©Copyright 2020
Sarah Kelley Hilton

Modeling the effects of site-specific amino-acid preferences on protein evolution.

Sarah Kelley Hilton

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Jesse D. Bloom, Chair

Erick Matsen

Maitreya Dunham

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Modeling the effects of site-specific amino-acid preferences on protein evolution.

Sarah Kelley Hilton

Chair of the Supervisory Committee:
Associate Member Jesse D. Bloom
Fred Hutchinson Cancer Research Center

An important goal in the study of protein evolution is understanding which genetic changes that fixed in nature were selected for and why. However, understanding the functional consequence of any given mutation is a challenge because the effects of amino-acid changes are highly idiosyncratic across sites in a protein. My graduate research has focused on developing computational tools and methods that integrate two existing methods, phylogenetics and deep mutational scanning, to understand the effect of site-specific amino-acid constraint on protein evolution. These two methods, one computational and one experimental, have complementary strengths and weaknesses. Phylogenetics provides methods to study natural sequences, which are subjected to natural selective pressures, in a principled manner; however, these methods are constrained to the genetic sequences we have sampled. Deep mutational scanning allows for the unbiased measurement of all single amino-acid changes to a protein, but the assay occurs in an artificial laboratory setting. The goal of my work is to leverage the strengths of each method, the comprehensiveness of the deep mutational scan with the realism of comparative sequence analysis, for a more complete and accurate understanding of site-specific protein constraint.

In Chapter 2, I develop a web-based visualization tool, `dms-view`, for interactive exploration of deep mutational scanning experiments. `dms-view` addresses common analysis challenges by allowing the user to easily and iteratively view site-level summary metrics,

individual mutation measurements, and the 3-D protein structure for site(s) of interest from a deep mutational scan. `dms-view` is a flexible tool that allows the user to explore the site-specific amino-acid preferences measured by a given deep mutational alongside external datasets, such as site-specific amino-acid frequencies observed in nature.

While tools like `dms-view` allow for qualitative comparison of natural selection and selection in the lab, more sophisticated methods are needed to make this comparison while account for sequencing sampling and shared evolutionary history. To this end, in Chapter 3, I implement a relatively new family of phylogenetic substitution models called Experimentally Informed Codon Models (ExpCMs) in a new Python software package called `phydms`. ExpCMs describe the selection on amino-acid changes using the empirical measurements from a deep mutational scan and therefore represent a bridge between selection in the laboratory and selection in nature. `phydms` implements the models in maximum-likelihood framework and includes auxiliary command line tools to facilitate fast and easy analysis.

In Chapter 4, I investigate the effect of the site-specific ExpCMs with empirical measurements on phylogenetic inference, specifically branch length estimation. A long-standing observation in phylogenetics is that long branches in phylogenetic trees are consistently underestimated. I found that site-specific ExpCMs estimated longer branches than a common site-uniform codon model but that this extension in branch length was limited by intraprotein epistasis. This work suggests that current phylogenetic models assumptions of independent evolution between sites and identical evolution among sites results inaccurate branch length estimation.

Overall, my graduate work has produced general computational methods and tools that integrate empirical measurements of site-specific amino-acid constraint with comparative sequence analysis to create a more complete picture of protein evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Graduate school has been an incredible experience due in no small part to the amazing people in my life. I am so glad my thanks and acknowledgment of them will live alongside my science as a cohesive unit.

First, thank you to Jesse Bloom for being an amazing graduate mentor. Jesse has an infectious enthusiasm about science. My desk in lab (Jesse's old desk!) is near a doorway. Throughout the years, out of the corner of my eye, I would see Jesse powerwalking down the hallway and then several minutes later powerwalking back to his office. Then, it would feel like I could hear a cartoon tire squeal as Jesse backs up and comes into lab asking a million questions a minute. "Sarah, did your job finish running?" "Did you read the paper, what do you think?" "Adam, have you heard back about the antibodies?" "Kate, have you read your neuts?" I never felt like he thought I was not working hard enough but rather that he genuinely cares about the questions, the data, and the process of science. Jesse, I learned so much from you about leadership. I have never worked with someone who solicits feedback so frequently or implements change based on that feedback so quickly. From day one, you made me feel like a valued scientist and member of the team and that has been incredibly empowering. Finally, thank you for teaching me to be a good computational biologist. I came into the lab with coding basics but everything I know about writing good, reproducible code in a collaborative environment I learned from Jesse. I plan on taking Jesse's diligence, enthusiasm, and leadership forward with me.

I want to thank all members of the Bloom lab, past and present, for their support over the years. This is a group of great scientists and very cool people. I've learned so much from you all about virology, proteins, immunology and several hundred other

random things! I'd rate our lunchtime (12pm on the dot!) conversations as 18,000 tacos out of three. I especially want to thank past lab members who helped me get settled in lab and taught me how to be a good labmate: Orr, Mike, Heather, Katherine, Juhye, and special thanks to my rotation mentor Hugh!

I want to thank my Genome Sciences cohort: Andy Lin, Ken Jean-Baptiste, Juhye Lee, Jolie Carlisle, Khrystyna North, Jonathan Packer, Steven Lee, and Eliah Overbey. From methods and logic to defense celebrations, I am very grateful I got to share the last five years with you.

Thank to everyone involved with the Fred Hutch Girls Who Code club. Thank you Sidney Bell for starting the club. Thank you to Emma Hoppe and Kate Crawford for taking over the club and working hard on improving it. Thank you Jeanne Chowning, Liza Ray, and Dave Vannier from Fred Hutch Education Outreach for your support and guidance. Thank you to the girls themselves! Hanging out with you after work was so fun and inspriring! Thank you the robot for working 40% of the time.

Thank you to Jon Mah. Jon, it has been so fun working with you over the past three years on phylogenetics projects. You are a great computational biologists and a great colleague. I cannot wait to see all of the exciting things you do at UCSD!

Thank you to John Huddleston, my co-lead on `dms-view`. John, I learned so much from about being a thoughtful, rigorous computational biologist. Since we started working together, I have written so many slack messages that start with "John Huddleston taught me to send notes right after a meeting, so here you go!" I will miss our early morning coffee meetings and quiet work time at the Yale building coffee shop.

