Experimentally informed site-specific substitution models deepen phylogenetic estimates of the divergence of viral lineages

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

[1]Division of Basic Sciences and Computational Biology Program,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA
E-mail: jbloom@fredhutch.org.

Abstract

Molecular phylogenetics is often used to estimate the time since the divergence of modern gene sequences. Such phylogenetic techniques often estimate substantially shallower divergence times than other methods. For instance, in the case of viruses there is independent evidence that molecular phylogenetics can underestimate deep divergence times. This discrepancy is thought to be caused in part by inadequate models of purifying selection leading to branch-length underestimation. Here we show that substitution models informed by experimental measurements of the purifying selection due to site-specific amino-acid preferences lengthen the deep branches on phylogenies of influenza virus hemagglutinin. This deepening of branch lengths is due to better modeling of the stationary state of the substitution models, and is independent of the branch-length-extension that results from modeling site-to-site variation in substitution rate. The deepening of branch lengths from experimentally informed site-specific substitution 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 models are limited by the inherent tension between the enhanced accuracy of accounting for site-specific amino-acid preferences and the fact that these preferences shift over long evolutionary times. Overall, our work underscores the importance of modeling how site-specific purifying selection affects the stationary state when estimating deep divergence times.

Introduction

[from JDB: what is the "age" of a virus? Maybe "divergence time of viral lineages"] skhcommentfrom JDB: what is the less than a million actually? "Old" is not the right phrase. Estimating the divergence time of viral lineages of a virus is essential to understanding its evolutionary history, including its emergence, spread, and past zoonoses. This estimation is commonly done using the concept a "molecular clock" to transform the branch lengths of the viral phylogenetic tree into age in years. However, this molecular dating technique often underestimates the age of many viruses, including measles, foamy virus, and ebola [(citations)], compared to other methods which are independent of the viral phylogeny. For example, SIV (the original source of HIV) is estimated to be less than a million years old based on the viral phylogeny (Sharp et al., 2000; Wertheim and Worobey, 2009; Worobey et al., 2010) but estimated to be several million years old based on the host tree or endogenous retroviral elements (Compton et al., 2013) [(other citations)]. Overall, there is a systematic and substantially large underestimation of of branch length on viral phylogenies. [long branches]

Branch length underestimation is due, in part, to strong purifying selection masking the evolutionary signal in the observed sequences. Purifying selection can lead to mutational saturation, where multiple unobserved, substitutions occur at a single site along a long branch and erase the divergence signal (Holmes, 2003). Furthermore, proteins do not have equal preference for all amino acids at all sites, this evident by a simple visual inspection of a multiple sequence alignment. How many and which amino acids tolerated at each site of the protein generate a site-specific expected rate of change. Failing to account for these site-specific constraints will lead to branch length underestimation. [you will have mutational saturation no matter what - this is a separate, addressable issue?] [talk about the high mutation rate in viruses?]

Substitution models that incorporate site-to-site rate variation have been developed to decrease the bias in long branch estimation. The most common strategy is to allow a single rate-controlling parameter to vary according to some statistical distribution, such as a Γ-distributed ω (dN/dS) (Yang et al., 2000). This flexibility in the value of ω accounts for the site-to-site rate variation by allow some sites to have a higher dN/dS value than others. While this modification is simple and only requires the addition of one extra parameter, it does not describe site-specificity in its stationary state. That is, at evolutionary equilibrium, this model still assumes that each site in the protein evolves identically.

An alternative approach is to model the site-specific amino-acid frequencies explicitly, such as those models in the mutation-selection family (Halpern and Bruno, 1998). In these models, each amino-acid at each site in the protein is described by its own parameter and these differences are reflected in the stationary state of the model. The rate of change at a given site is controlled by these amino acid profiles and can now vary from site to site, as expected based on observations in nature. Importantly, these rate variations are not constrained to an arbitrary statistical distribution but by parameters with a direct

biological interpretation.

Mutation-selection models are presumably more biologically relevant but pose more practical challenges than the $\Gamma\omega$ models. These models are highly parametrized with 19 free parameters (the 20 amino acid preferences are constrained to sum to one) per site leading to thousands of parameters for the length of a normal protein. One way to avoid overfitting is to implement the model as a mixture model in either a bayesian (Lartillot and Philippe, 2004) or maximum likelihood framework (Si Quang et al., 2008).

Alternatively, you can reduce the parameter space by defining the amino-acid frequencies *a priori*. We have shown previously that we can define an Experimentally Informed Codon Model (ExpCM) (Bloom, 2014a,b) from the mutation-selection family using measurements from deep mutational scanning (Fowler and Fields, 2014), a high-throughput functional assay. ExpCM are therefore defined by amino-acid preferences measured in a *single* genetic background and do not reflect any epistatic changes which may have occurred over the virus's evolutionary history. But they contain no more parameters than the traditional codon models while maintaining a site-specific stationary state. We hypothesize that the ExpCM will estimate longer branches than the traditional models due to the protein-specific description of purifying selection. [CAT model has been shown to work well (better) on saturated data.]

