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.]

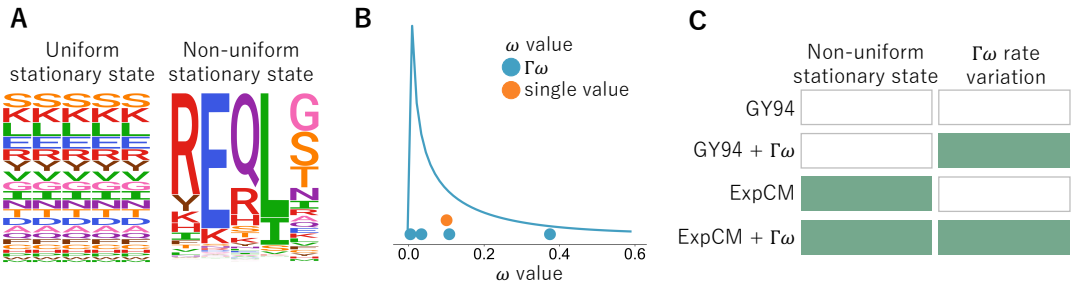


Figure 1: Different ways of modeling site-to-site variation in purifying selection. [Panels A and B need to be switched to match order in text. Then text updated too] (A) Substitution model stationary states can either be identical at every site in the protein or allow to vary from site to site. (B) 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. (C) Substitution models can both, one, or neither of these features. We used models from the GY94 and the ExpCM families.

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 (Figure 1B; Yang, 1994; Yang et al., 2000). Under 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 1A). 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 [cite Rodrigue / Lartillot papers, our papers, etc].

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 $19 \times L$ param-

eters for a protein of length L, which leads to $\sim 10^4$ 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. Here we will primarily use experimentally informed codon models (ExpCM) which define values for these parameters *a priori* from deep mutational scanning experiments so they do not need to be fit from phylogenetic data [include citations]. Therefore, for an ExpCM the number of remaining free parameters that are fit from the phylogenetic data are similar to a non-site-specific substitution model. Alternative strategies of obtaining parameters for site-specific stationary states via Bayesian or maximum-likelihood estimation are discussed in the last section of the Results.

These two strategies to account for site-to-site variation in purifying selection are not mutually exclusive. [I'm not sure if I love the rest of this paragraph. I feel like GY94 needs to be introduce more directly—perhaps in the paragraph beginning the "The most common strategy..." above. Also, it's a bit unclear what "most appropriate" means.] While differences in stationary state manifest as differences in non-synonymous rates across sites, these models are most appropriate for modeling purifying selection (Spielman and Wilke, 2015). Proteins which are under strong purifying and diversifying selection may be better modeled by a combination of the two. For this study, we used four models from the GY94 and ExpCM families, which together represent all possible combinations of these two strategies (Figure 1C).

