Experimentally Informed Site-Specific Substitution Models Substantially Deepen Viral 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

Molecular dating techniques is often used to estimate the divergence time of many viruses. However, these estimates are substantially younger than estimates from methods which do not rely on viral phylogenies. This discrepancy is thought to be 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 model 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. from JDB: also shows a way to actually do this in addition to showing importance.

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 (**???**) but estimated to be several million years old based on the host tree or endogenous retroviral elements (**?**) (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 (**?**). 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 incorportate 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) (**?**). 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 (**?**). 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 (**?**) or maximum likelihood framework (**?**).

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) (**??**) from the mutation-selection family using measurements from deep mutational scanning (**?**), 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

Substitution models

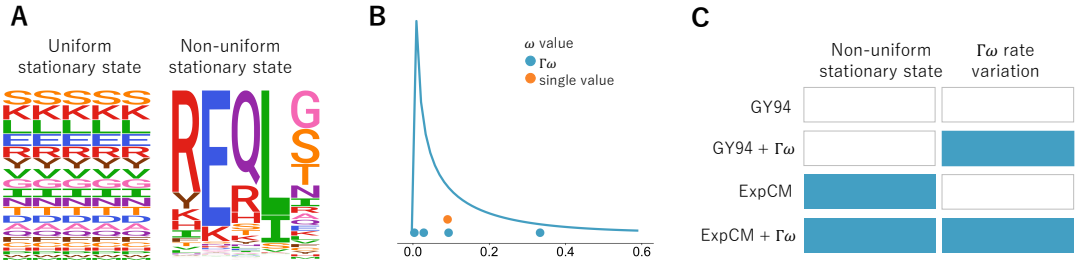


Figure 1: Comparison of substitution model features. Site-specific amino-acid profiles and Γ-distributed rate variation are both substitution model features which have been shown or theorized to lengthen branches. The models YNGKP M0, YNGKP M5, ExpCM, and ExpCM+$\Gamma\omega$ represent all possible combinations of these two features. Blue indicates presence and white indicates absence of a feature. from Table 2

Section Outline:

- There are different strategies to account for purifying selection and site-to-site rate variation.
  - $\Gamma\omega$ rate variation
    - single parameter controlling relative rate of non-synonymous to synonymous substitutions can vary across sites following some statistical distribution (usually Γ)
    - very common
    - practical implementation - discretized distribution with the mean of K bins.
  - stationary state
    - Explicitly model each amino-acid frequency at each site in the protein
    - a more mechanistic description, no arbitrary statistical distribution
    - practical limitation - how do you estimate these parameters without overfitting the data? We use experimental measurements to define the parameters *a priori*
- We can compare the effect of a model feature ($\Gamma\omega$ rate variation or non-uniform stationary state) while controlling the presence/absence of the other feature using models from the ExpCM and GY94 families.

