Experimentally Informed Site-Specific Substitution Models Substantially Deepen Divergence Estimates

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

$\leq$ 250 words, currently 159

Molecular dating techniques have been used to estimate the divergence date of many viruses. However, these estimates are consistently substantially younger than estimates from methods which are independent of branch length estimation. This discrepancy is caused, in part, by inadequate modeling of purifying selection leading to branch length underestimation. Here, we show that substitution models informed by empirical measurements of mutational constraint better than traditional models and extend branch lengths. We used models informed by deep mutational scanning experiments performed in two, highly diverged influenza virus hemagglutinin homologs to optimize the branch lengths of a phylogenetic tree. For each experimentally informed model, we observed extension in branch length from the experiment's focal sequence. This extension in branch length due to explicit modeling of site-specific purifying selection is observed in the presence and absence of standard methods for modeling site-to-site variation. Overall, this study underscores the importance of modeling purifying selection when estimating branch lengths and, by extension, divergence dates.

Introduction

Introduction outline 1. Estimating the divergence date of viruses is important and common. 2. There are numerous examples of molecular dating techniques contradicting other dating methods for viruses. 3. Why purifying selection could cause branch length underestimation. 4. Discussion of the attempts/methods made thus far to model site-to-site rate variation. 5. We use empirical measurements to model site-specific purifying selection. 6. In this study we compare branch lengths of trees optimized by GY94 and ExpCM models and see an extension in branch lengths with ExCM and with $\Gamma\omega$.

Estimating the age of viruses is important to understanding their evolutionary history, including co-evolution with their host species. Molecular dating techniques are commonly used to estimate the dates of viruses. However, these dates are consistently younger than the dates obtained by other methods which are independent of phylogenetic trees, often by orders of magnitude. *example HIV/SIV, foamy virus, or measles*

Failing to account for site-specific purifying selection results in an underestimation of branch lengths. It has long been observed that protein-coding genes have a non-uniform distribution of amino-acid frequencies across the sites in the protein. Multiple sequence alignments of homologs show some sites which are conserved and appeared to be mutationally constrained and other sites which are variable and appear to be mutationally tolerant. This site-specific purifying selection dictates the expectation of change at a given site which is translated into branch length. *The strength of purifying selection at a given site gives the expectation of change which directly relates to branch length. Specifically, failing to account for constraint will lead to an underestimation in branch length because your expectation of change will be too high.*

Different substitution models have been developed to address the effect of site-specific purifying selection. Most commonly, a rate parameter is described by some statistical distribution fit across the entire protein, such as the Γ-distributed ω in the YNGKP M5 (Yang et al., 2000). Other approaches have tried to explicitly model the site-specific amino-acid frequencies, such as mutation-selection models (Halpern and Bruno, 1998). While perhaps the most biologically-relevant, mutation-selection models are heavily parametrized and run the risk of overfitting in most practical applications. However, we have previously shown that we can model this site-specific purifying selection using empirical measurements of mutational constraint from a high-throughput assay called deep mutational scanning. We hypothesize that using these experimentally informed models will estimate longer branches than traditional models.

Here, we address this hypothesis by comparing the branch lengths on phylogenetic trees describing influenza virus hemagglutinin optimized by different substitution models. By comparing the trees optimized with and without the experimental descriptions of purifying selection, we show that failing to account for site-specific constraints results in shorter estimations of branch length. We used two different experimentally informed models, defined by experimental measurements from one of two

diverged homologs. For each model, the branchs from the experimental focal sequence are extended compared to traditional models. This branch length extension is seen even in the presence of model parameters which are traditionally used to describe site-to-site rate variation. These results underscore the importance of modeling purifying selection when estimating divergence dates and suggest that experimental measurement may be used to account for this selection while avoiding overparameterization.