Effect of stationary state and rate variation on branch length estimation

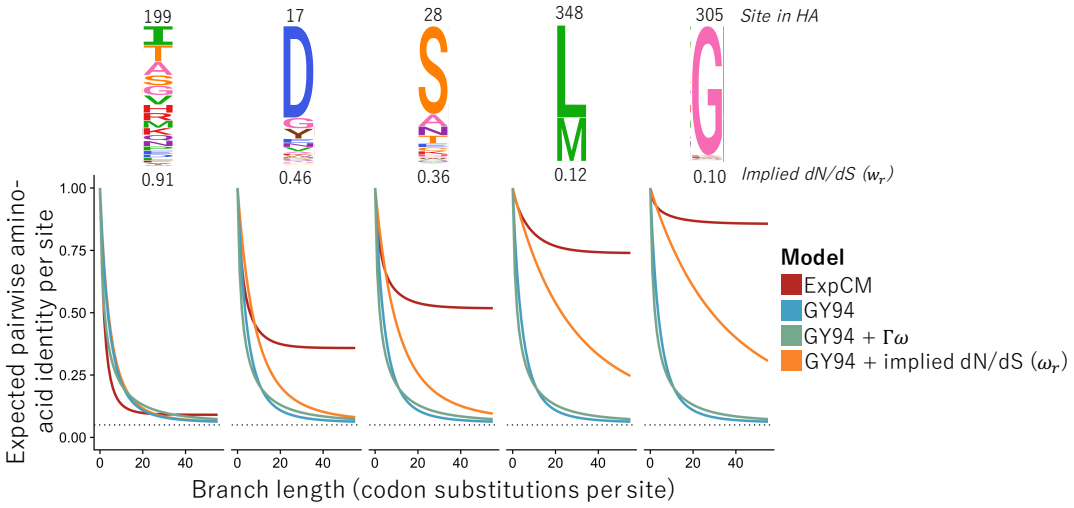


Figure 2: Effect of stationary state and rate variation on long branch estimation. The expected pairwise amino-acid identity for five sites in influenza hemagglutinin (HA) for four different substitution models. The logoplots show the HA amino-acid preferences from a deep mutational scan performed by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

Along a single branch, a substitution model transforms sequence divergence into branch length. Under the molecular clock assumption, this branch length is proportional to time. However, every substitution model has a stationary state, defining the expected sequence divergence given a very long time, at which this relationship between sequence divergence and time is lost. In Figure 2, the stationary state is represented by the long "tails" where the expected sequence divergence remains constant as time increases.

The specific parametrization scheme of a model will control features of its stationary state. Comparing GY94 to GY94+$\Gamma\omega$ will change how long it takes for the model to reach the stationary state but it does not change the expected sequence divergence at the stationary state. Furthermore, the GY94 and GY94+$\Gamma\omega$ stationary state is exactly the same at each site in the protein. Site-specific stationary states modulate both the amount of time to reach and the overall divergence level at the stationary state. These differences arise from whether or not the site is modeled as constrained or mutationally tolerant. The stationary state is important to branch length estimation because differences in expected divergence will lead to differences in branch length estimation and branch lengths cannot be estimated with any accuracy when two sequences have reached the divergence level of the stationary state.

The difference in branch length estimation between uniform and site-specific stationary state models will be most dramatic at constrained sites. Most uniform stationary state models define roughly equal frequency of each amino acid, in essence they assume that each site is mutationally tolerant. If the sequence divergence at constrained site is quite small, a site-uniform model will expect a small amount of time to have passed, a short branch, and a site-specific model will expect that a large amount of time has passed, long branch. At a mutationally tolerant site, both models will estimate roughly the same branch length for any given divergence.

This difference in branch length estimation cannot be recapitulated by more complex modeling of the ω parameter. We transformed the GY94 model into a site-specific model by inferring a unique ω value from the ExpCM for each site in the protein, following the work of Spielman and Wilke (2015). This model takes longer to reach stationary state than the GY94 model but, ultimately, has the same divergence level at stationary state. As this model shows, no matter how "well" a model accounts for site-to-site rate variation, it will underestimate long branches with a uniform stationary state model.

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

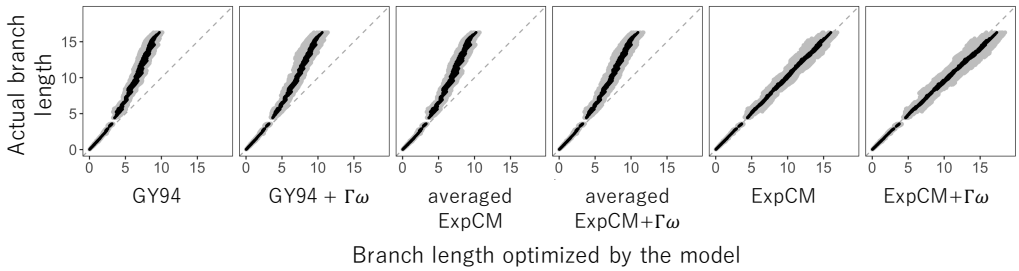


Figure 3: Model performance under simulated, site-specific data. Alignments were simulated under an ExpCM along an HA tree and the branches were re-optimized by a model from the ExpCM or YNGKP family. The averaged ExpCMs amino-acid frequencies defined by the average preference of that amino acid across 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.

