

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

Sarah K. Hilton<sup>1,2</sup> and Jesse D. Bloom<sup>1,2</sup>

<sup>1</sup>Division of Basic Sciences and Computational Biology Program,  
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

<sup>2</sup>Department of Genome Sciences, University of Washington, Seattle, WA  
E-mail: [jbloom@fredhutch.org](mailto:jbloom@fredhutch.org).

## Abstract

Molecular phylogenetics can be used to estimate the time since the divergence of modern gene sequences. For highly diverged sequences, such phylogenetic techniques often estimate substantially more recent times than other methods. In the case of viruses, there is independent evidence that the inferences from molecular phylogenetics are underestimates of the true deep divergence times. This discrepancy is caused in part by inadequate models of purifying selection leading to branch-length underestimation. Here we show that models informed by experimental measurements of purifying selection due to site-specific amino-acid preferences lengthen deep branches on phylogenies of influenza virus hemagglutinin. This deepening of branch lengths is due to more realistic stationary states of the models, and is independent of the branch-length-extension that results from modeling site-to-site variation in substitution rate. The branch-length extension from experimentally informed site-specific models is similar to that achieved by other approaches that allow the stationary state to vary across sites. However, the improvements from these site-specific 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"] skhcomment-  
from 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 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  $\Gamma$ -distributed  $\omega$  (dN/dS) (?). This flexibility in the value of  $\omega$  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.]

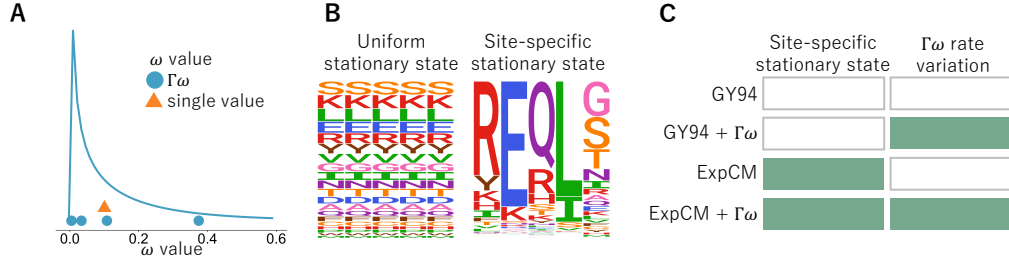
In order to test this hypothesis, we compared the branch lengths of a influenza virus HA phylogenetic trees optimized by different substitution models. We found that the ExpCM did extend the length of branches from the focal sequence on the tree [define focal] and that this extension was seen even in the context of  $\Gamma$ -distributed rate variation. Furthermore, we found this extension occurred even in the presence of  $\Gamma$ -distributed  $\omega$ , 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 that substitution models account for purifying selection

Proteins evolve under purifying selection to maintain their structure and function. This purifying selection is not homogenous across sites in a protein. It is also not homogenous among the different amino acids at a given site. For instance, some protein sites strongly prefer hydrophobic amino acids, others 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.

Here we consider how purifying selection is handled by codon models, which are the most



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

accurate of the three classes (nucleotide, amino acid, and nucleotide) of phylogenetic substitution models in widespread use (?). Codon models distinguish between two types of substitutions: synonymous and nonsynonymous. The relative rate of these substitutions is referred to as dN/dS or  $\omega$ . In their simplest form, codon substitution models fit a single  $\omega$  that represents the gene-wide average selection on nonsynonymous mutations relative to synonymous ones. Here we will use such substitution models in the form proposed by ?. Using the classification scheme of ?, when these models have a single gene-wide  $\omega$  they are called M0. We will refer to the M0 variant of the ? model simply as GY94. The gene-wide  $\omega$  is usually  $< 1$  [add citation, most of the site-specific  $\omega$  papers will talk about how gene-wide is not sensitive, so we can cite some of those], and crudely represents the fact that many amino-acid substitutions are under purifying selection.