Results and Discussion

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

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

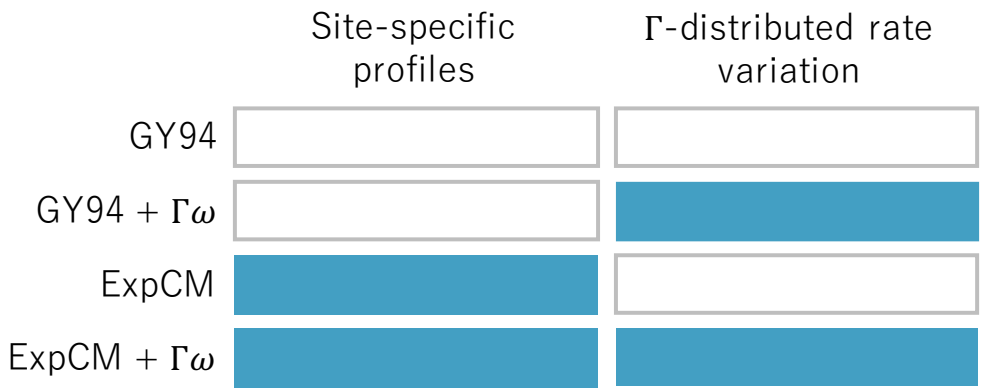

Figure 1: Comparison of substitution model features. Site-specific amino-acid profiles and Γ-distributed rate variation are both substitution model features which have been shown or theorized to lengthen branches. The models YNGKP M0, YNGKP M5, ExpCM, and ExpCM+Γω represent all possible combinations of these two features. Blue indicates presence and white indicates absence of a feature.

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

Conclusion

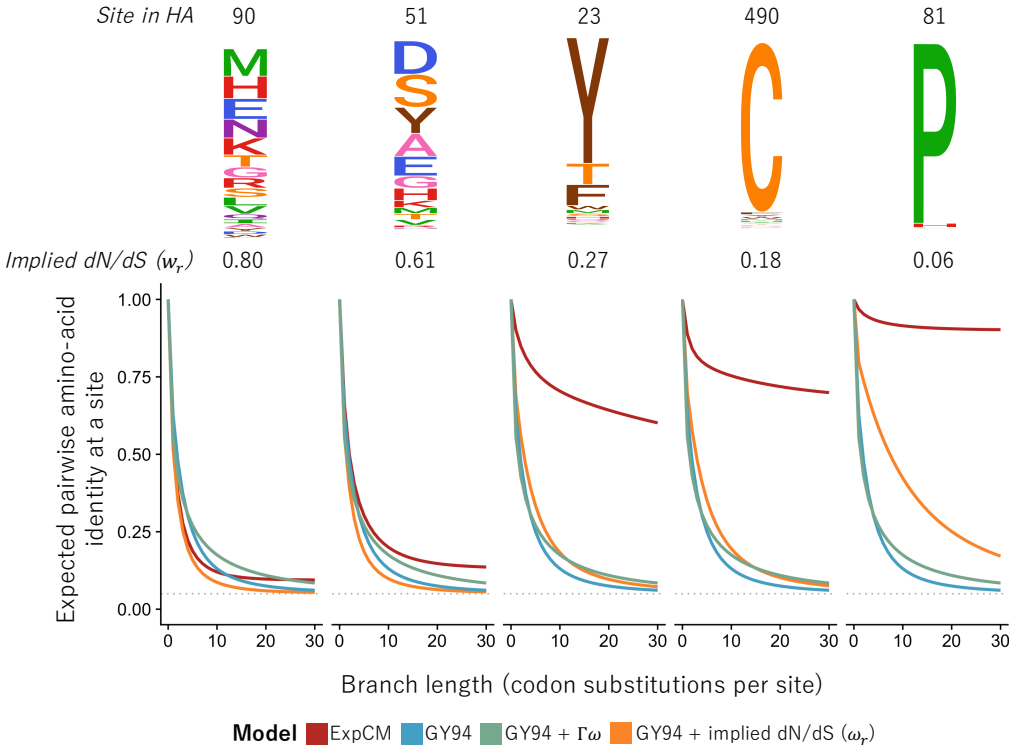

Figure 2: ExpCMs infer longer branches than YNGKP models for sites with mutational constraint. The expected pairwise identity trajectories were calculated using Equation 6 and models described in Table 2. The trajectories of the YNKGP M0 (blue) and YNGKP M5 (green) do not vary from panel to panel because neither model is site-specific. The deviation in trajectory of the ExpCM (red) from the YNGKP M0 (blue) increases from left to right as the mutational constraint of the amino-acid profile increases (logoplots, above). The deviation in trajectory of the YNGKP model with a site-specific ω value inferred from the ExpCM (yellow, Equation 5) is also positively correlated with the constraint of the site-specific amino-acid profiles but the effect size is smaller.