In order to test this hypothesis, we compared the branch lengths of a influenza virus HA phylogenetic trees optimized by different substitution models. We found that the ExpCM did extend the length of branches from the focal sequence on the tree [define focal] and that this extension was seen even in the context of Γ-distributed rate variation. Furthermore, we found this extension occurred even in the presence of Γ-distributed ω, indicating that they are both important for modeling purifying selection. This supports the conclusion that modeling purifying selection, especially in a model with a non-uniform stationary state, is important to estimating the branch lengths on phylogenetic trees.

Results and Discussion

Different ways hat substitution models account for purifying selection

[Some other comments on this section, which I think in general is pretty good: ω shows up in the figure, but is never mentioned in the text. I feel like GY94 models need to be explained at least in terms of what they and Γ distribution stand for. Just introducing nomenclature. I feel like the last paragraph should tie back to panel C of the figure.]

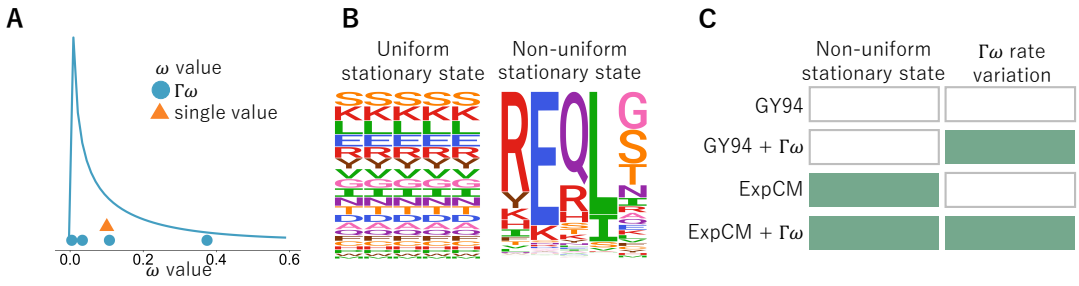


Figure 1: Different ways of modeling site-to-site variation in purifying selection. (A) The relative rate of non-synonymous change, ω, can be defined as one, gene-wide average or allowed to vary following some statistical distribution. In order to achieve computational tractability, the distribution is discretized into K bins and ω is set to the mean of each bin. A Γ distribution with $K = 4$ bins is shown here. (B) Substitution model stationary states can either be identical at every site in the protein or allow to vary from site to site. (C) Substitution models can both, one, or neither of these features and we use models from the GY94 and ExpCM families to represent all possible combinations.

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

The most common strategy to model rate variation is to allow the rate of non-synonymous change to vary among sites following some statistical distribution. For example, in the M5 of Yang et al. (2000) the implied dN/dS parameter, ω, follows a gamma distribution (GY94+$\Gamma\omega$, Figure 1A). Under the GY94+$\Gamma\omega$ and other such models, some sites have a higher rate of non-synonymous change than others. However, this rate is agnostic to the amino-acid identities themselves, and so all non-synonymous changes at at given site are treated equally.

In contrast, so-called "mutation-selection" models (Halpern and Bruno, 1998) account for purifying selection by explicitly defining a different set of amino-acid preferences at each site in the protein. This more mechanistic formulation, without an arbitrary statistical distribution, results in a site-specific stationary state (Figure 1B). These models better capture the site-to-site differences in amino-acid composition that is an obvious features of real proteins, and indeed they generally better describe actual evolution than models with only rate variation (Lartillot and Philippe, 2004; Le et al., 2008; Rodrigue et al., 2010; Hilton et al., 2017; Bloom, 2014a). The specificity of mutation-selection models comes at a cost in the form of an increased number of parameters. While codon substitution models with uniform stationary states typically have $<$20 parameters, mutation-selection models must specify 19 parameters for *each* site (the stationary state is for 20 amino acids whose frequencies are constrained to sum to one).

This corresponds to $19 \times L$ parameters for a protein of length L, or $\sim 10^4$ parameters for a typical size protein. As with all parameter-rich models, it is important to obtain values for these parameters using some method that avoids overfitting [cite rodrigue paper]. Here we will primarily use experimentally informed codon models (ExpCM) (Bloom, 2014a; Hilton et al., 2017; Bloom, 2017) which define values for these parameters *a priori* from deep mutational scanning experiments (Araya and Fowler, 2011; Fowler et al., 2010) so they do not need to be fit from phylogenetic data. Therefore, the number of remaining free parameters that are fit from the phylogenetic data for an ExpCM are similar to a non-site-specific substitution model. Alternative strategies of obtaining parameters for site-specific stationary states via Bayesian[cite] or maximum-likelihood estimation[cite] are discussed in the last section of the Results.

Finally, these two strategies to account for purifying selection, $\Gamma\omega$ rate variation and site-specific stationary states, are not mutually exclusive. Some proteins may be better modeled as a combination of the two. Therefore, we will use models from the GY94 (uniform stationary state) and ExpCM (site-specific stationary state) families with and without $\Gamma\omega$ rate variation (Figure 1C) to examine the effects of each strategy on branch length estimation.