Thank you to the HIV corner: Hugh Haddox, Adam Dingens, Kate Crawford, David Bacsik, and Jeremy Roop. You have been incredibly supportive colleagues and friends. Thank you for caring about the minutiae of my science as much as I do, thank you for great taste in beer, and thank you for willingness to go on adventures.

Thank you to my phyload comrades, Andy Magee and Will DeWitt. There are about 192 reasons why this project was one of my favorite parts of graduate school. It was great having the freedom to think of an idea, plan a study, and realize our mistakes as a group of graduate students. I especially want to thank Will DeWitt for his incredible support. Will, your serious advice about science, life, and pizza paired with your absolute goofiness has gotten me through a lot.

Thank you to my friends outside of biology: Muriel Moore, Taylor Soja, Devin Sawyers, Juliana Schiffgens, and Josh Veden. Thank you for listening to me babble about science for hours on end and, more importantly, thank you for talking to me about what interests you and preventing me from collapsing into a one-dimensional person. Thank to Taylor for many fortifying conversations about the process of getting your PhD. I could not have done this even half as gracefully without your empathy and can-do-it-ness. Thank you, Muriel Moore, for being an amazing friend and a truly ruthless editor. I will always be grateful for our conversations where you affirmed that I could handle graduate school and gave me the space to think critically about my role in the world as a scientist.

I want to thank my family who has been an incredible source of support over the last five years. Andrew, I have enjoyed our almost daily text conversation about how science, computer science, and coding are cool. I am excited it is not going to stop anytime soon. Thank you to my dad, Scott, for always encouraging my more quantitative interests when my default is to downplay them. I remember when I was in college and trying to avoid taking a computer science class you said to me ”I don’t understand why you think you won’t be good or like it.” You were the first and most consistent voice that a computational PhD was something I could do. Thank you to my mom, Judie. You taught me that wherever you are, you look for community. And once you find community you are obligated to try your best to make that community better. This is not something that is taught in science so I am grateful I had a strong foundation from you.

Finally, there are many more people who have made the last five years such a great adventure. Thank you!

Chapter 1
INTRODUCTION

Protein sequence evolution is governed both by the rate at which mutations arise and how selection acts upon these mutations. An important goal in the study of protein evolution is understanding which genetic changes that fixed in nature were selected for and why. A substitution could be positively selected for because it confers a new, beneficial phenotype, or a substitution could fix because it maintains the form and function of the protein and therefore is not purged by purifying selection. Understanding the functional consequence of any given mutation is a challenge because the effects of amino-acid changes are highly idiosyncratic across sites in a protein [32]. In some cases, proteins can evolve for billions of years while maintaining the same function [65] and in other cases a single amino-acid change results in a dramatic phenotypic change [115, 113, 2]. As Zukerlandl and Pauling noted over 50 years ago, "it is the type rather than number of amino acid substitutions that is decisive" [166]. This observation is a succinct summary of the large, sprawling challenge that measuring or inferring the site-specific amino-acid preferences of a protein poses. Here I will describe my graduate work developing tools, methods, and models focused on learning and understanding the effect of site-specific amino-acid preferences on protein evolution.

While the methods and principles I used are relevant to protein evolution generally, I focused my work on understanding the site-specific constraint on Influenza A Virus (influenza). Influenza evolution is of great interest from a public health perspective. The WHO estimates 290,000 - 650,000 deaths each year are due to seasonal influenza (`https://www.who.int/influenza/surveillance_monitoring/bod/en/`). Understanding how in-

fluenza evolves could inform vaccine design [142], help develop new therapies [57], or understand the dynamics of zoonotic crossovers [119, 83, 47].

Beyond the public health interest, the basic biology of influenza makes it an interesting test case for protein evolution methods and questions. Like all RNA viruses, influenza has a high mutation rate ($10^{-5} - 10^{-4}$ nucleotide mutations/site/genome [133, 91, 117, 11]) that provides a large substrate for selection to act upon. Furthermore, influenza proteins experience both strong purifying selection and diversifying selection. This push and pull is particularly evident for the influenza surface protein hemagglutinin (HA). On one hand, HA is under strong purifying selection to main its essential function of allowing viral entry by binding and fusing to host cell membranes [148]. This selection is reflected by the high structural conservation of genetically-diverged HA homologs [49, 114]. On the other hand, HA is under strong diversifying selection as the major target of the humoral immune response to influenza [56, 10]. Finally, influenza is well suited for the phylogenetic methods I discuss below because recombination events are rare [15]. This is in stark contrast to the high recombination rates in other RNA viruses, such as coronaviruses [164] and lentiviruses [64].

Here I will introduce the two methods I worked with to understand the site-specific constraint on influenza proteins: an experimental assay called deep mutational scanning and the statistical framework of molecular phylogenetics. Deep mutational scanning [40] is a high-throughput assay which measures the effect of every single amino-acid change on some protein function measured in the laboratory. A deep mutational scan produces a *complete* map of the site-specific amino-acid preferences for a given protein and a given function. Molecular phylogenetics is a set of computational and statistical methods, algorithms, and models for comparative sequence analysis. Specifically, I worked on integrating these two methods in the form of a phylogenetic substitution model defined almost exclusively from deep mutational scanning measurements. The goal of this work is to leverage the strengths of each method, the comprehensiveness of the deep mutational scan with the realism of comparative sequence analysis, for a more complete and

accurate understanding of site-specific protein constraint.

1.1 Deep Mutational Scanning (DMS): a high-throughput functional assay to measure site-specific amino-acid preferences in the laboratory

It is widely appreciated that laboratory experiments are useful to understand natural protein evolution [24, 53]. Specific hypotheses of mutational effects can be tested in the laboratory with precise and controlled experiments and the effect of the mutation in the lab can then be compared to its evolutionary fate in nature [24, 53]. However, such targeted experiments are limited in the breadth of site-specific constraint they are able to capture. They require either defining a specific set of mutations *a priori* or randomly sampling a subset of possible mutations.