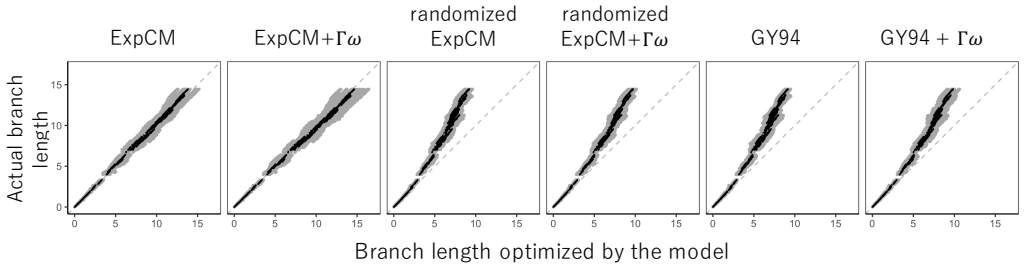

Figure 3: Branch lengths simulated under an ExpCM are underestimated by YNGKP models and long branches are disproportionally affected. Alignments were simulated under an ExpCM (**??**) along an HA tree and the branches were re-optimized by a model from the ExpCM or YNGKP family. The randomized ExpCMs have amino-acid profiles shuffled among the sites These randomized models are still site-specific but the relationship between the site and the experimental data is broken. Grey points represent the length of one branch and the black points are the mean branch lengths over eight 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.

Materials and Methods

Experimentally informed codon model (ExpCM)

YNGKP M0

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

The expected pairwise amino-acid identity at a site r over time t for a given model is

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

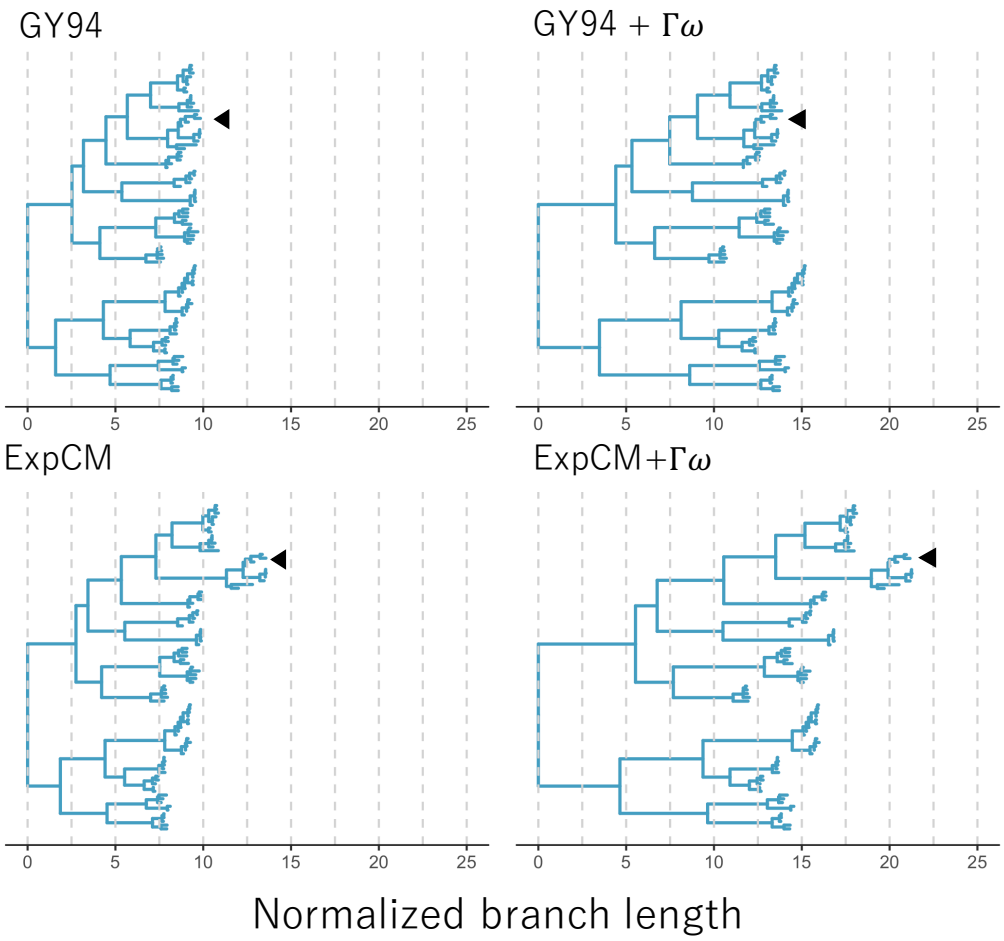

Figure 4: Trees optimized with an ExpCM defined by H1 preferences lengthen branches from the focal H1 sequence compared to YNGKP models. The branch lengths of a base topology inferred using the GTR-CAT model were optimized by **(A)** an ExpCM defined by H1 preferences, **(B)** an ExpCM+$\Gamma\omega$ defined by H1 preferences, **(C)** YNKGP M0, and **(D)** YNGKP M5. The branch lengths are normalized to the distance between A/South Carolina/1/1918 and A/Solomon Islands/3/2006 and colored to indicate the distance from the H1 focal sequence (black triangle).

