the preferences inferred by fitting the pbMutSel model to the full HA tree.



Supplementary figure 5: Overall divergence for subtrees. We created two subalignments for each HA used in the deep mutational scanning experiments. The "low divergence" alignments had $\geq$59% amino-acid identity to either the H1 or H3 reference sequence. The "intermediate divergence" alignments had $\geq$46% amino-acid identity to the reference sequences.

Model	Tree diameter (average codon substitutions per site)	Percentage of GY94 tree diameter
GY94	12.04	100%
ExpCM(H1)	14.70	122%
ExpCM(H3)	16.28	135%
ExpCM(H1+H3 avg)	19.21	160%
GY94 + $\Gamma\omega$	19.15	159%
ExpCM(H1) + $\Gamma\omega$	24.75	206%
ExpCM(H3) + $\Gamma\omega$	25.03	208%
ExpCM(H1+H3 avg) + $\Gamma\omega$	30.78	256%

Supplementary table 1: Branch length extension as measured by tree diameter. We calculated the tree diameter, the distance between the two most divergent tips, for the trees in Figure 4. For each tree, the diameter is reported as a raw value and as a percentage of the GY94 model tree, the smallest of the eight trees.

Model	Parameters
GY94	$\kappa = 3.17$, $\omega = 0.10$, $\phi_{1,A} = 0.32$, $\phi_{1,C} = 0.14$, $\phi_{1,G} = 0.28$, $\phi_{2,A} = 0.38$, $\phi_{2,C} = 0.18$, $\phi_{2G} = 0.20$, $\phi_{3,A} = 0.36$, $\phi_{3,C} = 0.19$, $\phi_{3,G} = 0.21$
GY94 + $\Gamma\omega$	$\alpha_\omega = 0.51$, $\beta_\omega = 3.92$, $\kappa = 3.49$, $\phi_{1,A} = 0.32$, $\phi_{1,C} = 0.14$, $\phi_{1,G} = 0.28$, $\phi_{2,A} = 0.38$, $\phi_{2,C} = 0.18$, $\phi_{2,G} = 0.20$, $\phi_{3,A} = 0.36$, $\phi_{3,C} = 0.19$, $\phi_{3,G} = 0.21$
ExpCM(H1)	$\beta = 1.56$, $\kappa = 3.64$, $\omega = 0.24$, $\phi_A = 0.378$, $\phi_C = 0.17$, $\phi_G = 0.23$

Supplementary table 2: Model parameters fit to a low divergence tree. We fit GY94 models and an ExpCM defined by H1 deep mutational scanning preferences to the "low divergence from H1" tree in Figure 5. We used these model parameters calculate the expected pairwise sequence identity in Figure 2 and simulate the sequences in Figure 3.

Supplementary file 1: List of alignable sites between H1 HA and H3 HA. This files provides a conversion between the numbering scheme we use in the paper (sequential numbering of just the alignable sites) to sequential numbering of the H1 HA reference sequence A/Wilson Smith/1933 and the H3 HA reference sequence A/Perth/2009.

Supplementary file 2: Amino acid preferences measured by the deep mutational scanning of the H1 HA strain A/WSN/1933 (Doud and Bloom, 2016). This file only contains measurements for the alignable sites between H1 and H3 HAs. Conversion from this numbering scheme to sequential numbering of A/WSN/1933 is in Supplementary file 1.

Supplementary file 3: Amino acid preferences measured by the deep mutational scanning of the H3 HA strain A/Perth/2009 (Lee et al., 2018). This file only contains measurements for the alignable sites between H1 and H3 HAs. Conversion from this numbering scheme to sequential numbering of A/Perth/2009 is in in Supplementary file 1.

Supplementary file 4: The HA sequences for the full HA tree. The sequences in this alignment contain only the alignable sites between H1 and H3 HAs. Conversion from this numbering scheme to sequential numbering of A/Perth/2009 is in in Supplementary file 1.

[Notes from Sarah to Jesse]

[**(1) Within subtype amino acid identity.** When describing the structure of the HA tree we comment that within a subtype HA sequences are $\geq 68\%$ identical on an amino acid level. To calculate this number I didn't want to look just at the tree of 92 sequences we used for the analysis because we might have missed the most diverged pair of sequences within a subtype. I also didn't want to take a brute force approach and calculate the pairwise identity of *every* sequence I downloaded fro IVR or build a tree with all of these sequences. To get the number 68% I took 1 million samples of pairs of sequence for each subtype and calculated the mean divergence between these 18 million samples. Do you think this is sufficient? Is there an alternative way which makes more sense?]

[This seems fine, except you should take the *minimum* rather than the *mean* divergence between these samples. Did you actually take minimum and just have typo in your comment above?]

[**(2) The asymptotic sequence divergence of the GY94 model.** In the discussion of the decay to stationary state plot, we comment that the asymptotic sequence divergence of the GY94 model is 0.059. I got this value by simply putting a very large value of t (time) into the equation. I tried

to find the true asymptotic divergence by solving the limit of the equation as t approaches infinity. When I did this I got a value of 0.0193. This is too low. I've attached my notes, is there a place where my math is obviously wrong?]

[It is not true that all elements of $\mathbf{D}$ are < 0. In fact, the principle eigenvector of a irreducible and acyclcic stochastic matrix is 1, and since $\mathbf{D}$ is a stochastic matrix minus the identity matrix, that means that it has one eigenvalue that is 0 and the rest are < 0. I'm not sure how all the math works out, but I know that it should end up being the case that $\left[\lim_{t\to\infty} e^{t\mathbf{P_r}}\right]_{xy} = p_{r,x}$ (see https://en.wikipedia.org/wiki/Stochastic_matrix).

So the correct formula for the asymptotic sequence divergence at site r must be

$$\sum_a \sum_{x\in a} p_{r,x} \sum_{y\in a} p_{r,y}. \qquad \text{(Equation 10)}$$

You can see that this is true without even doing all the complicated matrix exponential limits: at stationary state, the probability of being x at site r is $p_{r,x}$.]

[**(3) pbMutSel re-run.** I am going to re-run the pbMutSel analysis making two major changes. The first is I am going to specify the number of bins for the gamma distribution so there are exactly the same number of bins as in the ExpCM+$\Gamma\omega$ ($K = 4$). The second change I am going to make is to run two parallel chains. I will use the amount of congruency between the parallel chains along with the trace files to assess the convergence of the model. The server is going to be undergoing some maintenance this weekend so I might not have the results until mid next week. However, when I was first getting the pbMutSel model up and running I ran two chains in parallel on a similar (but not identical) tree and the results were very congruent. That is, I am not too worried about the results changing in the pbMutSel section.]

[OK, sounds fine.]