Recent technological developments in the form of a high-throughput functional assay called deep mutational scanning has made it possible to experimentally measure the effects of all single amino-acid mutations to a protein [40]. The basic technique involves creating a library of genetic variants, each containing a single amino-acid step away from wildtype, and then performing bulk selection to select for functional variants. Deep sequencing of the starting and selected library is used to calculate the relative enrichment or depletion of each variant, creating a set of site-specific amino-acid preferences. Over the past five years, this technique has been applied to dozens of different proteins. These studies have varied research goals from basic evolutionary questions [26, 130, 132] to applied clinical work [128, 42]

A strength of deep mutational scanning is that it is a general platform for measuring site-specific constraint and is flexible in the experimental design. For example, the laboratory selection can be modified to measure a specific aspect of the protein. Studies have used deep mutational scanning to look at protein abundance [78], protein function [39, 82, 84], or protein binding [81]. Furthermore, studies have deep mutational scans on proteins in their native genomic contexts [111, 27, 132, 82] and in more experimentally tractable systems [78, 129, 68]. However, no matter the exact experimental setup, the

goal and results of a deep mutational scan are the same: a map of site-specific amino-acid preferences.

In recent years, deep mutational scanning has increasingly been used to study the site-specific constraints of viral proteins. There have been deep mutational scanning studies for influenza [121, 27, 26, 154, 6, 75, 155, 153], HIV [50, 38, 3], Zika Virus [122], and Hepatitis C [101]. For influenza specifically, there are deep mutational scans for five of eight influenza virus proteins. These viral deep mutational scans were designed to test the effect of amino-acid changes under a variety of selective pressures relevant to viruses, such as viral growth [121, 27, 75] or escape from antibodies [74, 25], innate immune factors [6], or anti-viral drugs [101].

Deep mutational scanning provides a comprehensive map of amino-acid preferences, but the results may be limited in their relevance to natural sequence evolution by two experimental restrictions. The first restriction is that the experimental setup must be tractable in a laboratory setting at scale. For example, Doud and Bloom (2016) [27] discuss how their use of a lab-adapted influenza strain or how growth in cell culture might be inadequate to capture the complexity of a natural infection. A second restriction is due to the mutagenesis strategy of a deep mutational scan. In a deep mutational scan, each variant carries one or very few amino-acid changes. This constrains the deep mutational scan to a single "focal" sequence and blinds the scan to the effect of genetic background on modulating the effect of specific amino-acid mutations. However, both theoretical work [96] and experimental work using deep mutational scanning [51, 26, 131] shows that site-specific amino-acid preferences "shift" over time as a result of intraprotein epistasis. The restriction in both genetic background and experimental setup could limit a deep mutational scan's ability to explain the long-term evolution of a given protein.

The important question is not in what are all the *possible* ways could experimental artifacts limit the relevance of a deep mutational scan to natural sequence evolution. Rather, it is whether we can identify this effect. A common and easy analysis is to simply compare the site-specific amino-acid frequencies from a deep mutational scan to the amino-

acid frequencies observed in nature. However, this qualitative analysis is not tractable or particularly informative for the complete map of amino-acid preferences from a deep mutational scan. Furthermore, the calculation of amino-acid frequencies from natural sequences are confounded by shared evolutionary history of the sequences and sampling concerns. More sophisticated methods are needed to adequately make this comparison.

1.2 Molecular phylogenetics: statistical methods to explore evolution through comparative natural sequence analysis

The desired framework is provided by molecular phylogenetics, a statistical method for comparative sequence analysis. In both maximum-likelihood and Bayesian phylogenetic approaches, an evolutionary substitution model is used to calculate the likelihood of the observed sequences given a phylogenetic tree [36, 63]. Using phylogenetic methods, biologically-interpretable parameters can be estimated from a set of extant sequences. For example, estimated tree topologies describe the sequences' shared evolutionary history, while branch lengths describe the timing of major evolutionary events on the tree and substitution model parameters describe the rate of specific mutation types.

Part of phylogenetic inference is defining a substitution model, which describes the probability of character state change along a tree. Codon models are the most sophisticated substitution model for protein-coding genes[5] and common formulations include those described by Goldman and Yang [44, 160] and Muse-Gaut [87]. These models describe the instantaneous rate of change from one codon to another via single nucleotide changes. Importantly, these models differentiate between a single nucleotide change that preserves the site's amino-acid identity (synonymous change) and one that changes it (nonsynonymous). Specifically, the parameter ω represents the ratio of non-synonymous to synonymous mutations along the tree. The estimated value of this parameter is interpreted as indication of purifying($\omega < 1$), neutral($\omega \sim 1$) or positive selection ($\omega > 1$). In order to account for the nonhomogeneous amino-acid substitution rate across a protein, ω is commonly allowed to vary across sites according to a statistical distribution such as the

Γ-distribution [89, 160]. However, this additional model complexity affects the *rate* only and does not affect the stationary state frequencies of the model. That is, even with a Γ-distributed ω, these common codon models assume the same set of amino-acid preferences for each site in the protein. This assumption is not only a clear oversimplification of the current understanding of protein evolution but also may affect phylogenetic inference itself.

Mutation-selection models, first described by Halpern and Bruno (1998) [52], explicitly relax the assumption of identical evolution within the framework of a codon model. Mutation-selection models define a set of amino-acid preferences at each site in the protein, resulting in a model with a site-specific stationary state. Recent work has shown that mutation-selection models are often better descriptions of natural sequence evolution than site-uniform codon models, as assayed by Bayesian or maximum-likelihood criteria [72, 73, 102, 141, 110, 11, 12, 59]. However, the more explicitly account for site-specific purifying selection comes at the cost of a large increase in the number of parameters. Mutation-selection models have issues with overfitting [105] and convergence [125] for maximum-likelihood and Bayesian approaches respectively, which may limit their ability to accurately estimate or infer the site-specific amino-acid preferences from natural sequences.

The mechanistic parametrization of site-specific constraint by mutation-selection models compliments the empirical site-specific amino-acid constraints measured by deep mutational scanning. An alternative to estimating or inferring the site-specific amino-acid preferences, a mutation-selection model can be parametrized with measurements from a deep mutational scan. These Experimentally Informed Codon Models (ExpCMS) [11] account for site-specific constraint with fewer free parameters to be estimated from the natural sequence data than the site-uniform codon models. As a phylogenetic substitution model, ExpCMs represent a bridge between selection in the laboratory and selection nature and facilitates principled comparison of the two.

As a relatively new model, analysis with ExpCMs have been restricted to a small num-

ber of proteins. Previously, ExpCMs have been shown to be a better description of natural sequence evolution than site-uniform models for β-lactamase, influenza HA, influenza NP, and Gal4 [14]. This limited application of ExpCMs is due to a small number of available deep mutational scans at the time of publication and inadequate computational implementations for fast and efficient inference.