Effect of stationary state and rate variation on branch length estimation

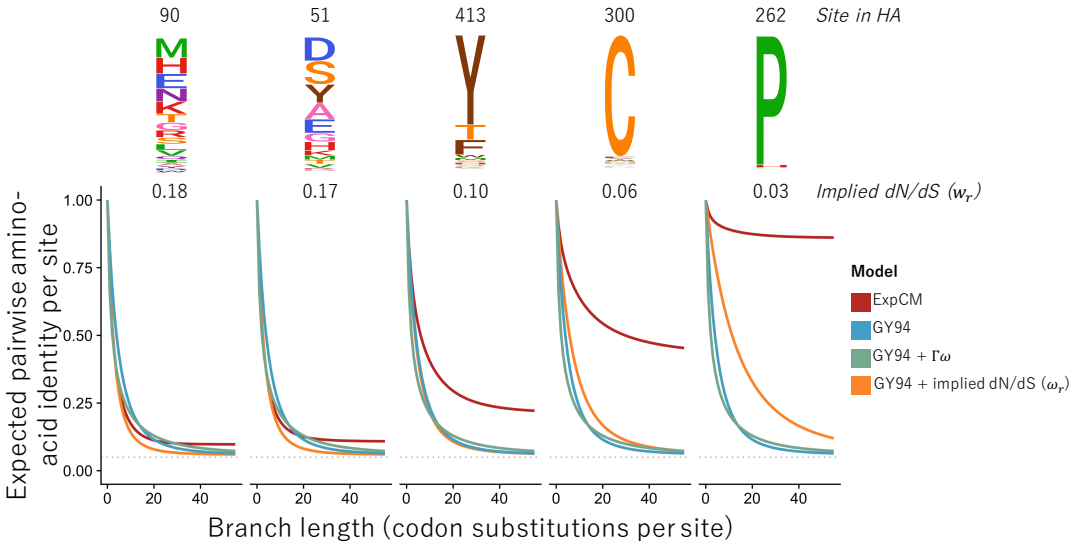


Figure 2: Effect of stationary state and rate variation on long branch estimation. The expected pairwise identity trajectories were calculated using Equation 6 and models described in Table 2. The trajectories of the YNKGP M0 (blue) and YNGKP M5 (green) do not vary from panel to panel because neither 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.

Section Outline:

- Substitution models are only useful for inferring branch lengths when they can accurately transform sequence divergence to time (branch length)
- There is a saturation point in every substitution model where the temporal signal is lost (long tails in figure).
  - In the long tails the expected sequence divergence remains constant as time increases.
  - The lost of temporal signal will affect long branches specifically.
- Different model features (Figure 1) can affect *how long* it takes to reach the saturation point and the *sequence divergence* at the saturation point
  - Adding $\Gamma\omega$ rate variation affects the "path" to the saturation point but affects the sequence divergence at the saturation point minimally. Importantly, the saturation point is identically

for every site in the protein. These $\Gamma\omega$ parameters might not the best to show the effect of $\Gamma\omega$ on long branch estimation.

  - Adding a non-uniform stationary state creates a unique saturation point for each site in the protein. Sites which can tolerate many mutations have saturation points which are quite similar to the GY94 saturation point. However, constrained sites show a dramatic difference.
  - This difference cannot be recapitulated by more complex modeling of the ω parameter. We can transform the GY94 model into a completely site-specific value by inferring a unique ω value from the ExpCM for each site in the protein. This model shows a substantial difference but does not recover the full effect of the ExpCM.

- No matter how "well" you model site-to-site rate variation, you will underestimate *very* long branches with a uniform stationary state.

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

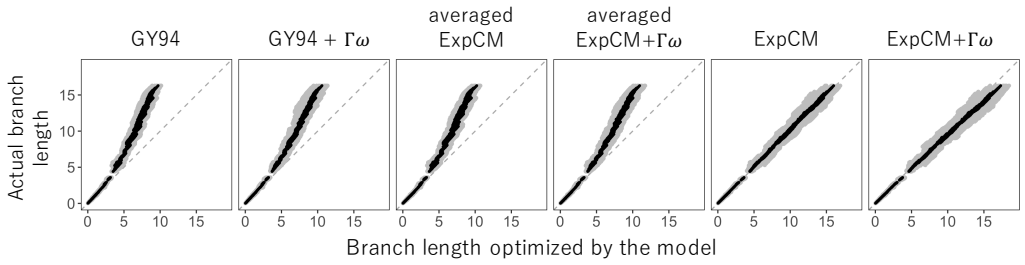


Figure 3: Model performance under simulated, site-specific data. Fonts are small, and tick labels definitely too small. I would show averaged. Alignments were simulated under an ExpCM (**??**) along an HA tree and the branches were re-optimized by a model from the ExpCM or YNGKP family. The randomized 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.