Effect of stationary state and rate variation on branch length estimation

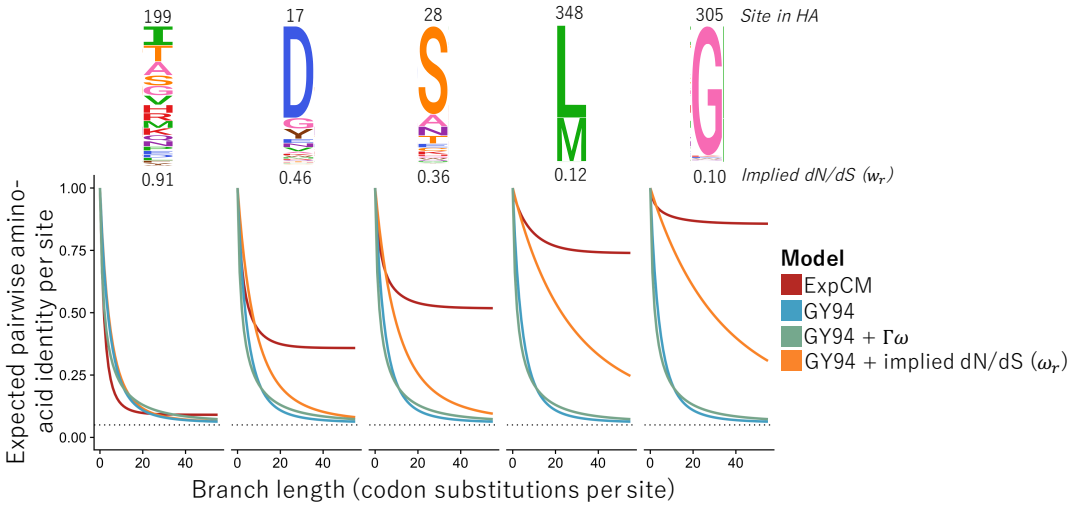


Figure 2: Effect of stationary state and rate variation on long branch estimation. The expected pairwise amino-acid identity for five sites in influenza hemagglutinin (HA) for four different substitution models calculated using Equation 6. The logoplots show the HA amino-acid preferences from a deep mutational scan performed by Doud and Bloom (2016). The implied dN/dS value was calculated from the ExpCM following Spielman and Wilke (2015) (Equation 5). [Maybe change long branch length to expected asymptotic sequence divergence.]

Given a single branch, a substitution model transforms sequence divergence into branch length, which is proportional to time under a molecular clock assumption. All phylogenetic substitution models [all?] are stochastic matrices and, as such, reach stationary state. Stationary state describe the expected sequence composition after a very long evolutionary time and this stationary state sequence composition is independent to time. In Figure 2, the stationary state is represented by the long "tails" where the expected sequence divergence remains constant as time increases.

By definition, the expected sequence composition varies from site to site for the site-specific stationary state models (Figure 2 ExpCM, red) but is constant across sites for the uniform stationary state models (Figure 2 GY94, blue). The biggest difference in stationary state between the models is at constrained sites, such as site 305 in Figure 2. At such a site, the ExpCM would estimate a much longer branch than the GY94 given the observed divergence of a set of sequences.

Rate variation, as modeled by $\Gamma\omega$, affect the rate at which stationary state is approached by not the sequence composition at stationary state itself. Stationary state is a fundamental feature of stochastic matrices and is invariant to multiplicative terms, such as ω. In Figure 2, GY94+$\Gamma\omega$ (green) takes longer to reach the stationary state than GY94 (blue) but stationary state for each model is identical.

This means even complex modeling of the ω to account for purifying selection will not affect the expectation at evolutionary equilibrium. We inferred a site-specific ω value from the ExpCM (Spielman and Wilke, 2015) and applied these values to the GY94 model, GY94+$\Gamma\omega$ (Figure 2 yellow). At constrained sites, this model takes even longer to reach stationary state than the GY94+$\Gamma\omega$ but the eventual stationary state is identical to both the GY94 and GY94+$\Gamma\omega$ stationary state.

No matter how "well" a model accounts for site-to-site rate variation, it will underestimate long branches with a uniform stationary state.

Failure to account for site-specificity leads to branch length underestimation.

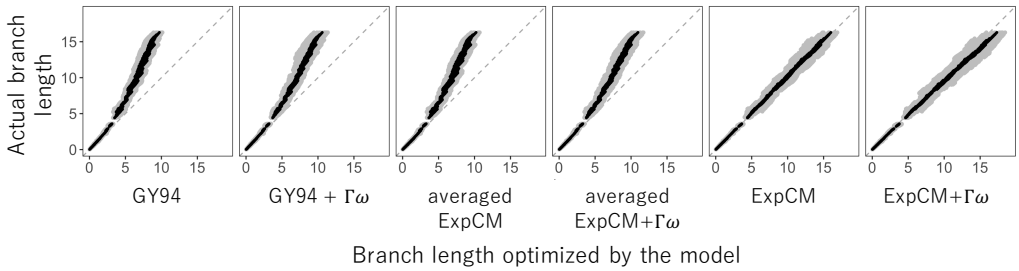


Figure 3: Model performance under simulated, site-specific data. Alignments were simulated under an ExpCM along an HA tree and the branches were re-optimized by a model from the ExpCM or YNGKP family. The averaged ExpCMs amino-acid frequencies defined by the average preference of that amino acid across all sites in the protein. While these models extract information from the deep mutational scanning experiment, they are not site-specific. Grey points represent the length of one branch and the black points are the mean branch lengths over ten simulations. The grey, dashed line is the reference line $y = x$, depicting the behavior of a model which is an unbiased estimator of the simulated branch length.