1.3 Layout of Dissertation

In this dissertation, I describe my graduate work developing computational tools, methods, and models to explore the effect of site-specific constraint on protein evolution. My work relies on and builds on two existing methods to study protein evolution, the experimental assay deep mutational scanning and site-specific phylogenetic substitution models. I've developed tools both for general analysis of deep mutational scans and specific analysis using ExpCMs, a phylogenetic model defined by deep mutational scanning data. I used these tools to explore protein-specific questions and questions about phylogenetic inference.

In Chapter 2, I developed a web-based visualization tool for deep mutational scanning data called `dms-view` (`https://dms-view.github.io`). The goal of `dms-view` is to allow researchers performing deep mutational scans to interactively explore their data and easily share analyses with others. Specifically, this tool links summary information at the site level of the protein with specific mutation-level measurements, all within the context of the 3-D protein structure. This link is important to interpreting the results of a deep mutational scan because sites that are discordant in linear, genomic space but contact in three-dimensional space. Additionally, while `dms-view` I designed with deep mutational scanning experiments in mind, it is a flexible tool that can display data from many different sources. In this way, the user can contextualize the data from their specific deep mutational scan with other data, such as amino-acid frequencies from natural sequences.

In Chapter 3, I wrote a python package, `phydms` [59], to implement and perform phy-

logenetic inference with Experimentally Informed Codon Models (ExpCMs). ExpCMs are mutation-selection models which use empirical measurements from deep mutational scans to model selection on amino-acid changes. ExpCMs provide a principled way to compare selection in the laboratory to selection in nature because the vast majority of the parameters used to describe the natural selection come from the laboratory experiments. I designed `phydms` to easily facilitate such analyses. For example, you can ask if an ExpCM is a better descriptor of natural sequence evolution that a standard site-uniform model or identify sites which are evolving faster or slower than expected. `phydms` performs these analyses much faster than previous implementations of the model and with methods and tools to make the analysis easier.

In Chapter 4, I investigated the effect of modeling site-specific amino-acid preferences using ExpCMs had on the phylogenetic inference of branch length estimation. A long-standing observation in the field of phylogenetics is that long branches on phylogenetic trees are consistently underestimated. In their original paper describing mutation-selection models, Halpern and Bruno hypothesized that modeling site-specific amino-acid constraint would alleviate this underestimation [52]. I tested this hypothesis by comparing influenza phylogenetic tree branch lengths estimated by site-specific ExpCMs to the branch lengths estimated by a site-uniform codon model. I found that the site-specific ExpCMs do estimate longer branches. However, since I used two ExpCMs defined by deep mutational scans from two diverged homologs, I was able to see how intraprotein epistasis limited branch length extension. This work supports previous work suggesting that phylogenetic inference, and branch length estimation specifically, is affected not only by the assumption that sites evolved identically but also the assumption that they evolve independently.

Chapter 2

DMS-VIEW: INTERACTIVE VISUALIZATION TOOL FOR DEEP MUTATIONAL SCANNING EXPERIMENTS

Sarah K. Hilton*, John Huddleston*, Allison Black, Khrystyna North, Adam S. Dingens, Trevor Bedford and Jesse D. Bloom

(* equal contribution)

2.1 Summary and Purpose

The high-throughput technique of deep mutational scanning (DMS) has recently made it possible to experimentally measure the effects of all amino-acid mutations to a protein [40]. Over the past five years, this technique has been used to study dozens of different proteins [33] and answer a variety of research questions. For example, DMS has been used for protein engineering [151], understanding the human immune response to viruses [74], and interpreting human variation in a clinical setting [128, 42]. Accompanying this proliferation of DMS studies has been the development of software tools [13, 112] and databases [33] for data analysis and sharing. However, for many purposes it is important to also integrate and visualize the DMS data in the context of other information, such as the 3-D protein structure or natural sequence-variation data.

Here we describe `dms-view` (`https://dms-view.github.io/`), a flexible, web-based, interactive visualization tool for DMS data. `dms-view` is written in JavaScript and D3 (`https://d3js.org`), and links site-level and mutation-level DMS data to a 3-D protein structure. The user can interactively select sites of interest to examine the DMS measurements in the context of the protein structure. `dms-view` tracks the input data and user selections in the URL, making it possible to save specific views of interactively generated visualizations to share with collaborators or to support a published study. Importantly, `dms-view` takes a flexible input data file so users can easily visualize their own DMS data in the context of protein structures of their choosing, and also incorporate additional information such amino-acid frequencies in natural alignments.

Users can access `dms-view` at `https://dms-view.github.io/`. The tool consists of a data section at the top and a description section at the bottom. The data section displays the user-specified data in three panels: the site-plot panel, the mutation-plot panel, and the protein-structure panel (1A). When sites are selected in the site-plot panel, the individual mutation values are shown in the mutation-plot panel and highlighted on the protein structure. The user can toggle between site- and mutation-level metrics, which

are defined in the user-generated input file. The description section is at the bottom of the page, and allows the user to add arbitrary notes that explain the experimental setup, acknowledge data sources, or provide other relevant information.

Please visit the documentation at `https://dms-view.githubio/docs` to learn more about how to use the tool, how to upload a new dataset, or view case studies.

2.2 Example: Mapping influenza A virus escape from human sera

Using a DMS approach, Lee and colleagues (2019) [74] measured the ability of every single amino-acid mutation in the influenza virus surface protein hemagglutinin to escape neutralization by human sera. For more information on the experimental setup, see the paper [74] or the GitHub repo (`https://github.com/jbloomlab/map_flu_serum_Perth2009_H3_HA`).

We visualized the serum mapping data using `dms-view`. To explore this dataset, please visit `https://dms-view.github.io/`. In the `dms-view` visualization of these data, the conditions are the different human or ferret sera used for the selections. The site- and mutation-level metrics are different summary statistics (`https://jbloomlab.github.io/dms_tools2/diffsel.html`) measuring the extent that mutations escape from immune pressure.

Lee and colleagues asked two questions in their paper which can be easily explored using `dms-view`.

1. **Are the same sites selected by sera from different people?** To explore this question, we compared the site-level and mutation-level metric values for a specific set of sites between different conditions.