Section Outline:

- To test the effect of substitution model on data with site-specific amino acid frequencies, we simulated sequences under an ExpCM and re-inferred the branch lengths using different substitution models.
- Uniform stationary state models underestimate long branches

  - GY94 and GY94 + $\Gamma\omega$ both underestimate long branches
  - The *averaged* ExpCMs also underestimate long branches. These control models have uniform stationary states but still derive some information from the experiments.
  - Non-uniform stationary state models accurately estimate long branches

- If sequences have site-specific amino acid frequencies then non-uniform stationary state models, even with $\Gamma\omega$, will underestimate long branches.

empirical Data

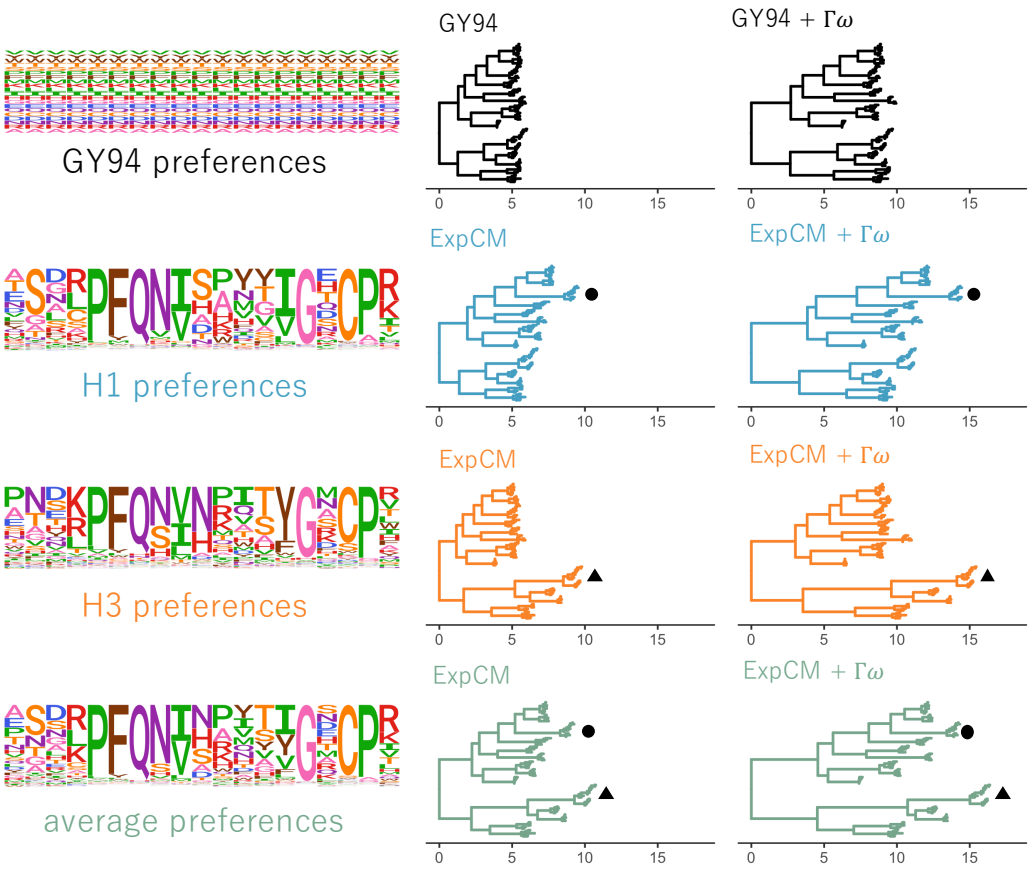


Figure 4: Trees optimized with an ExpCM defined by H1 preferences lengthen branches from the focal H1 sequence compared to YNGKP models. Keep branch lengths in substitutions per site. Make a panel A show a prefs snippet, B H1 trees, C H3 trees, D H1+H3. Maybe increase very small font sizes a bit. The 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).

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 don't *know* the model

  - These sequences are from fairly diverged proteins. Some sequences share only 40% identity on the amino acid level
  - HA has small, closely related subtypes with long branches between subtypes and groups.

- 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 ExpCM does not have a uniform effect across the whole tree.
  - This is not too surprising. DMS makes accurate measurements in a given genetic background.
  - You would expect epistatic effects over such a large divergence distance. "Shifting preferences"

Competing effects of shifting preferences and long branches.