To test the effect of substitution models on branch length estimation given sequences with site-specific amino-acid frequencies we simulated sequences under an ExpCM and re-inferred the branch lengths using the different substitution models described in Figure 1C. As expected, the ExpCM and ExpCM+$\Gamma\omega$ estimated the branch lengths of the simulated sequences accurately. The variance in the estimation of an given branch increases between the simulations increased as branch length increased but this error is not biased towards over- or underestimation. The models with a uniform stationary state consistently underestimate the long branches. This underestimation is seen with both the GY94 and the GY94+$\Gamma\omega$, indicating that accounting for rate variation via the rate parameter cannot prevent branch length underestimation at these long branches. The ExpCM can also be transformed into a uniform stationary state model by setting the amino-acid at every site to the average value for that amino acid across all sites. In contrast to the GY94 models, this average control does not have preference for amino acid and does extract some information from the experiments, but like the GY94 models the stationary state is uniform across sites. These models performed similarly to the GY94 and GY94+$\Gamma\omega$ and underestimated long branches.

These simulations show that when the sequences have site-specific amino-acid frequencies, models with uniform stationary states cannot accurately estimate branch lengths and, will in fact, underestimate long branches.

Table 1: Model comparison for Fig. Figure 4.

Model	ΔAIC	Log Likelihood	Parameters
ExpCM + $\Gamma\omega$ (H1+H3 average)	0.00	-48751.16	ω = (0.19, 0.50, 0.90, 1.86), $\beta = 1.70$, $\kappa = 3.73$
ExpCM (H1+H3 average)	949.70	-49227.01	$\beta = 1.78$, $\kappa = 3.40$, $\omega = 0.15$
ExpCM + $\Gamma\omega$ (H1)	1306.44	-49404.38	ω = (0.13 , 0.44, 0.91, 2.16), $\beta = 1.12$, $\kappa = 3.64$
ExpCM + $\Gamma\omega$ (H3)	1737.40	-49619.86	ω = (0.09, 0.33, 0.72, 1.77), $\beta = 1.28$, $\kappa = 3.71$
ExpCM (H1)	2555.58	-50029.95	$\beta = 1.22$, $\kappa = 3.24$, $\omega = 0.13$
ExpCM (H3)	3196.50	-50350.41	$\beta = 1.45$, $\kappa = 3.38$, $\omega = 0.12$
GY94 + $\Gamma\omega$	4719.34	-51105.83	ω = (0.00, 0.03, 0.08, 0.26), $\kappa = 3.18$
GY94	7624.64	-52559.48	$\kappa = 2.89$, $\omega = 0.07$