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

A model formulation that accounts for the fact that purifying selection depends on the specific

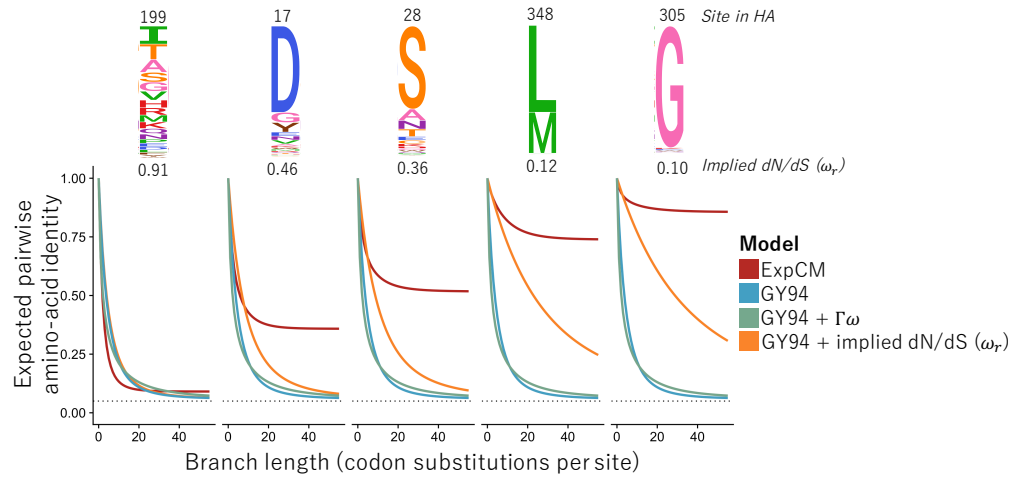
amino-acid mutation is provided by so-called “mutation-selection” models (????). These models by explicitly define a different set of amino-acid preferences at each site in the protein. This more mechanistic formulation results in a site-specific stationary state (Figure 1B). These models capture the site-to-site variation in amino-acid composition that is an obvious features of real proteins, and generally better describe actual evolution than models with only rate variation (????).

However, the increased realism of mutation-selection models comes at the cost of an increased number of parameters. Codon substitution models with uniform stationary states have only a modest number of parameters that must be fit from the phylogenetic data. For instance, a GY94+ $\Gamma\omega$  model with the commonly used F3X4 stationary state has 12 parameters: two describing the shape of the gamma distribution over  $\omega$ , a transition-transversion rate, and 9 parameters describing nucleotide composition of the stationary state. However, mutation-selection models must additionally specify 19 parameters defining the stationary state for *each* site (there are 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 9500 parameters for a 500-residue protein. It is challenging to obtain values for these parameters without overfitting the data (?). 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. Therefore, the number of remaining free parameters 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.

Importantly, these two strategies for modeling purifying selection are not mutually exclusive. Mutation-selection models such as ExpCM can still incorporate an  $\omega$  parameter, which now represents the relative rate of nonsynonymous to synonymous substitution *after* accounting for the constraints due to the site-specific stationary state. This  $\omega$  parameter for an ExpCM can be drawn from a distribution just like for GY94-style models. We will denote such models as ExpCM+ $\Gamma\omega$ . Figure 1C shows the full spectrum of models that incorporate all combinations of gamma-distributed  $\omega$  and site-specific stationary states.

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

Given a single branch, a substitution model transforms sequence identity into branch length. Under a molecular-clock assumption, this branch length is proportional to time. The transformation from sequence identity to branch length is trivial when the sequence identity is high. For instance, when where there has only been one substitution, then the sequence identity will simply be  $\frac{L-1}{L}$  for a gene of  $L$  sites, and even a simple exponential model (?) will correctly infer the short branch length of  $1/L$  substitutions per site. However, as more substitutions accumulate then it becomes



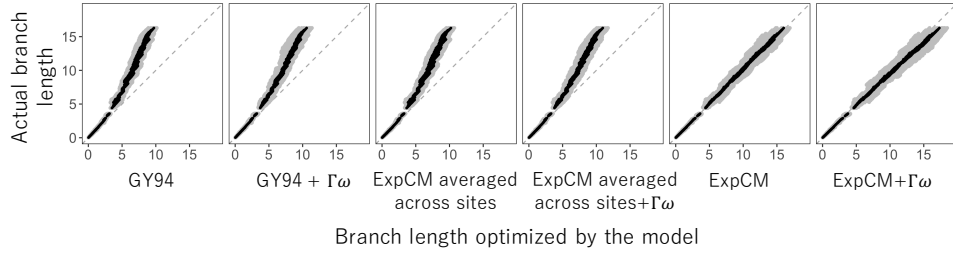
**Figure 2: Effect of stationary state and  $\Gamma\omega$  rate variation on predicted asymptotic sequence divergence.** The logo plots at top show the amino-acid preferences for some sites in an H1 influenza hemagglutinin protein as measured by ?. The graphs show the expected amino-acid identity at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