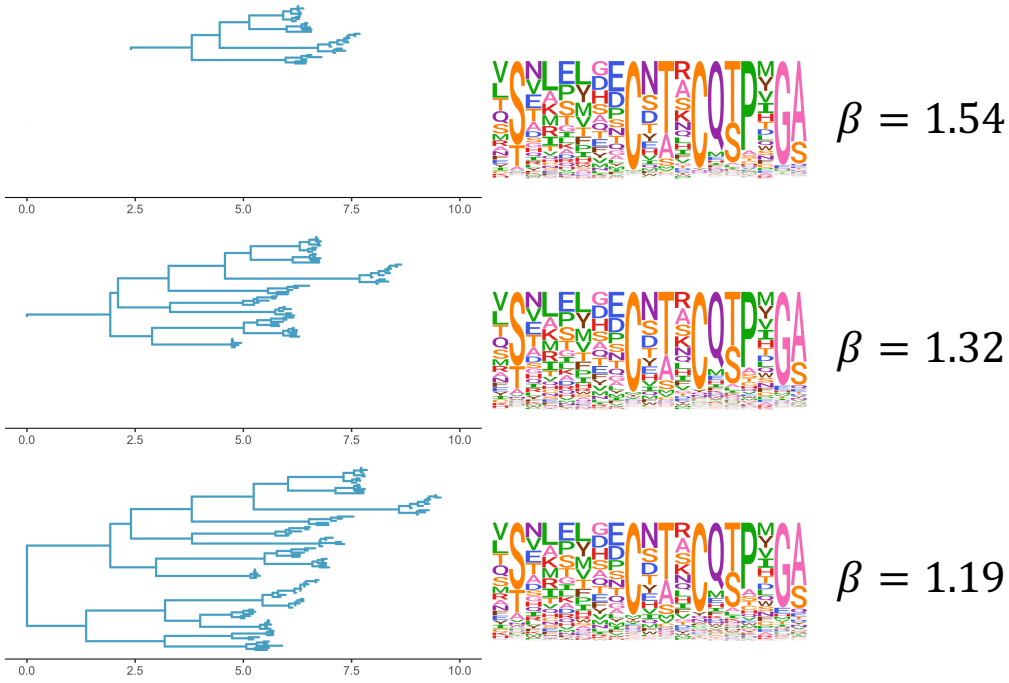


Figure 5: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree of 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 investigated the competing effects of shifting preferences and long branch estimation. That is, our preferences are most relevant to sequences they are close to but the effect of a non-uniform stationary is only observed on long branches.
  - We broke the large tree into small, sub trees with differing maximum divergence levels.
  - We estimated the branch lengths using the same models we used for empirical data above.
- The "relevance" of the preferences is inversely correlated with sequence divergence.
  - We used the ExpCM stringency parameter as the measure of preference "relevance". The larger the β value the more "relevant" the preferences.

  - As the maximum divergence of the tree increases the β value decreases.

- All models estimate short branches relatively equally.

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) (**????**) 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 (**???**) 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 (**?**) 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 (**?**)

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

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 **?** 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

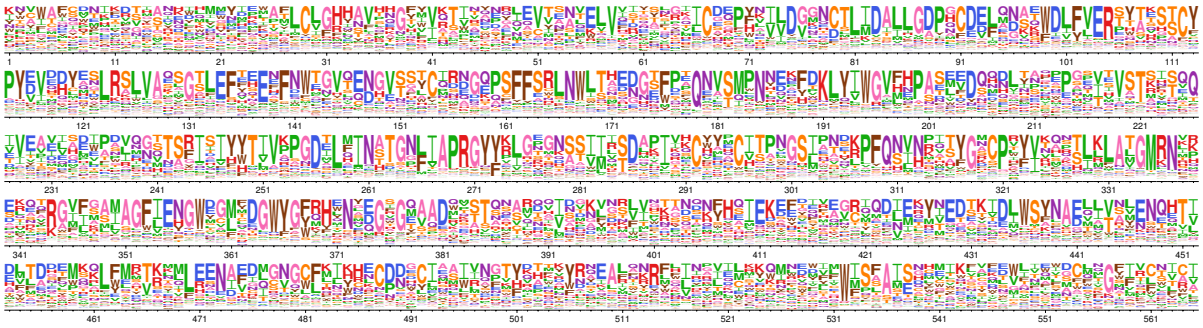
$$\sum_a \sum_{x \in a} p_{r,x} \sum_{y \in a} [M_r\,(t)]_{xy} \qquad \text{(Equation 6)}$$