Next, we wanted to test the effect of substitution model on branch length estimation given sequences with site-specific amino-acid frequencies. To this end, we simulated sequences under an ExpCM and re-inferred the branch lengths using the models in Figure 1C.

Models with uniform stationary state underestimated the branch lengths of the simulated sequences. This underestimation is most pronounced on the long branches, with the GY94 estimating branches which are only $\sim \frac{2}{3}$ of the true branch length (Figure 3). Accounting for the site-to-site rate variation by $\Gamma\omega$ does not prevent this underestimation as the GY94+$\Gamma\omega$ estimates roughly the same branch length as the GY94. We also estimated branch lengths using an average ExpCM. These models have a uniform stationary state defined by the average preference of amino acid across all sites. Even these models, which share some information with the model used to simulate the sequences, underestimate the long branches.

As expected neither the ExpCM nor the ExpCM+$\Gamma\omega$ underestimate the long branches. However, the variance between simulations increases as the branch length increases, though this error is not biased towards over- or underestimation. This increase in error despite a perfect model match simply underscores the difficulty of estimating extremely long branches. Despite this universal difficulty, it is clear that models with uniform stationary states will always drastically underestimate the long branch lengths.

Table 1: Model comparison for Fig. Figure 4.[GY94+Γω has one ω of 0 because of rounding.]

Model	Stationary State	Γω	ΔAIC	Log Likelihood	ω (implied dN/dS)	stringency parameter
ExpCM + Γω (H1+H3 avg)	yes	yes	0	-487510	0.19, 0.50, 0.90, 1.86	1.70,
ExpCM (H1+H3 avg)	yes	no	950	-492270	0.15	1.78
ExpCM + Γω (H1)	yes	yes	13060	-494040	0.13 , 0.44, 0.91, 2.16	1.12
ExpCM + Γω (H3)	yes	yes	17370	-49620	0.09, 0.33, 0.72, 1.77	1.28
ExpCM (H1)	yes	no	2556	-50030	0.13	1.22
ExpCM (H3)	yes	no	3197	-50350	0.12	1.45
GY94 + Γω	no	yes	4719	-51106	0.00, 0.03, 0.08, 0.26	N/A
GY94	no	no	7625	-52560	0.07	N/A

empirical Data

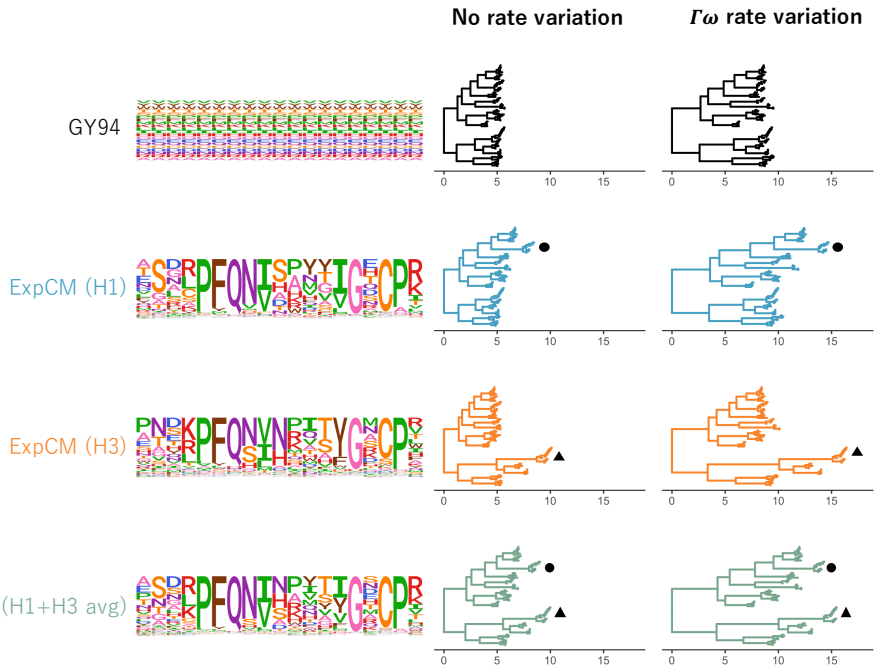


Figure 4: Trees optimized with an ExpCM defined by H1 preferences lengthen branches from the focal H1 sequence compared to GY94 models.).

Next, we want to examine the effect of substitution model on branch length estimation for actual protein coding sequences. We assume protein coding sequences have site-specific amino-acid preferences, like the simulations, but unlike the simulations, we do not know the underlying model.

To this end, we used sequences from the influenza virus surface protein hemagglutinin (HA). HA is a good model for long branch estimation because the 18 HA subtypes are all separated by long branches.

Furthermore, HA homologs are incredibly diverse and the most diverged homologs on the tree are only 42% identical on the amino-acid level. We used a tree with 87 sequences from 14 of the 18 subtypes, excluding subtypes with very few sequences.