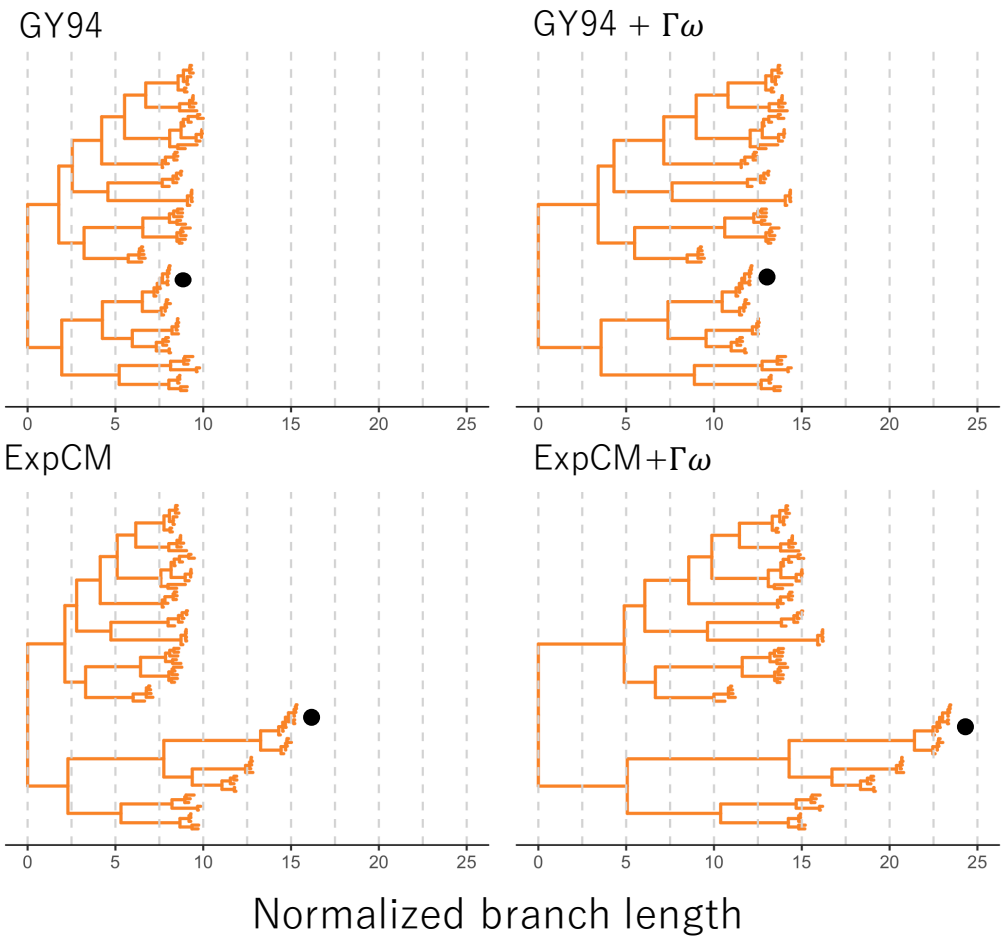

Figure 5: Trees optimized with an ExpCM defined by H3 preferences lengthen branches from the focal H3 sequence compared to YNGKP models. The branch lengths of a base topology inferred using the GTR-CAT model were optimized by **(A)** an ExpCM defined by H3 preferences, **(B)** an ExpCM+$\Gamma\omega$ defined by H3 preferences, **(C)** YNKGP M0, and **(D)** YNGKP M5. The branch lengths are normalized to the distance between A/South Carolina/1/1918 and A/Solomon Islands/3/2006 and colored to indicate the distance from the H3 focal sequence (black circle).