2. **Where on the protein structure are the highly selected sites located?** To explore this question, we selected specific sites of interest to be visualized on the 3-D protein structure

2.2.1 Comparing site-level and mutation-level metric values for specific sites between conditions

To address whether or not the same sites are selected by different human sera using `dms-view`, we selected the mostly highly targeted sites for the human sera condition “Age 21 2010” 1A (144, 159, 193, and 222). We then use the condition dropdown menu to toggle between the other sera. The highlighted sites remain highlighted after the condition is changed so we can easily see if the same sites are targeted in other conditions.

In 1B, we can see that there is no overlap of the sites selected by the human sera “2010-age-21” the human sera “2009-age-53”. These data are the default data for `dms-view`, so to explore this question in more detail please see `https://dms-view.github.io/`.

2.2.2 View sites on the protein structures

To address where on the protein structure the targeted sites are located, we selected the most highly targeted sites (144, 159, 193, and 222) for the human sera condition “Age 21 2010” to highlight them on the protein structure.

In 1A, we can see that these sites cluster on the ”head” of the hemagglutinin, which is known to be a common target of the human immune system [21].

2.3 Code Availability

- dms-view is available at `https://dms-view.github.io/`.
- Source code is available at `https://github.com/dms-view/dms-view.github.io/`.
- Documentation (`https://dms-view.github.io/docs/`) and case studies (`https://dms-view.github.io/docs/casestudies/` are also available.

2.4 Acknowledgements

This work started as the final project for UW class CSE 512 Data Visualization as a part of the UW eScience Advanced Data Science Option curriculum and we would like to thank Dr. Jeffrey Heer, Halden Lin, and Jane Hoffswell for their input on the initial design. Thank you to Bloom and Bedford lab members for their generosity providing feedback, data, and time for testing. This work was supported in part by the following grants of the NIAID of the NIH: R01AI127983, R01AI141707, and R01AI140891. JDB is an Investigator of the Howard Hughes Medical Institute.

Chapter 3

PHYDMS: SOFTWARE FOR PHYLOGENETIC ANALYSES INFORMED BY DEEP MUTATIONAL SCANNING

A version of this chapter has been previously published as:

Sarah K. Hilton, Michael B. Doud, and Jesse D. Bloom. “phydms: Software for phylogenetic analyses informed by deep mutational scanning.” PeerJ. 5:e3657 (2017)

3.1 Abstract

It has recently become possible to experimentally measure the effects of all amino-acid point mutations to proteins using deep mutational scanning. These experimental measurements can inform site-specific phylogenetic substitution models of gene evolution in nature. Here we describe software that efficiently performs analyses with such substitution models. This software, `phydms`, can be used to compare the results of deep mutational scanning experiments to the selection on genes in nature. Given a phylogenetic tree topology inferred with another program, `phydms` enables rigorous comparison of how well different experiments on the same gene capture actual natural selection. It also enables re-scaling of deep mutational scanning data to account for differences in the stringency of selection in the lab and nature. Finally, `phydms` can identify sites that are evolving differently in nature than expected from experiments in the lab. As data from deep mutational scanning experiments become increasingly widespread, `phydms` will facilitate quantitative comparison of the experimental results to the actual selection pressures shaping evolution in nature.

3.2 Introduction

It is widely appreciated that experiments in the lab can inform understanding of protein evolution in nature [24, 53]. Efforts to synthesize experiments with evolutionary data have typically involved generating protein variants of interest, assaying their functionality in the lab, and qualitatively comparing the measured functionality of each variant to its evolutionary fate in nature [24, 53]. The recent advent of high-throughput deep mutational scanning techniques [40] has greatly expanded the potential of such research. For instance, numerous recent papers have reported measuring the effects of *all* amino-acid mutations on the functionality of a range of proteins [81, 111, 39, 92, 82, 11, 138, 132, 26, 68, 84, 27, 79, 50, 38, 77, 17]. This flood of data necessitates new methods for comparing experimental measurements to evolution in nature, since simple qualitative inspection is

insufficient when measurements are available for tens of thousands of mutants.

A solution is provided by the methods of molecular phylogenetics. Longstanding phylogenetic algorithms enable calculation of the statistical likelihood of an alignment of naturally occurring gene sequences given a phylogenetic tree and a model for the evolutionary substitution process [35, 36]. Deep mutational scanning data can be incorporated into this statistical framework via the substitution model [11]. Such an experimentally informed codon model (ExpCM) of substitution can be used to test whether a deep mutational scanning experiment provides evolutionarily relevant information [11], compare the stringency of selection in nature and the lab [12], assess how well different experiments describe natural selection on the same gene [26], and identify sites that are evolving differently in nature than expected from experiments in the lab [14].

However, a hindrance to such analyses has been the lack of appropriate software. Prior work using an ExpCM has re-purposed existing software such as `HyPhy` [98] or `Bio++` [48] to optimize the phylogenetic likelihood. Because these existing software packages are not designed for such site-specific models, the resulting analyses have been slow and cumbersome. Other software packages [135, 136, 110, 107] that handle site-specific codon substitution models are designed to treat the effects of mutations as unknowns to be inferred rather than as values that have been measured *a priori*.

Here we describe `phydms`, software for **phy**logenetics informed by **d**eep **m**utational **s**canning. We show that `phydms` is $\sim$100-fold faster than existing alternatives for performing analyses with an ExpCM, and demonstrate how it can be used to quantitatively relate measurements from deep mutational scanning with selection in nature. Readers who are interested in technical details of how `phydms` works should read the METHODS section; readers who are primarily interested in simply using `phydms` may prefer to jump directly to the RESULTS section.

3.3 Methods

3.3.1 Substitution models

Experimentally informed codon model (ExpCM)

The basic ExpCM implemented in `phydms` is identical to those in [14]. We recap this ExpCM to introduce nomenclature needed to understand the extensions described in the next few subsections.

In an ExpCM, rate of substitution $P_{r,xy}$ of site r from codon x to y is written in mutation-selection form [52, 80, 124] 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 [55] 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 (see the RESULTS section for more details on these preferences). The $F_{r,xy}$ terms are

defined in terms of these experimentally measured amino-acid preferences as

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

where $\mathcal{A}\left(x\right)$ 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. As shown in Figure 2, Equation 3 implies that mutations to more preferred amino acids are favored, and mutations to less preferred amino acids are disfavored. The functional form in Equation 3 was derived by [52] and under certain (probably unrealistic) population-genetic assumptions; under these assumptions, β is related to the effective population size. When $\beta > 1$, natural evolution favors the same mutations as the experiments but with greater stringency. 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.