We optimized the branch lengths on the HA tree using the models in Figure 1C. We used two different ExpCMs defined by DMS preference sets measured independently in two different HA genetic backgrounds. One DMS set was measured in an H1 background (Figure 4, black circle) and the other DMS preference set was measured in an H3 background (Figure 4, black triangle). The two focal sequences of the DMS are only 42% identical on the amino-acid level. By comparing the two ExpCMs, we can examine the effect of the ExpCM focal sequence on branch length estimation. Additionally, because the models we use represent all possible combinations of $\Gamma\omega$ rate variation and site-specific stationary state, we were able to examine the effect of each of these features while controlling for the presence of the other.

Both site-specific stationary states and $\Gamma\omega$ rate variation increase branch length estimates (Figure 4). The addition of either site-specific stationary state or $\Gamma\omega$ rate variation leads to an increase but the longest branches were estimated with the ExpCM+$\Gamma\omega$, which includes both.

However, the branch length extension for the ExpCM compared to the GY94 (or ExpCM+$\Gamma\omega$ compared to the GY94+$\Gamma\omega$) was not uniform across the entire tree. The branches with the largest difference in branch length are near the focal sequence of the deep mutational scans (Figure 4, circle for H1 and triangle for H3). This local effect of the ExpCM indicates that the stationary state described by the DMS preferences is most accurate near the focal sequences of the scan.

This result is not entirely surprising. It is expected that protein homologs as diverged as HA would be affected by epistatic interactions, when the effect of a mutation at one site is dependent on the amino-acid identity at another site. One of the strengths of deep mutational scanning is the ability to accurately measure the effect of a single amino-acid mutation. However, these measurements are in the context of a single genetic background and are therefore completely blind to the effect of epistatic interactions. Epistatic interactions would cause the DMS measurements to be different in different genetic backgrounds, resulting in "shifting" preferences across the tree, and would explain the strong local effect we see with the ExpCM. When we average the H1 and H3 preferences, or balanced sampling the two different preference sets, we see estimation of longer branches from both the H1 and H3 clades compared to the GY94 model. This is further evidence of shifts across the tree.

Competing effects of shifting preferences and long branches.

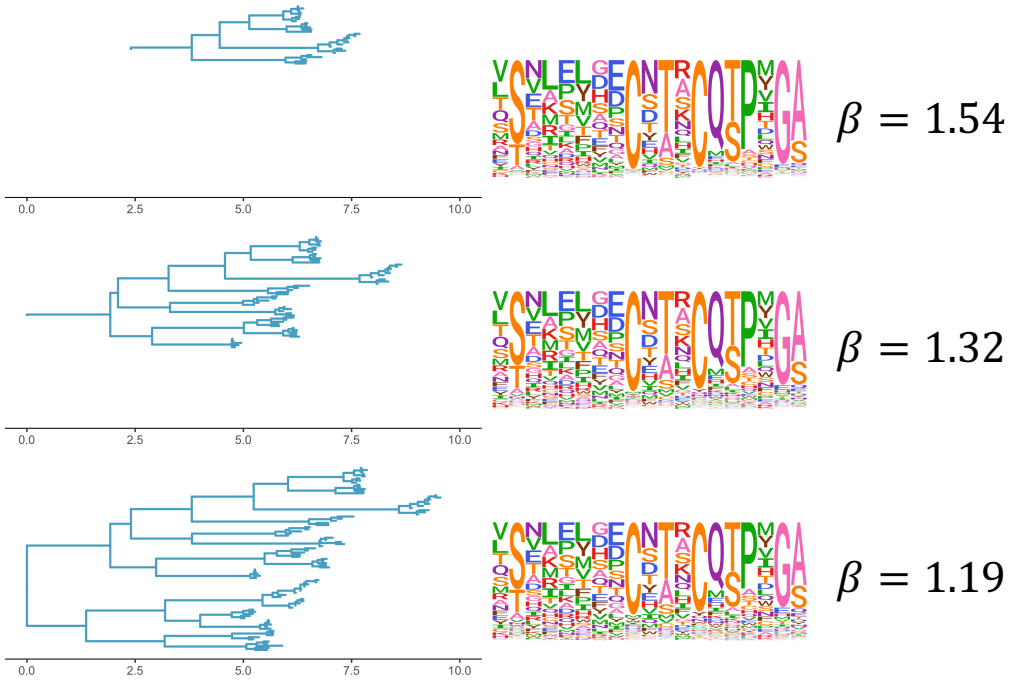


Figure 5: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree of the "high" alignment. The colors denote the alignment and the black circle denotes the focal H3 sequence. **(C)** The value of the ExpCM and ExpCM+$\Gamma\omega$ stringency parameter β decreases as the divergence from the focal H3 sequence increases. **(D)** Comparisons of branch lengths optimized by the four substitution models for the varying degrees of divergence. Black points represent branches from the focal H3 sequence and grey points represent all other branches. The branch lengths are in average number of codon substitutions per site.

We looked closer at the effect of shifting preferences on branch length estimation. We examined the behavior of the ExpCM on trees with varying levels of overall sequence divergence from the ExpCM focal sequence. For each tree, we asked "does the site-specific stationary state model estimate longer branches than the uniform stationary state model?" and "is the preference set defining the ExpCM relevant?"