progressively more likely for multiple changes to occur at the same site. In this regime, the accuracy of the substitution model becomes critical for transforming the sequence identity into branch length. Any time-homogenous substitution model predicts that after a very large number of substitutions, two related sequences will approach some asymptotic sequence identity. For instance, if all 20 amino acids are equally likely in the stationary state, then this asymptotic sequence identity will be 0.05. If the substitution model underestimates the asymptotic sequence divergence then it will also underestimate long branch lengths, since it will predict that sequences that have evolved for a very long time should be more diverged than is actually the case.

Figure 2 shows how different substitution models predict sequence identity to decrease as a function of branch length for model parameters fit to a phylogeny of H1 influenza hemagglutinin (HA) genes. The GY94 model predicts the same behavior for all sites, since it does not have any site-specific parameters. This model predicts an asymptotic sequence divergence of [?], which is slightly higher than 0.05 since some of the 20 amino acids are favored due to more redundant codons and biases towards certain nucleotides. Intuitively, this asymptotic sequence identity of [?] seems low, since like many proteins HA has a highly conserved structure and function that imposes constraints that cause some sites to only sample a small subset of the 20 amino acids among all known HA homologs[citation, I think Juhye has something she cites that looks at really diverged HAs].

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

However, mutation-selection models such as ExpCM's have site-specific stationary states. Therefore, they predict that different sites will have different asymptotic sequence identities (Figure 2)—a prediction that accords with the empirical observation that some sites are much more



**Figure 3: Branch lengths inferred on data simulated under a model with site-specific amino-acid preferences.** We simulated alignments along a phylogenetic tree of HA genes (see Figure ??) using an ExpCM parameterized by the actual site-specific amino-acid preferences measured by deep mutational scanning of an H1 HA (?). We then inferred the branch lengths of this tree on the simulated alignments. The inferred branch lengths for various models are plotted on the x-axis, and the actual branch lengths used in the simulations are on the y-axis. We performed 10 simulations and inferences, and gray points show each inferred branch length from each simulation, and black points show the average of each branch length across simulations. The grey dashed line at  $y = x$  represents the behavior of an unbiased estimator. [Because of order we will discuss in text, switch order of ExpCM and ExpCM averaged across sites.]

variable than others in alignments of highly diverged sequences. For instance, Figure 2 shows that at sites such as 348 and 305 in the H1 HA, an ExpCM but not a GY94-style model predicts that the divergence will always be low. When sites with highly constrained amino-acid preferences such as these are common, an ExpCM can estimate a very long branch length at modest sequence identities that GY94 model might attribute to a much shorter branch.

### Simulations demonstrate how failure to model site-specific stationary states leads to branch-length underestimation.

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

The models with a uniform stationary state underestimated the lengths of long branches separating the simulated sequences (??). The GY94 model estimated branches lengths which are only about 60% of the true values for the longest branches. Accounting for site-to-site variation in  $\omega$  did not fix the fundamental problem: the GY94+ $\Gamma\omega$  did slightly better, but still substantially



**Table 1:** Performance comparison for the models in ???. All ExpCMs describe the evolution of HA better than the GY94 models, as evaluated by the Akaike information criteria ( $\Delta AIC$ , ?) The mean  $\omega$  value from each  $K = 4$  bins is shown for the models with  $\Gamma\omega$  rate variation, otherwise simply the fit, gene-wide  $\omega$ . Each ExpCM has a stringency parameter  $> 1$ . [Let's get rid of the 2nd and 3rd column]