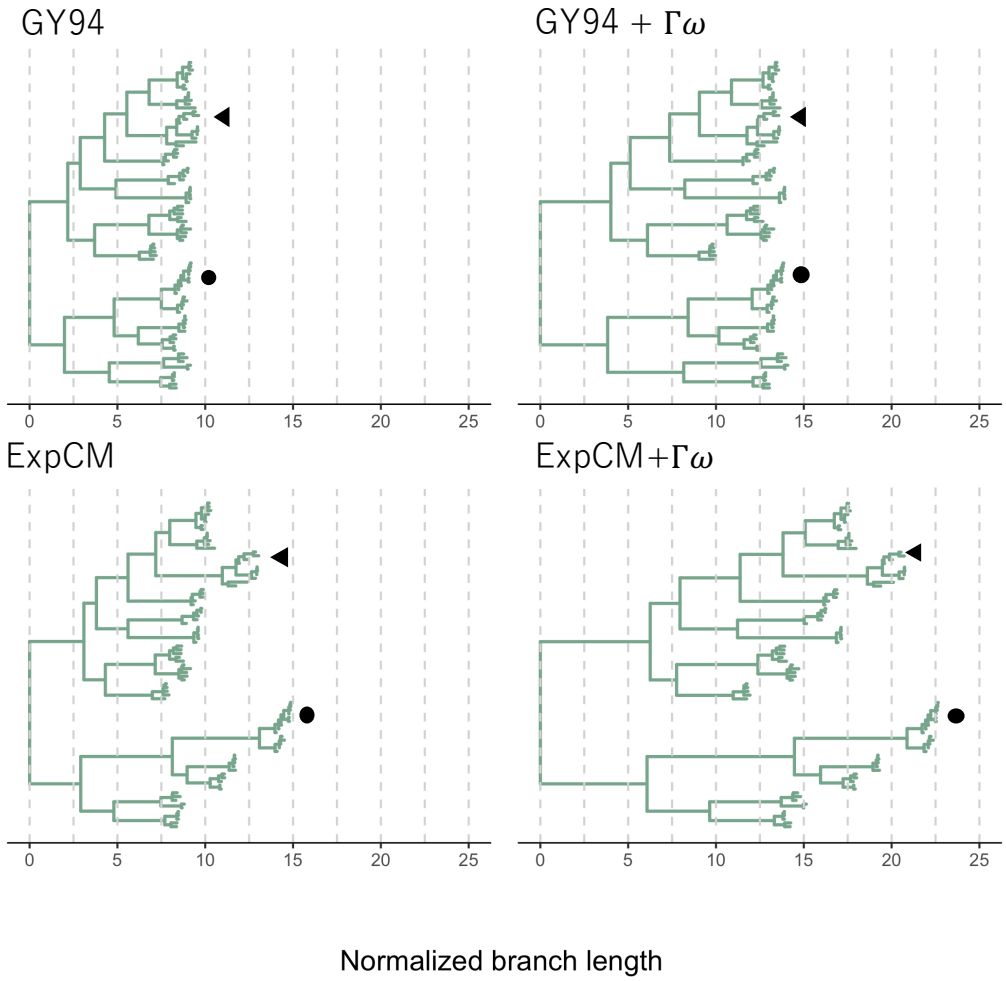

Figure 6: Trees optimized with an ExpCM defined by the average of H1 and H3 preferences lengthen branches from both the focal H3 sequence and the focal H1 sequence compared to YNGKP models. The branch lengths of a base topology inferred using the GTR-CAT model were optimized by **(A)** an ExpCM defined by the average preferences, **(B)** an ExpCM+Γω defined by the average preferences, **(C)** YNKGP M0, and **(D)** YNGKP M5. The branch lengths are normalized to the distance between A/South Carolina/1/1918 and A/Solomon Islands/3/2006. The black triangle indicates the H1 focal sequence and the black circle indicates the focal sequence.