ExpCM with empirical nucleotide frequency parameters

Phylogenetic substitution models commonly set the nucleotide frequency parameters (ϕ_w in the case of an ExpCM) so that the model's stationary state equals the empirical frequencies of the characters in the alignment. Setting the frequency parameters in this way reduces the number of parameters that must be optimized by maximum likelihood. Empirically setting the nucleotide frequency parameters is easy for substitution models where the stationary state only depends on these parameters.

However, the situation for an ExpCM is more complex. The ϕ_w values give the expected nucleotide frequencies in the *absence* of selection on amino acids, but in an ExpCM there is site-specific selection on amino acids. Therefore, the stationary state of an ExpCM also depends on other quantities: the stationary state frequency $p_{r,x}$ of codon x

at site r is [14]

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

where x_k indicates the nucleotide at position k in codon x. As this equation makes clear, the stationary state of an ExpCM depends on the preferences $\pi_{r,a}$ and stringency parameter β as well as the nucleotide frequency parameters ϕ_w.

So for an ExpCM, setting ϕ_w empirically means choosing their values such that the alignment frequency g_w of nucleotide w is as expected given the stationary state $p_{r,x}$. This will be the case if the following equation holds for all w:

$$g_w = \frac{1}{L} \sum_r \sum_x \frac{1}{3} N_w\left(x\right) p_{r,x} \qquad \text{(Equation 5)}$$

where L is the length of the gene in codons, r ranges over all codon sites, x ranges over all codon identities, and $N_w\left(x\right)$ is the number of occurrences of nucleotide w in codon x. We could not analytically solve this system of equations for ϕ_w in terms of g_w, so we instead used a non-linear equation solver to determine the values as detailed in Supplemental file 1. Calculating ϕ_w empirically in this fashion is the default for `phydms`. If you instead want to fit the ϕ_w values, use the `--fitphi` option.

ExpCM with gamma-distributed nonsynonymous-to-synonymous rate parameter

A common extension to traditional non-site-specific codon substitution models is to allow the dN/dS ratio ω to come from several discrete categories by making the overall likelihood at each site a linear combination of the likelihood computed for each category [89, 160]. Such models are not site-specific since sites are not assigned to a category during likelihood optimization, but they do capture the idea that the strength of selection on nonsynonymous mutations varies across sites.

One variant of this approach draws ω from a discrete gamma distribution. This variant is referred to as the M5 variant [160] in `PAML` [158]. We implemented a similar approach

for an ExpCM, following [156] to draw the ω in Equation 3 from the means of equally weighted gamma-distributed categories (by default there are four categories). This option can be used via the `--gammaomega` switch to `phydms`, and adds one free parameter, since there are two parameters controlling the gamma distribution (a shape and inverse-scale parameter) rather than a single ω. This option increases the runtime by $\sim$5-fold.

Using a gamma-distributed ω may lead to less of an improvement in fit for an ExpCM than for non-site-specific models, since much of the site-to-site variation in the selection is already captured by the amino-acid preferences. However, it can still lead to substantial improvements if a subset of sites are under diversifying selection or if the preferences do not fully capture selection on nonsynonymous mutations.

Traditional YNGKP (or Goldman-Yang) models

To enable comparison of an ExpCM with non-site-specific substitution models, `phydms` implements several of these more traditional models. Here these models are referred to as YNGKP as they are variants of the Goldman-Yang style models described by Yang, Nielsen, Goldman, and Krabbe-Pedersen [160]. Note that sometimes in the literature these models are called GY94 rather than YNGKP; however here we use the name YNGKP to explicitly emphasize that we are using the model variants delineated by [160] rather than the original variants described in [44]. The M0 and M5 YNGKP models are implemented in `phydms`. The M0 variant optimizes a single dN/dS ratio (ω) and so is comparable with the basic ExpCM, while the M5 variant draws ω from a gamma distribution and so is comparable to an ExpCM with the `--gammaomega` option. The equilibrium codon frequencies are calculated empirically after correcting for stop codons as described by [97] (the CF3X4 method). The M0 variant has 11 parameters (9 empirical nucleotide frequencies plus ω and κ), while the M5 variant has 12 parameters (ω is replaced by the two gamma-distribution parameters).

YNGKP models are less computationally expensive than an ExpCM since they are not

site-specific. Therefore, YNGKP models are faster than the ExpCM in `phydms`. However, `phydms` is *not* optimized for maximal speed with YNGKP models, so if you are only using those models then consider using `PAML` [158] or `HyPhy` [98].

3.3.2 Gradient-based optimization of the likelihood

Given one of the substitution models described above and a fixed phylogenetic tree topology, `phydms` numerically optimizes the model parameters and branch lengths to their maximum likelihood values via the Felsenstein pruning algorithm [36]. Numerical optimization generally requires fewer steps if the gradient of the objective function with respect to free parameters is computed explicitly [43], although this advantage can be offset by the cost of computing the gradient. We were unable to find clear published comparisons of the efficiency of phylogenetic optimization with and without an explicit gradient, although [67] describe how the gradient (and Hessian matrix of second derivatives) can be computed.

We chose to use gradient-based optimization for `phydms` under the supposition that it might be more efficient. The first derivatives with respect to branch lengths and virtually all the model parameters can be computed analytically, propagated through the matrix exponentials using the formula provided by [66], and evaluated along the tree by applying the chain rule to the Felsenstein pruning algorithm. For the ExpCM empirical nucleotide frequencies ϕ_w and the gamma-distributed ω, we used the numerical finite-difference method to compute small portions of the derivatives for which we could not derive analytic results. Supplemental file 1 details how `phydms` computes the likelihood and its gradient.

For the optimization, we used the limited-memory BFGS optimizer with bounds [19, 165, 85]. This optimizer uses the gradient, although this can be turned off with the `--nograd` option to `phydms` (doing so is *not* recommended as the accuracy of `phydms` without gradients has not been extensively tested, and the limited-memory BFGS optimizer may not perform well without gradients). Rather than optimizing model parameters and branch lengths simultaneously, `phydms` takes an iterative approach. First the model

parameters are simultaneously optimized along with a single scaling parameter that multiplies all branch lengths. After this optimization has converged, all branch lengths are simultaneously optimized while holding the model parameters constant. This process is repeated until further optimization leads to negligible improvement in the likelihood. Note that simultaneous optimization of all branch lengths appears to be the minority approach in phylogenetics software [18] and is said by [157] to be less efficient than one-at-a-time optimization; however, we found it to work effectively on the trees that we tested. The rationale for iterating between model parameters and branch lengths is that optimization of the former is more costly in terms of the gradient computation. If you simply want to scale branch lengths by a single parameter rather than optimize them, you can use the `--brlen scale` option. In other contexts, scaling but not individually optimizing branch lengths has been shown to reduce runtime with little effect on final model parameters if the initial tree is reasonably accurate [157, 99].