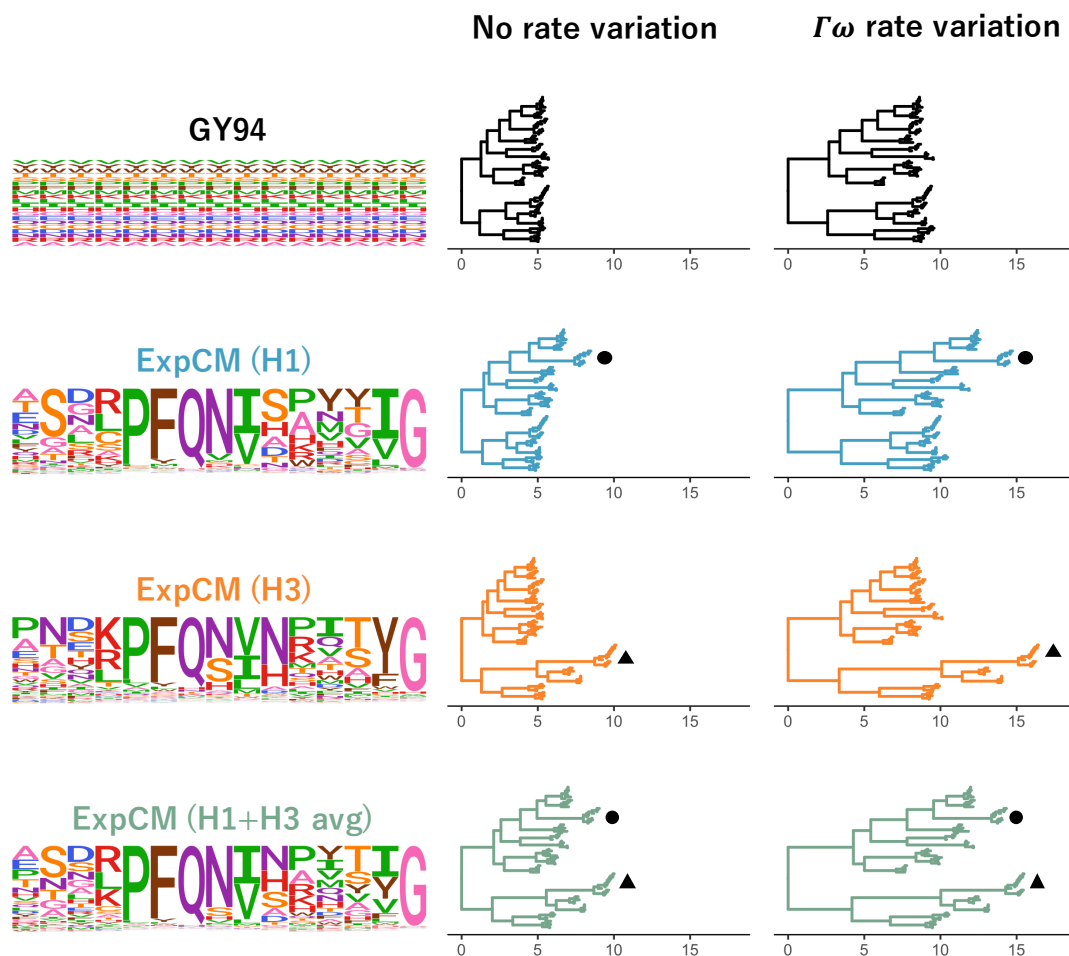
Model	Site-specific stationary state	$\Gamma\omega$	$\Delta AIC$	Log Likelihood	$\omega$ (implied dN/dS)	stringency parameter
ExpCM + $\Gamma\omega$ (H1+H3 avg)	+	+	0	-48751	0.19, 0.50, 0.90, 1.86	1.70
ExpCM (H1+H3 avg)	+	-	950	-49227	0.15	1.78
ExpCM + $\Gamma\omega$ (H1)	+	+	1306	-49404	0.13, 0.44, 0.91, 2.16	1.12
ExpCM + $\Gamma\omega$ (H3)	+	+	1737	-49620	0.09, 0.33, 0.72, 1.77	1.28
ExpCM (H1)	+	-	2556	-50030	0.13	1.22
ExpCM (H3)	+	-	3197	-50350	0.12	1.45
GY94 + $\Gamma\omega$	-	+	4719	-51106	0.00, 0.03, 0.08, 0.26	-
GY94	-	-	7625	-52560	0.07	-

underestimated the length of the longest branches. However, there was no systematic underestimation of long branches by the ExpCM and ExpCM+ $\Gamma\omega$  models, which accurately accounted for the site-specific amino-acid preferences used in the simulations. The improved performance of the ExpCM's is due to their site-specific modeling of the stationary state: if we parameterize these models by site-specific amino-acid preferences that have been averaged across HA sites, then they perform no better than GY94 models. Therefore, models with uniform stationary states are fundamentally incapable of accurately estimating the length of long branches in phylogenies of sequences that have evolved under strong site-specific amino-acid preferences.