The effect of site-specific stationary state on branch length estimation is seen most strongly on long branches. As the full HA tree represents "high" divergence from the DMS focal sequence, we took subsets the tree to create trees with either "intermediate" or "low" divergence from the DMS focal sequence. The majority of the branch length estimates are very similar between the GY94+$\Gamma\omega$ and the ExpCM+$\Gamma\omega$ for the "low" and "intermediate" alignments. Only when the tree has a high overall divergence from the

DMS focal sequence, and therefore the longest branches, is there a difference between the two models. This result is not unsurprising. One of the original motivations for the "mutation-selection" models was the effect the model would have on long branches specifically.

We wanted to determine the accuracy or relevance of the ExpCM on these subtrees. Due to the local branch length extension in Figure 4, we expect that the preferences are most accurate on the low divergence tree. We use the ExpCM stringency parameters as a proxy for accuracy of the preference set. The stringency parameter rescales the raw preferences from the deep mutational scan. Large stringency parameter values (> 1) indicate that selection in nature prefers the same amino acids as selection in lab but with a higher stringency. Or that the preferences measured in lab are relevant for describing the evolution of that protein. The stringency parameter for ExpCM(H1 prefs) has the highest value for the low divergence alignment and declines as the overall divergence increases. The ExpCM(H3 prefs) show the same pattern. The inverse relationship between stringency parameter and divergence supports the hypothesis that epistatic interactions degrade the ExpCM stationary state as defined by DMS preferences. [All epistatic? Could there just be a change?]

Current models which have one stationary state across the entire tree, site-specific or not, will not be able to address this tension.

`phylobayes`

Finally, we compared the branch lengths estimated between two different models with site-specific stationary states. We compared the branch lengths estimated by ExpCM+$\Gamma\omega$ (avg) with the branch lengths estimated by the mutation-selection model implemented in `phylobayes`. The `phylobayes` mutation-selection model estimates the amino-acid preferences which define the stationary state in a Bayesian framework rather than defining them from the DMS *a priori*. Branch length estimates are very similar between the ExpCM+$\Gamma\omega$ (avg) and the `phylobayes` mutation-selection model. The equivalence between the two models shows the general importance of site-specific stationary states. The local accuracy of the DMS preferences and the ExpCM can still be seen. The branch length estimates from either focal sequence (H1 and H3) are longer by the ExpCM than by the `phylobayes` mutation-selection model. This indicates that the DMS preferences are still more accurate for the local branches around the focal sequence than the global preferences estimated by `phylobayes`.

[What is the average difference in branch length between the two models?]

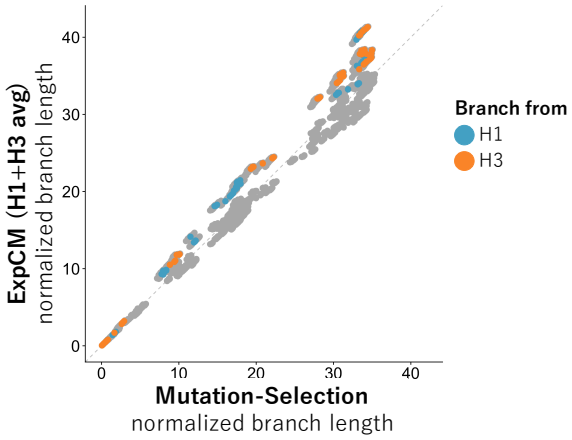


Figure 6: Comparison of ExpCM and phylobayes [y=x line too faint?]

Conclusion

1. We don't allow any of the models to vary by lineage.

Materials and Methods

Substitution models

GY94 models

ExpCMs

We recap the **Exp**erimentally Informed **C**odon **M**odel (ExpCM) (Bloom, 2014a,b, 2017; Hilton et al., 2017) to introduce nomenclature.

In an ExpCM, rate of substitution $P_{r,xy}$ of site r from codon x to y is written in mutation-selection form (Halpern and Bruno, 1998; McCandlish and Stoltzfus, 2014; Spielman and Wilke, 2015) as

$$P_{r,xy} = Q_{xy} \times F_{r,xy} \qquad \text{(Equation 1)}$$

where Q_{xy} is proportional to the rate of mutation from x to y, and $F_{r,xy}$ is proportional to the probability that this mutation fixes. The rate of mutation Q_{xy} is assumed to be uniform across sites, and takes an HKY85-like (Hasegawa et al., 1985) form:

$$Q_{xy} = \begin{cases} \phi_w & \text{if } x \text{ and } y \text{ differ by a transversion to nucleotide } w \\ \kappa\phi_w & \text{if } x \text{ and } y \text{ differ by a transition to nucleotide } w \\ 0 & \text{if } x \text{ and } y \text{ differ by } > 1 \text{ nucleotide.} \end{cases} \qquad \text{(Equation 2)}$$

The κ parameter represents the transition-transversion ratio, and the ϕ_w values give the expected frequency of nucleotide w in the absence of selection on amino-acid substitutions, and are constrained by $1 = \sum_w \phi_w$.