3.3.3 Design and implementation of `phydms`

The `phydms` software is written in `Python`. Most of the numerical computation is performed with `numpy` and `scipy`, and a few parts of the code are written in compiled `C` extensions created via `cython`. The limited-memory BFGS optimizer used by `phydms` is the one provided with `scipy.optimize`. The most computationally costly part of the optimization performed by `phydms` is the matrix-matrix multiplication performed when computing exponentials of the transition matrix, and the second most costly part is the matrix-vector multiplication performed while implementing the Felsenstein pruning algorithm. Both these steps are performed using `BLAS` subroutines called via `scipy`.

In addition to the core `phydms` program, the software is distributed with auxillary programs that make it easy to prepare alignments (`phydms_prepalignment`) and run multiple models for comparison (`phydms_comprehensive`). Importantly, `phydms` currently does *not* infer phylogenetic tree topologies, but rather optimizes branch lengths and model parame-

ters given a topology. The tree topology must therefore be inferred using another program such as `RAxML` [127] with a simpler substitution model.

3.3.4 Visualization of the results with logoplots

It is often instructive to visualize the amino-acid preferences that are used to inform an ExpCM, as these preferences determine the unique properties of the models. In addition, visualization can help understand how the stringency parameter β optimized by `phydms` re-scales the preferences to increase concordance with natural selection. To aid such visualizations, `phydms` comes with an auxiliary program (`phydms_logoplot`) that renders the amino-acid preferences in the form of logoplots via the `weblogo` libraries [23]. The RESULTS section shows example logoplots.

Computer code

The `phydms` software is freely available on GitHub at `https://github.com/jbloomlab/phydms`. Detailed documentation is at `http://jbloomlab.github.io/phydms`. Analyses in this paper used versions of `phydms` ranging from 2.0.0 to 2.0.5.

3.4 Results

3.4.1 Testing `phydms` on two different genes

In the next few subsections, we describe example applications of `phydms` to real-world data sets. Specifically, we use `phydms` to compare deep mutational scanning measurements to natural sequence evolution for two genes: influenza hemagglutinin (HA) and β-lactamase. We choose these genes because there are multiple published deep mutational scanning datasets for each.

Analysis with an ExpCM requires three pieces of input data: the experimentally measured amino-acid preferences, an alignment of naturally occurring gene sequences, and

a phylogenetic tree topology. The tree topology can be inferred from the sequence alignment. But like most other software for codon-based phylogenetic analyses [98, 158], `phydms` is not designed to infer the tree topology. Instead, it provides easy ways to infer the tree topology using `RAxML` [127].

To prepare the required input data, we followed the workflow in Figure 3. The deep mutational scanning experiments on HA [138, 27] directly reported amino-acid preferences. However, the two deep mutational scanning experiments on β-lactamase [39, 132] reported enrichment ratios for each mutation rather than amino-acid preferences. There is a simple relationship between enrichment ratios and amino-acid preferences: the preferences are the enrichment ratios after normalizing the values to sum to one at each site, enabling easy conversion between the two data representations (Figure 3).

We also created codon-level alignments of naturally occurring HA and β-lactamase sequences using `phydms_prepalignment`. The alignments were trimmed to contain only sites for which amino-acid preferences were experimentally measured. Table 1 summarizes basic information about these alignments.

3.4.2 Test if deep mutational scanning is informative about natural selection

A first simple test is whether the deep mutational scanning experiment provides any information that is relevant to natural selection on the gene in question. This can be determined by testing whether an ExpCM that uses the experimental data outperforms a substitution model that is agnostic to the site-specific preferences measured in the experiments.

To perform such a test, we used `phydms_comprehensive` to fit several substitution models to the alignment of HA sequences. This program automatically generates a phylogenetic tree topology from the alignment using `RAxML` [127]. It then fits an ExpCM (in this case informed by the deep mutational scanning data in [27]) as well as several substitution models that do not utilize site-specific experimental information. The analysis was

performed by running the following command on the input data in Supplemental file 2:

```
phydms_comprehensive results/ HA_alignment.fasta HA_Doud_prefs.csv
```

Table 2 lists the four tested substitution models: the ExpCM, an ExpCM with the amino-acid preferences averaged across sites, and the M0 and M5 variants of the standard Goldman-Yang style substitution models [160]. (Because these variants were originally described by Yang, Nielsen, Goldman, and Krabbe-Pedersen, they are referred to here as YNGKP models; note that other literature sometimes uses the alternative acronym GY94.) The ExpCM with averaged preferences is a sensible control because the averaging eliminates any experimental information specific to individual sites in the protein. Because the models have different numbers of free parameters, they are best compared using Akaike Information Criterion (AIC) [100], which compares log likelihoods after correcting for the number of free parameters. Table 2 shows that the ExpCM has a much smaller AIC than the other models (ΔAIC > 2000 for all other models). Therefore, the experimentally measured amino-acid preferences contain information about natural selection on HA, since a substitution model informed by these preferences greatly outperforms models that do not utilize the experimental information.

3.4.3 Re-scale deep mutational scanning data to stringency of natural selection

Even if a deep mutational scanning experiment measures the authentic natural selection on a gene, the stringency of selection in the experiment is not expected to match the stringency of selection in nature. Differences in the stringency of selection can be captured by the ExpCM stringency parameter β. If selection in nature prefers the same amino acids as the selection in lab but with greater stringency, β will be fit to a value > 1. Conversely, if selection in nature does not prefer the lab-favored mutations with as much stringency as the deep mutational scan, β will be fit to a value < 1. Table 2 shows that an ExpCM for HA informed by the experiments in [27] has $\beta = 2.11$, indicating that natural selection favors the experimentally preferred amino acids with higher stringency than selection in the lab.