empirical Data

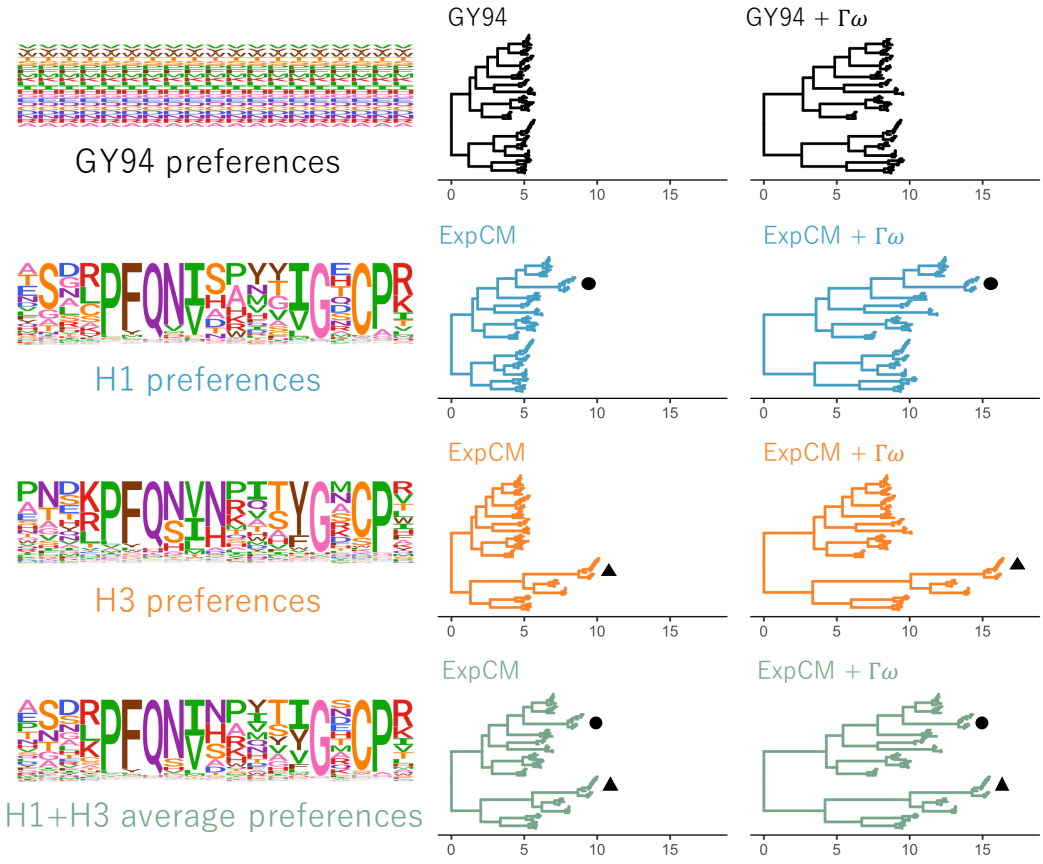


Figure 4: Trees optimized with an ExpCM defined by H1 preferences lengthen branches from the focal H1 sequence compared to GY94 models.).

[Section Outline:]

- We tested the effect substitution model choice on branch length estimation using influenza HA sequences.
  - These sequences presumably evolve with site-specific frequencies but unlike the simulations we do not *know* the model
  - These sequences are from fairly diverged proteins. Some sequences share only 40% identity on the amino acid level
  - The HA tree has closely related subtypes with long branches between subtypes. This is good because we are looking for the effect on long branches.
- Both $\Gamma\omega$ and non-uniform stationary states result in an extension of branch lengths.
  - The addition of $\Gamma\omega$ extends branches independent of preferences.
  - The addition of preferences extends branches *from the focal sequence* independent of $\Gamma\omega$.
- The extension of branch length from the focal sequence shows that the ExpCM does not have equal "relevance" across the whole tree.
  - This result is not entirely surprising. DMS measure the effects of single amino-acid changes in the context of a single genetic background.
  - It is expected that there would be epistatic changes across such a highly diverged phylogenetic tree. "Shifting preferences".
  - "Sampling" from both preference sets via preference averaging extends the branch length from both focal sequences.

Competing effects of shifting preferences and long branches.

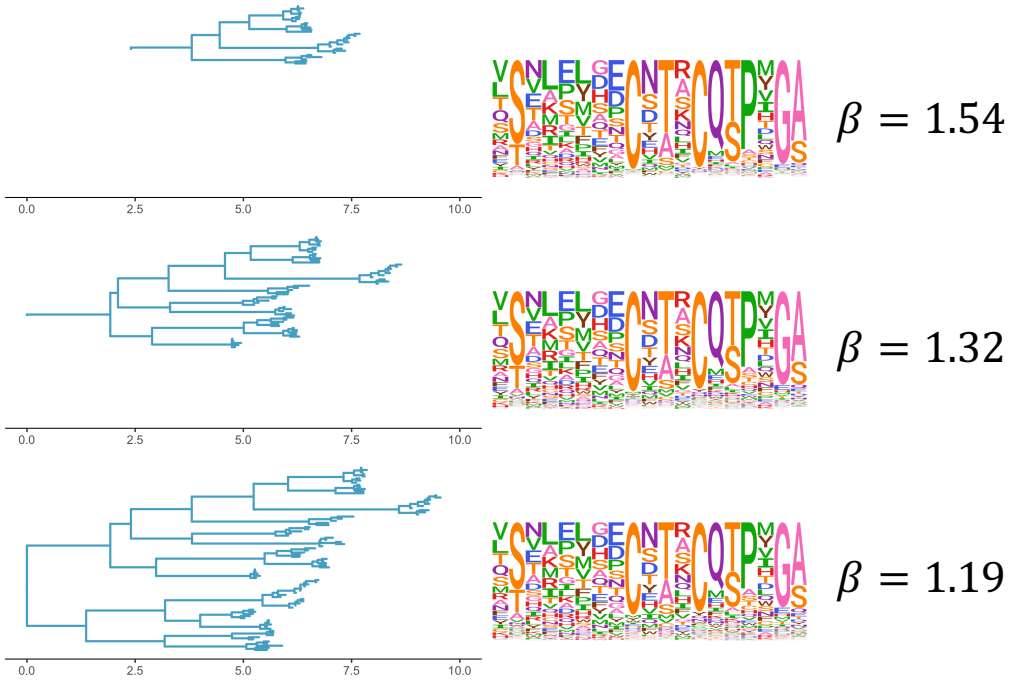


Figure 5: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree of 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.