The deep mutational scanning data 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. The $F_{r,xy}$ terms are defined in terms of these experimentally measured amino-acid preferences as

$$F_{r,xy} = \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \\ \omega \times \frac{\ln\left[\left(\pi_{r,\mathcal{A}(y)} / \pi_{r,\mathcal{A}(x)}\right)^{\beta}\right]}{1 - \left(\pi_{r,\mathcal{A}(x)} / \pi_{r,\mathcal{A}(y)}\right)^{\beta}} & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y) \end{cases} \qquad \text{(Equation 3)}$$

where $\mathcal{A}(x)$ is the amino-acid encoded by codon x, β is the stringency parameter, and ω is the relative rate of nonsynonymous to synonymous substitutions after accounting for the amino-acid preferences. The ExpCM has six free parameters (three ϕ_w values, κ, β, and ω). The preferences $\pi_{r,a}$ are *not* free parameters since they are determined by an experiment independent of the sequence alignment being analyzed.

The ExpCM stationary state frequency $p_{r,x}$ of codon x at site r is (Bloom, 2017)

$$p_{r,x} = \frac{\left(\pi_{r,\mathcal{A}(x)}\right)^{\beta} \phi_{x_0} \phi_{x_1} \phi_{x_2}}{\sum_z \left(\pi_{r,\mathcal{A}(z)}\right)^{\beta} \phi_{z_0} \phi_{z_1} \phi_{z_2}}, \qquad \text{(Equation 4)}$$

Theoretical effect of model choice on branch length

Effect of model choice on natural sequences

ExpCM + $\Gamma\omega$ and YNGKP M5

Spielman ω_r values inferred from the ExpCM

We inferred the average nonsynonymous fixation rate from the ExpCM following Spielman and Wilke (2015) as

$$\omega_r = \frac{\sum_x \sum_{y \in N_x} p_{r,x} \times P_{r,xy}}{\sum_x \sum_{y \in N_x} p_{r,x} \times Q_{xy}} \qquad \text{(Equation 5)}$$

where $p_{r,x}$ is the stationary state of the ExpCM at site r and codon x, $P_{r,xy}$ is the substitution rate from codon x to codon y at site r, Q_{xy} is the mutation rate from codon x to codon y, and N_x is the set of codons that are nonsynonymous to codon x and differ from codon x by only one nucleotide.

Expected pairwise amino-acid identity

Do I need to talk about the branchScale scaling I used? The expected pairwise amino-acid identity at a site r over time t for a given model is

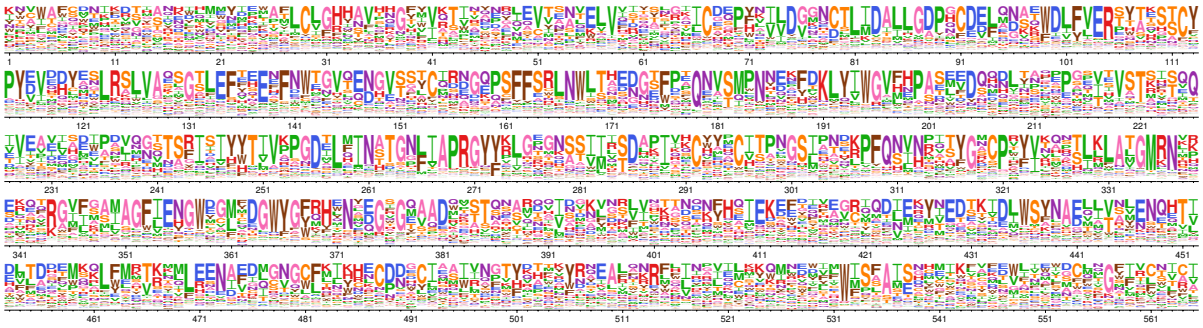
$$\sum_a \sum_{x \in a} p_{r,x} \sum_{y \in a} \left[M_r\left(t\right)\right]_{xy} \qquad \text{(Equation 6)}$$

where a is all amino acids, $p_{r,x}$ is the stationary state of the model at site r and codon x, and $\left[M_r\left(t\right)\right]_{xy}$ is the transition rate from codon x to codon y at site r given time t.

Supplemental Information

Model Parameters for the simulations

Supplementary figure 1: H1 preferences measured by Doud and Bloom (2016) rescaled with the ExpCM stringency parameter optimized in ??A ($\beta = 1.19$) [I need to change the β value when the new `phydms` results finish running.]



Supplementary figure 2: H3 preferences measured by *lee* rescaled with the ExpCM stringency parameter optimized in ??A ($\beta = 1.46$) [I need to change the β value when the new `phydms` results finish running.]

Table 2: ExpCM parameters used to simulate sequences in Fig. **??**.