### Site-specific models estimate longer branches on real data.

The foregoing section shows the superiority of ExpCM's to GY94 models for estimating long branches on phylogenies simulated with ExpCM models. But how do these models perform on real rather than simulated data? Real genes do evolve under site-specific constraint, but this constraint is almost certainly not modeled in all its complexity by an ExpCM. However, if ExpCM's do a substantially better job of modeling the constraint, we might still expect them to estimate more accurate branch lengths.

To test the models on real data, we used actual sequences of influenza HA. The topology of HA phylogenetic trees (??) makes these sequences an interesting test case for branch-length estimation. HA consists of a number of different subtypes. Sequences within a subtype have [?] amino-acid identity, but sequences in different subtypes have as little as 38% identity. However, HA proteins from all subtypes (with the exception of an unusual subtype of bat influenza virus that we exclude from this analysis [?]) have a highly conserved structure that performs a highly



**Figure 4: Effect of site-specific stationary state and  $\Gamma\omega$  rate variation on HA branch length estimation.** The HA tree is optimized by models from the ExpCM and GY94 families. The amino-acid preferences defining the model (ExpCM) or implied by the model (GY94) are shown as logoplots. The amino-acid preferences measured by the deep mutational scans defining the ExpCMs were measured in either an H1 background, an H3 background, or were the average measurements from the H1 and H3 scans. The circle denotes the H1 clade and the triangle denotes the H3 clade.

conserved function[citation, Juhye can probably help]. We used RAxML with nucleotide model of substitution (GTRCAT) to infer a phylogenetic tree for 87 HA sequences drawn from 14 of the 18 subtypes (we excluded bat influenza and three other rare subtypes). For the rest of this paper, we fix the tree topology to this RAxML-inferred tree. Although the nucleotide model used with RAxML is probably less accurate than codon models, the modular subtype structure of the HA phylogeny means that most of the phylogenetic uncertainty probably lies in the length of the long branches separating the subtypes rather than in the tree topology itself.

Given this fixed HA tree topology, we estimated the branch lengths on the tree using each of the substitution models in Figure 1C. There are two all amino-acid deep mutational scanning datasets for HA [citations], and each parameterizes a different ExpCM. One scan measured the site-specific amino-acid preferences in an H1 background (??, circle) and the other measured the preferences in an H3 background (??, triangle). These two deep mutational scans sample the breadth of HA diversity as their respective focal sequences are only  $\sim 42\%$  identical on the amino acid level. By comparing these two ExpCMs, we can examine the effect of the experiment's focal sequence on branch length estimation, along with the effect of site-specific stationary state and  $\Gamma\omega$  rate variation.