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

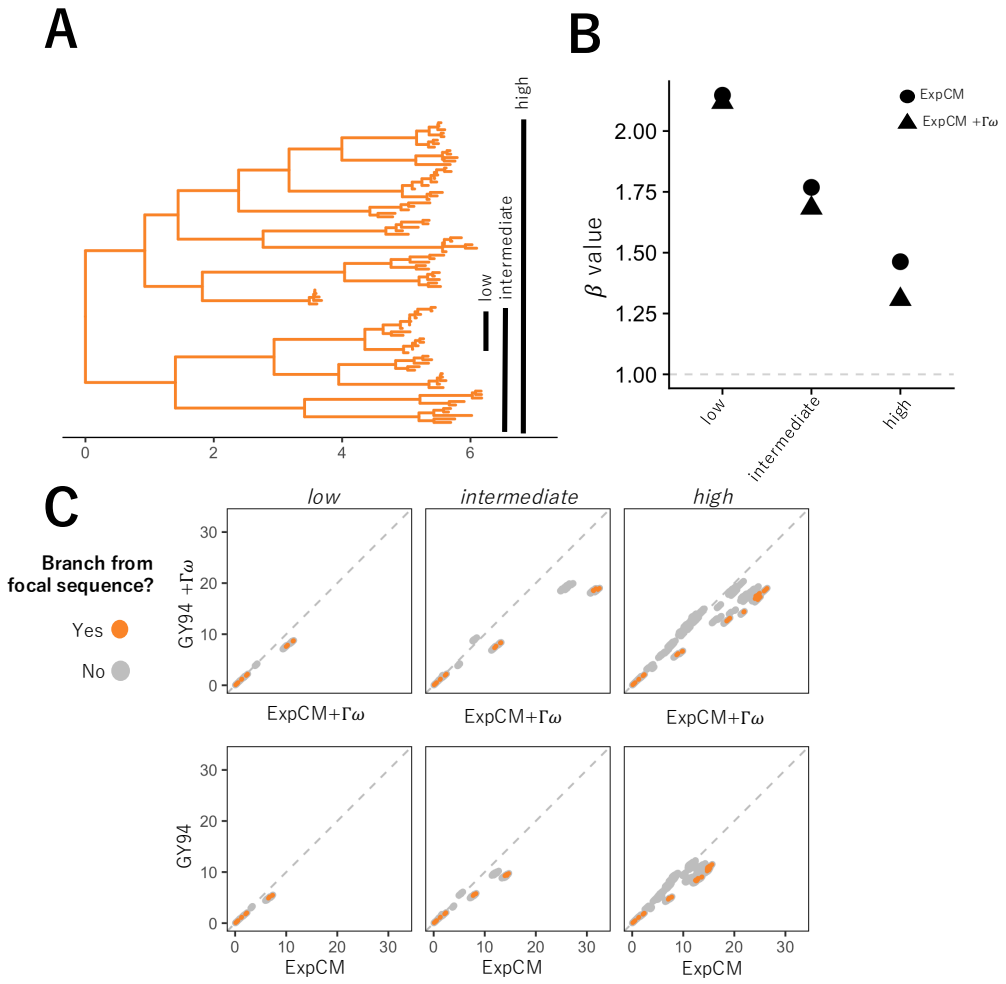
Table 1: ExpCM parameters used to simulate sequences in Fig. **??**.

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

Table 2: 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	$\kappa = 2.9984$
	$\omega = 0.09076$

Figure 6: The average of the H1 preferences measured by ? and the H3 preferences measured by *Lee* rescaled with the ExpCM stringency parameter optimized in ??A ($\beta = 1.77$)
I need to change the β value when the new `phydms` results finish running.



Supplementary figure 3: 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.