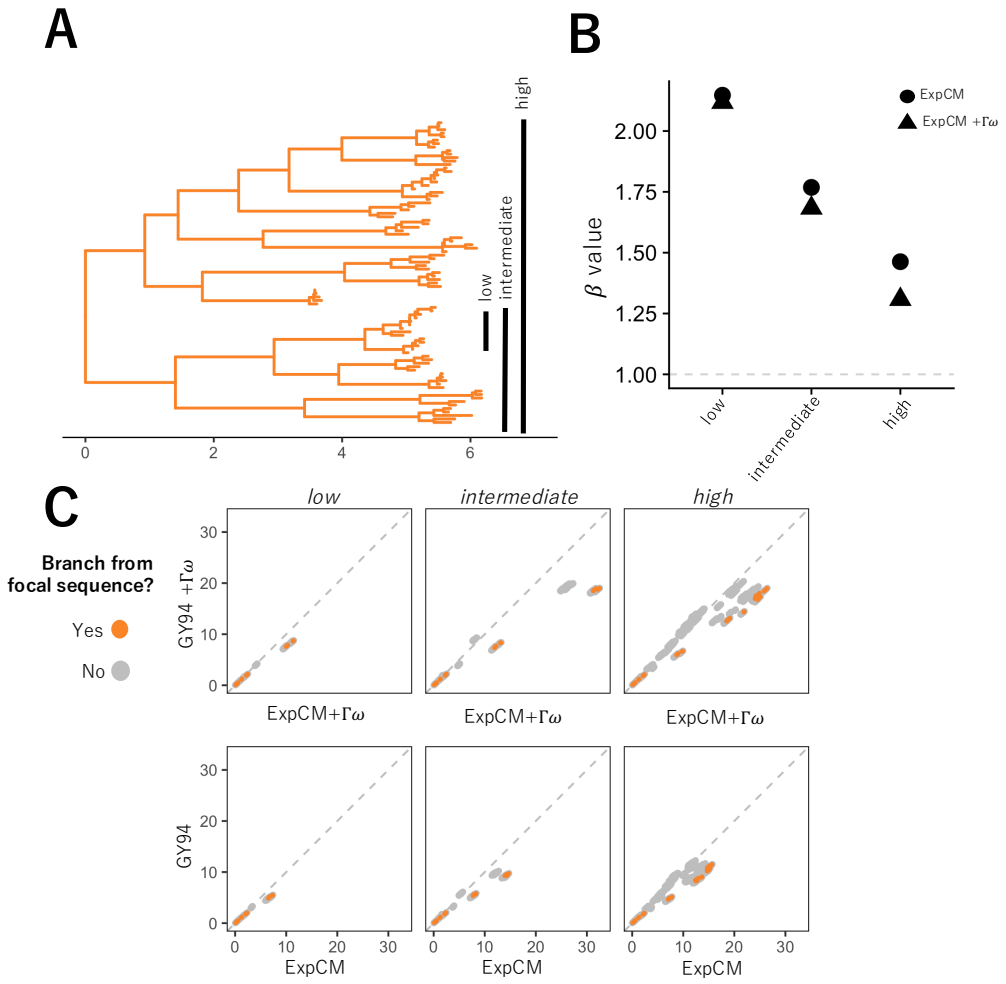
Parameter	Value
β	1.54
κ	3.60
ω	0.20
ϕ_A, ϕ_C, ϕ_G	0.38, 0.17, 0.23

Table 3: Model parameters used in Fig. Figure 2.

Model	Parameters
ExpCM	$\beta = 1.54196$
	$\kappa = 3.47184$
	$\omega = 0.219225$
YNGKP M0	$\kappa = 2.9984$
	$\omega = 0.09076$
YNGKP M5	$\kappa = 2.9984$
	$\omega = 0.09076$

[I need to change the β value when the new `phydms` results finish running.]

Supplementary figure 3: The average of the H1 preferences measured by Doud and Bloom (2016) and the H3 preferences measured by *Lee* rescaled with the ExpCM stringency parameter optimized in ??A ($\beta = 1.77$)



Supplementary figure 4: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. **(A)** An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree of the "high" alignment. The colors denote the alignment and the black circle denotes the focal H3 sequence. **(C)** The value of the ExpCM and ExpCM+$\Gamma\omega$ stringency parameter β decreases as the divergence from the focal H3 sequence increases. **(D)** Comparisons of branch lengths optimized by the four substitution models for the varying degrees of divergence. Black points represent branches from the focal H3 sequence and grey points represent all other branches. The branch lengths are in average number of codon substitutions per site.

References

Araya CL, Fowler DM. 2011. Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology*. 29:435–442.

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.

Compton AA, Malik HS, Emerman M. 2013. Host gene evolution traces the evolutionary history of ancient primate lentiviruses. *Phil. Trans. R. Soc. B*. 368:20120496.

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

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S. 2010. High-resolution mapping of protein sequence-function relationships. *Nat. Methods*. 7:741–746.

Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nature methods*. 11:801–807.

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

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.

Holmes EC. 2003. Molecular clocks and the puzzle of rna virus origins. *Journal of virology*. 77:3893–3897.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC evolutionary biology*. 7:S4.

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.

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

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

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.

Sharp P, Bailes E, Gao F, Beer B, Hirsch V, Hahn B. 2000. Origins and evolution of aids viruses: estimating the time-scale.

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

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

Wertheim JO, Worobey M. 2009. Dating the age of the siv lineages that gave rise to hiv-1 and hiv-2. *PLoS computational biology*. 5:e1000377.

Worobey M, Telfer P, Souquière S, et al. (11 co-authors). 2010. Island biogeography reveals the deep history of siv. *Science*. 329:1487–1487.

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