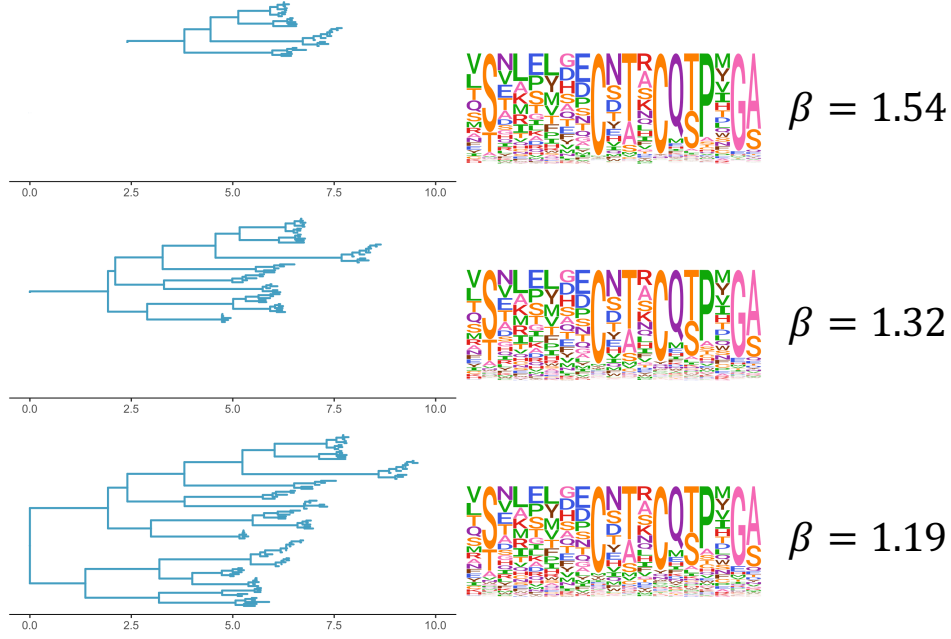
[The results in Figure ? show the distinct effects of modeling variation in  $\omega$  and variation in the stationary state... (Basically, more clearly delineate the two effects)] Models with either site-specific stationary state or  $\Gamma\omega$  rate variation estimated longer branches than models without (??). Furthermore, we can isolate the effect of each method. For example, the increase in branch length estimated by GY94+ $\Gamma\omega$  compared to GY94 and by ExpCM+ $\Gamma\omega$  compared to ExpCM shows that  $\Gamma\omega$  rate variation increases branch length in the presence and absence of a site-specific stationary state. We also see an increase in branch length with the addition of a site-specific stationary state in the presence and absence of  $\Gamma\omega$  rate variation (??, GY94 vs. ExpCM and GY94+ $\Gamma\omega$  vs. ExpCM+ $\Gamma\omega$ ). The model with both methods, ExpCM+ $\Gamma\omega$ , estimated the longest branches of all the models. However, the branch length extension from any of the ExpCM compared to any of the GY94 models 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 (??, circle for H1 and triangle for H3) This local effect of the ExpCM indicates that the stationary state described by the deep mutational scanning preferences is most accurate near the focal sequences of the scan.

The local effect of the ExpCM is not entirely surprising. Proteins are affected by epistatic interactions, where the effect of a mutation at one site is dependent on the amino-acid identity at another site. While one of the strengths of deep mutational scanning is the ability to accurately measure the effect of all single amino-acid mutations, these measurements are done in the context of a single genetic background and do not take into account epistatic interactions. Consequently, the stationary state defined by these preferences is most accurate for only a subset of the tree. An

ExpCM with a more general stationary state, defined by the average of the H1 and H3 preferences, lengthen the preferences from both the H1 and H3 clades (??).

### **Competing effects of shifting preferences and long branches.**

[I know I was pressing on this a lot, but I think that the three trees in Figure 5 turn out not to be useful as they aren't aligned in a helpful way. So I would suggest instead creating something that is a hybrid between the current Supplementary figure 4A and the Figure 5B. It shows the sequence subsets, and then just shows that the stringency parameter goes down for more diverged alignments. I like the Supplementary figure 4A as it shows the trees most clearly. I like the Figure 5B because it gives the stringency parameters but also shows their actual effects in logos, which is more intuitive than just plotting them on a graph.] [Two panels one for H1 and one for H3. Each panel has big tree with the subtrees labeled and the same prefs snippet as fig 4 with the different beta values for each subtree.]



**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  $\beta$  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.

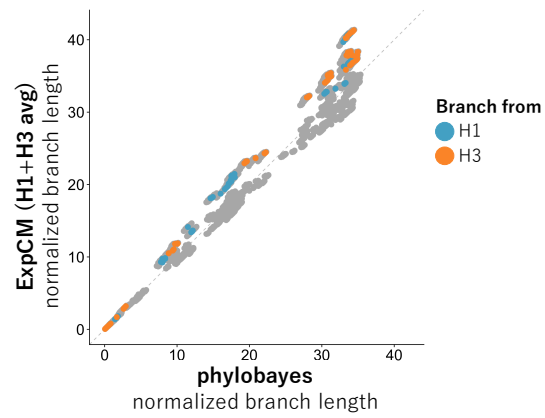
We wanted to determine the accuracy or relevance of the ExpCM on these subtrees. Due to the local branch length extension in ??, we expect that the preferences are most accurate on the low divergence tree. We use the ExpCM stringency parameters as a proxy for accuracy of the preference set. The stringency 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 **Experimentally Informed Codon Model (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

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

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:

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

The  $\kappa$  parameter represents the transition-transversion ratio, and the  $\phi_w$  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

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

where  $\mathcal{A}(x)$  is the amino-acid encoded by codon  $x$ ,  $\beta$  is the stringency parameter, and  $\omega$  is the relative rate of nonsynonymous to synonymous substitutions after accounting for the amino-acid preferences. The ExpCM has six free parameters (three  $\phi_w$  values,  $\kappa$ ,  $\beta$ , and  $\omega$ ). 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 (?)