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

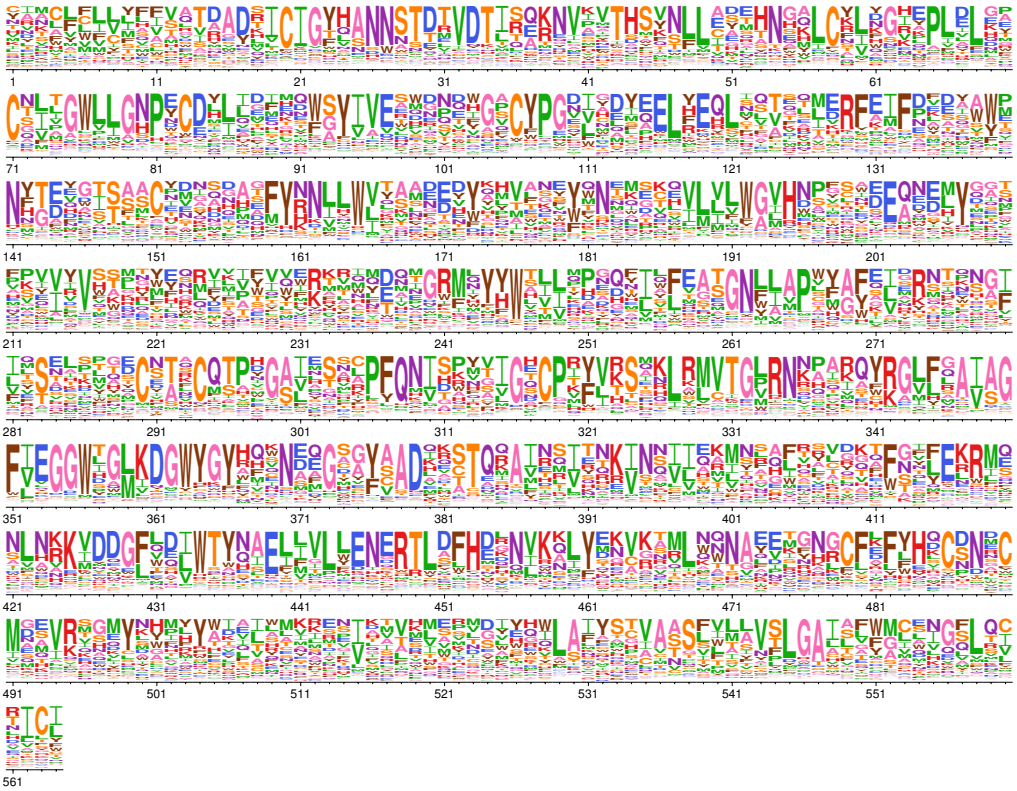

Figure 7: H1 preferences measured by Doud and Bloom (2016) rescaled with the ExpCM stringency parameter optimized in Figure 4A ($\beta = 1.21$)

Table 1: 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

Supplemental Information

Model Parameters for the simulations

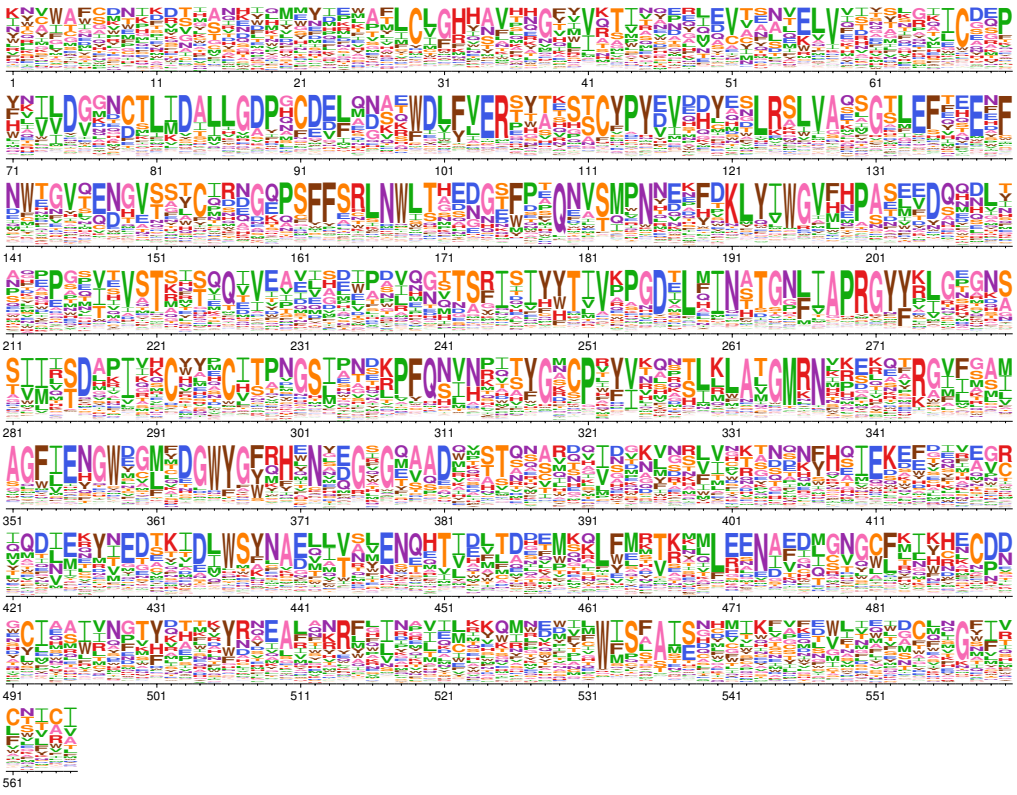

Figure 8: H3 preferences measured by *lee* rescaled with the ExpCM stringency parameter optimized in Figure 5A ($\beta = 1.46$)

Figure 9: 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 Figure 6A ($\beta = 1.82$)