[Section Outline:]

- We investigated the competing effects of "shifting preferences" and long branches by comparing ExpCMs with trees of increasing overall divergence from the focal sequence.
- We use the fit stringency parameter β as a measure of relevance.
  - The stringency parameter rescales the preferences. A stringency parameter with a value greater than one indicates that selection in nature prefers the same amino acids as selection in lab but with greater strength.

  - There is an inverse relationship between the stringency parameter and overall sequence divergence. As the overall divergence of the tree increases, the stringency parameter decreases.

- [hypothetical data] There is also a positive relationship between the stringency parameter and branch lengths
  - We optimized the same tree with ExpCMs with different stringency parameters.
  - The ExpCMs with higher stringency values estimate longer branches than the ExpCMs with lower stringency values.

`phylobayes`

[Section Outline]

- `phylobayes` implements the mutation-selection model in a bayesian framework as a way to avoid overfitting.
- `phylobayes` and ExpCM (H1+H3 average) estimate branches of roughly equal length
  - the ExpCM estimate longer branches from the focal sequences
  - `phylobayes` estimates longer branches from the non-focal sequences

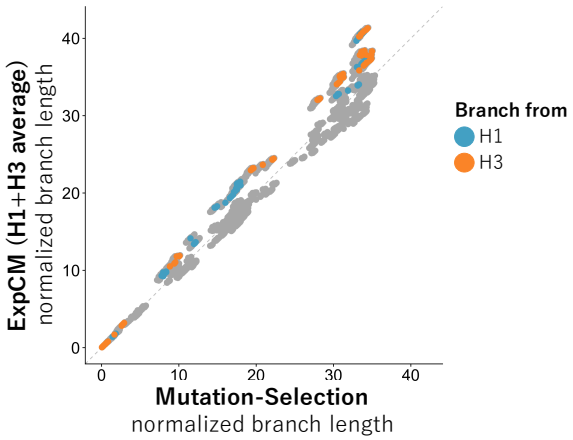


Figure 6: Comparison of ExpCM and phylobayes [y=x line too faint?]

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

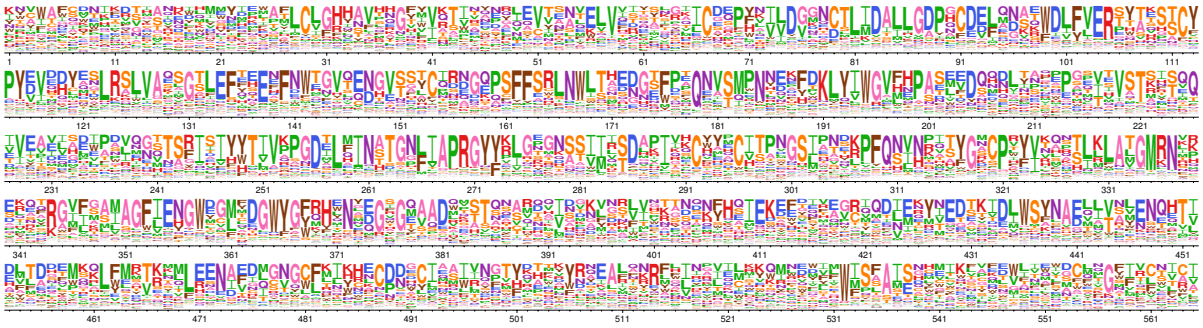
$$\sum_a \sum_{x \in a} p_{r,x} \sum_{y \in a} \left[M_r\left(t\right)\right]_{xy} \qquad \text{(Equation 6)}$$

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.]