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

## Theoretical effect of model choice on branch length

### Effect of model choice on natural sequences

#### ExpCM + $\Gamma\omega$ and YNGKP M5

#### Spielman $\omega_r$ values inferred from the ExpCM

We inferred the average nonsynonymous fixation rate from the ExpCM following ? as

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

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

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

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



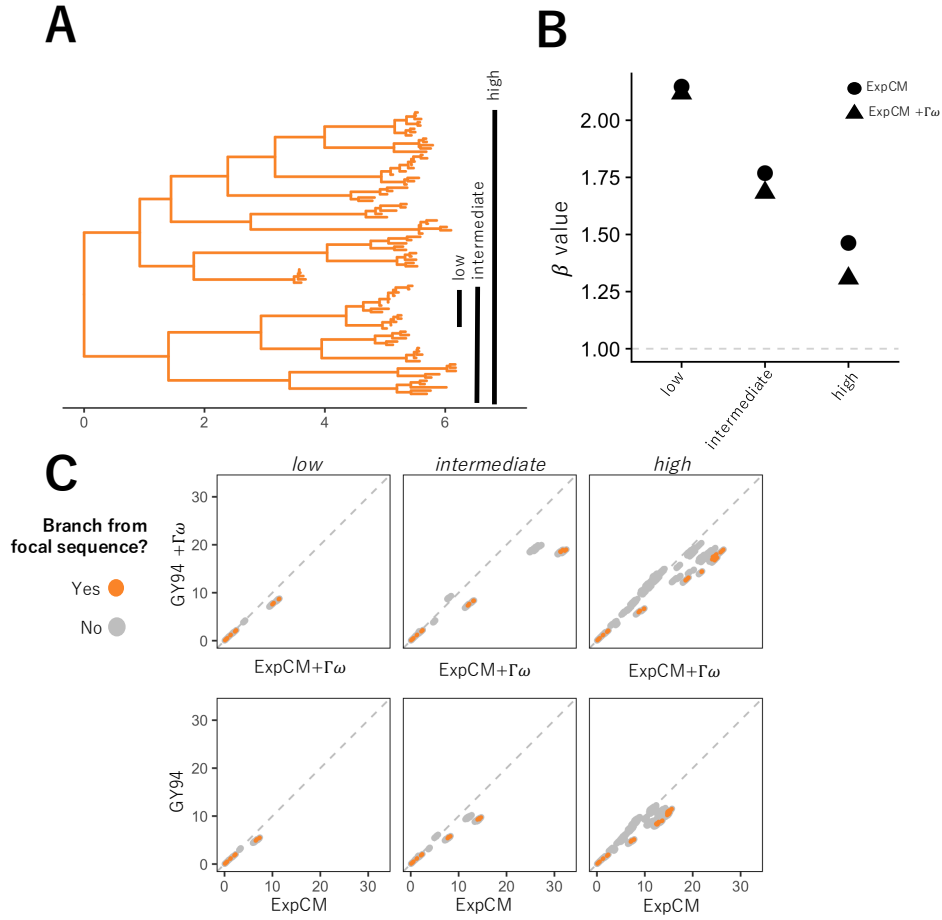
**Supplementary figure 1: H1 preferences measured by ? rescaled with the ExpCM stringency parameter optimized in ??A ( $\beta = 1.19$ )** [I need to change the  $\beta$  value when the new phydms results finish running.]



**Supplementary figure 2: H3 preferences measured by *lee* rescaled with the ExpCM stringency parameter optimized in ??A ( $\beta = 1.46$ )**



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



**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  $\beta$  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

- Arenas M. 2015. Trends in substitution models of molecular evolution. *Frontiers in genetics*. 6:319.
- 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.
- Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*. 11:725–736.
- 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.
- 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.
- Nielsen R. 2006. Statistical methods in molecular evolution. Springer.
- Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*. 53:793–808.
- Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*. 193:557–564.
- 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.
- Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*. 190:1101–1115.
- 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. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*. 25:568–579.
- Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155:431–449.

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