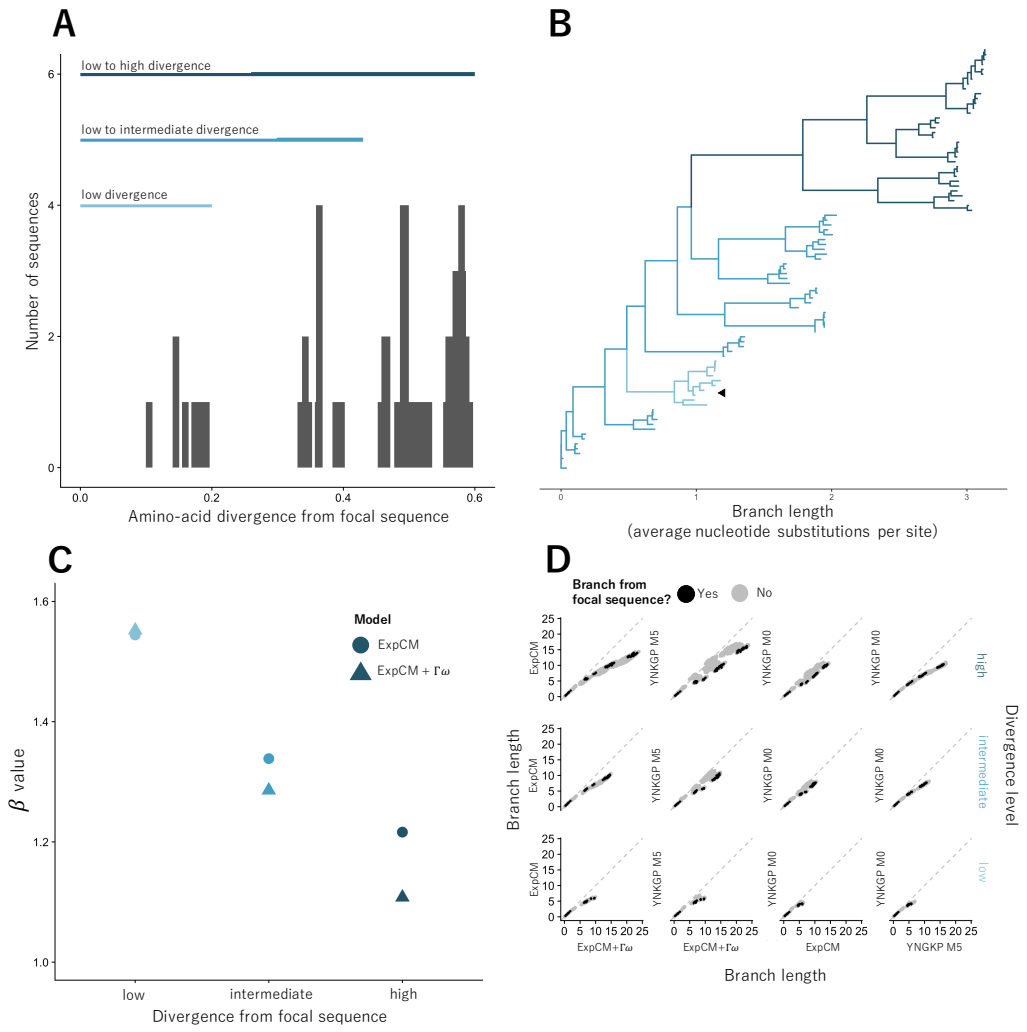

Figure 10: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H1 sequence, referred to as "low", "intermediate", and "high". **(B)** A phylogenetic tree of the "high" alignment was constructed using the GTR-CAT model. 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 H1 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.