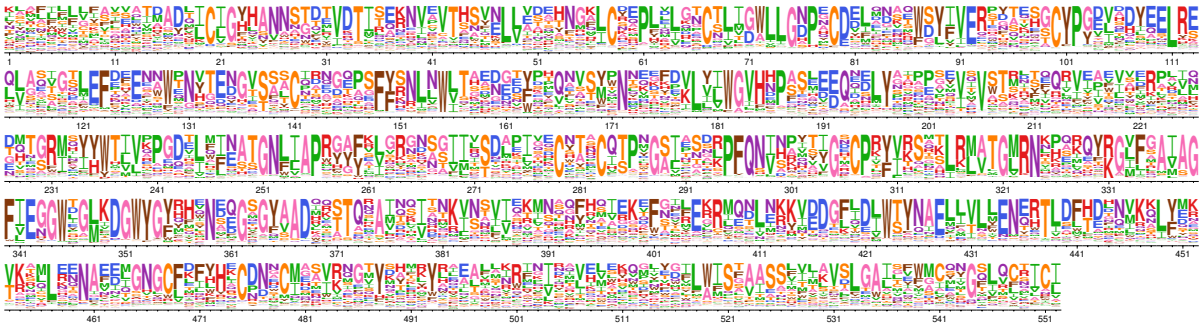
Supplementary figure 2: H3 preferences measured by *lee* rescaled with the ExpCM stringency parameter optimized in ??A ($\beta = 1.46$) [I need to change the β value when the new `phydms` results finish running.]

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

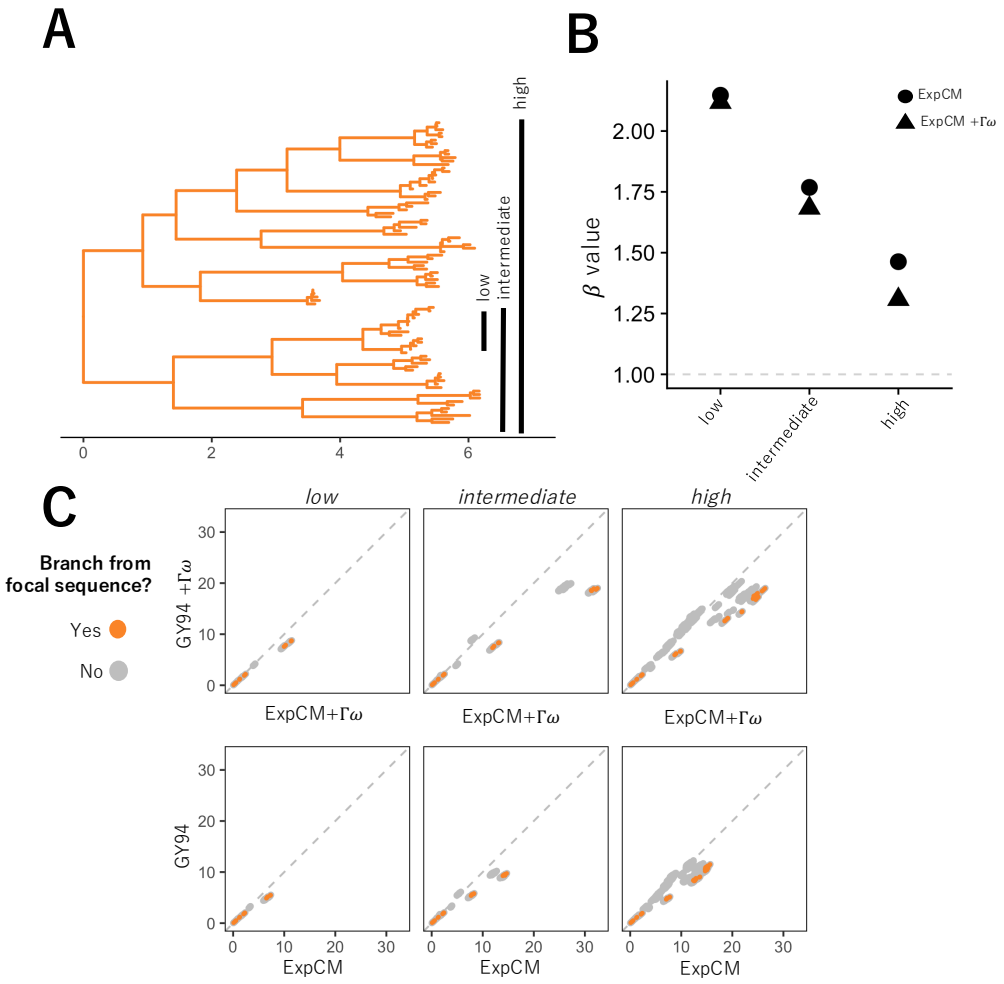
Table 3: Model parameters used in Fig. Figure 2.

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

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

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

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.