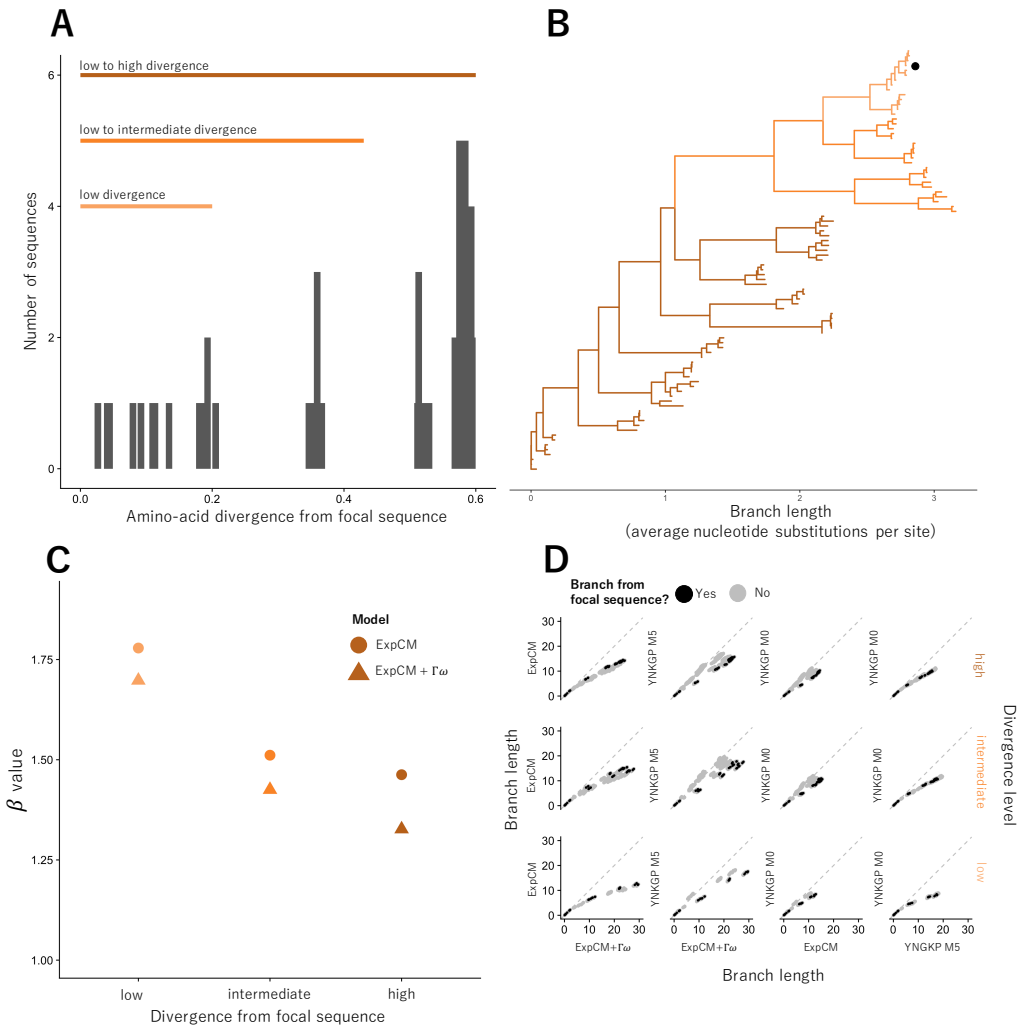

Figure 11: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree 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.

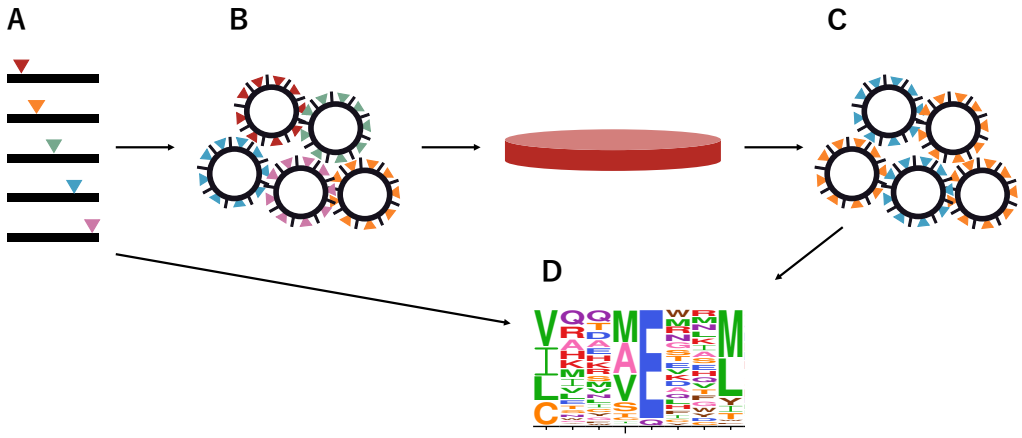

Figure 12: Schematic of deep mutational scanning. **(A)** All single codon mutations are introduced into the wildtype HA gene. **(B)** Each virus in the mutant virus library contains one HA variant. **(C)** The mutant virus library is passaged in cell culture to select for functional variants. **(D)** Deep sequencing quantifies the frequency of each variant before and after selection. The preference for each amino acid at each site (as quantified by the deep sequencing) is represented as a logoplot.

Table 2: Model parameters used in Fig. **??**.

Model	Parameters
ExpCM	$\beta = 2.0$, $\kappa = 1.0$, $\omega = 1.0$, $\phi_A = \phi_C = \phi_T = 0.25$, $\pi_{r,A(X)}$:**?**
YNGKP M0	$\kappa = 1.0$, $\omega = 1.0$, $\phi_{rw} = 0.25$
YNGKP M5	$\kappa = 1.0$, $\alpha_\omega = 0.36$, $\beta_\omega = 1.9$, $\phi_{rw} = 0.25$

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.

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

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. *bioRxiv*. p. 121830.

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

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

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