The effect of this stringency re-scaling of the preferences can be visualized using `phydms_logoplot` as shown in Figure 4. Re-scaling by the optimal stringency parameter of 2.11 exaggerates the selection for experimentally preferred amino acids. Conversely, if the analysis had fit a stringency parameter < 1, this would have flattened the experimental measurements, and when $\beta = 0$ all information from the experiments is lost (Figure 4). Because selection in the lab can probably never be tuned to exactly match that in nature, stringency re-scaling is a valuable method to standardize measurements across experiments.

3.4.4 Compare how well different experiments capture natural selection

The amino-acid preferences for HA and β-lactamase have each been measured by two independent experiments. For each gene, which of these experiments better captures natural selection?

We can address this question by comparing ExpCM's informed by each experiment. For β-lactamase, this means comparing the preferences measured by [132] to those measured by [39]. We did this with `phydms_comprehensive` by running the following command on the input data in Supplemental file 4:

```
phydms_comprehensive results/ betaLactamase_alignment.fasta
betaLactamase_Stiffler_prefs.txt betaLactamase_Firnberg_prefs.txt
```

Table 3 shows that an ExpCM informed by the data of [132] outperform an ExpCM informed by the data of [39], with a ΔAIC of 96.2. Therefore, the former experiment better reflects natural selection on β-lactamase. However, both experiments are informative, as both greatly outperform traditional YNGKP models.

We made a similar comparison of the two deep mutational scans of HA. As summarized in Table 4 (and detailed in Supplemental file 5), the deep mutational scanning of [27] better describes the natural evolution than the experiments of [138] (ΔAIC of 44.2). Again, both experiments are clearly informative, as both greatly outperform the YNGKP

models.

3.4.5 Identify sites of diversifying selection

In some cases, a few sites may evolve differently in nature than expected from the experiments in the lab. For instance, sites under diversifying selection for amino-acid change will experience more nonsynonymous substitutions than expected given the experimentally measured amino-acid preferences. Such sites can be identified by using the `--omegabysite` option to fit a parameter ω_r that gives the relative rate of nonsynonymous to synonymous substitutions after accounting for the experimentally measured preferences for each site r [14]. If the preferences capture all the selection on amino acids, then we expect $\omega_r = 1$. Sites with $\omega_r > 1$ are under diversifying selection for amino-acid change, while sites with $\omega_r < 1$ are under additional purifying selection not measured in the lab.

We tested for diversifying selection in HA by running the following command on the data in Supplemental file 6:

```
phydms HA_alignment.fasta HA_RAxML_tree.newick ExpCM_HA_Doud_prefs.csv
results/ --omegabysite
```

The results are visualized in Figure 5. While most sites are evolving with ω_r not significantly different from one, some sites show evidence of $\omega_r > 1$. As described in [14], these sites may be under diversifying selection due to immune pressure. Overall, this analysis shows how `phydms` can identify sites evolving differently in nature than expected from experiments in the lab.

3.4.6 `phydms` *has superior computational performance to existing alternatives*

Our rationale for developing `phydms` was to enable the analyses described above to be performed more easily than with existing software. To validate the improved computational performance, we compared `phydms` (version 2.0.0) to alternative programs that have been

used to fit an ExpCM. The comparisons used the HA sequences described in Table 1 with an ExpCM informed by the deep mutational scanning in [27], and were performed on a single core of a 2.6 GHz Intel Xeon CPU.

Table 5 shows the results. With default settings, `phydms` took 10 minutes to optimize the model parameters and branch lengths. This runtime could be decreased by scaling the branch lengths by a single parameter rather than optimizing them individually (`--brlen scale` option); other work has shown that when the initial tree is reasonably accurate, this approximation can improve runtime while only slightly affecting model fit [157, 99]. Fitting the nucleotide frequency parameters ϕ_w (`--fitphi` option) rather than determining them empirically doubled the runtime. The log likelihood and values of the model parameters β and ω were nearly identical for all three of these settings. The gradient-based optimization is important: using `phydms` without gradients (`--nograd` option) increased the runtime over 5-fold while also yielding a poorer log likelihood.

Two alternative programs have previously been used to fit an ExpCM. Bloom (2014a) [11] and Bloom (2014b) [12] used a `Python` program (`phyloExpCM`) to run `HyPhy` to optimize an ExpCM similar to the ones used here. Bloom (2017)[14] used an old version of `phydms` to fit an ExpCM identical to the ones here using the `Bio++` libraries [48]. We ran both these programs on the HA data set, using `phyloExpCM` version 0.3 with `HyPhy` version 2.22, and `phydms` version 1.3.0 with `Bio++`. Table 5 shows that these programs were $\sim$100-fold and $\sim$200-fold slower than `phydms` with default settings. A small portion of the slower runtime is because these earlier implementations cannot calculate empirical nucleotide frequency ϕ_w parameters; however they remain much slower than `phydms` even when these parameters are fit. Note that Table 5 may overestimate the computational advantage of `phydms` over `HyPhy` in some situations, since `HyPhy` code but not `phydms` can in principle be written to enable the use of multiple cores. Divining the reasons for the performance differences was not possible, as the programs differ completely in their implementations. But reassuringly, all programs yielded similar model parameters β and ω despite independent implementations of the likelihood calculations and the optimization.

The analyses above used relatively small alignments of 34 or 50 sequences (Table 1). To test how the performance of `phydms` changed with alignment size, we analyzed HA alignments ranging from 34 to 108 sequences. As shown in Table 6, the runtime increased with alignment size, but remained under an hour even for the largest alignment. The inferred model parameter values also remained relatively constant as the size of the HA alignment increased (Table 6).

3.5 Discussion

We have described a new software package that facilitates efficient analyses with phylogenetic substitution models informed by deep mutational scanning experiments. This software, `phydms`, can quantitatively compare deep mutational scanning measurements to selection on genes in nature. It can re-scale deep mutational scanning data to account for differences in the stringency of selection between the lab and nature, identify sites evolving differently in nature than expected from the experiments, and compare how well different experiments on the same gene describe natural selection.

The ability to perform these comparisons is useful because the rationale for many deep mutational scanning experiments is to provide information about the effects of mutations on genes in nature. For instance, there are many ways to design an experiment, and it is often not obvious which choice is best if the goal is to make the experiment reflect natural selection. Using `phydms`, it is possible to quantitatively compare how well different experiments describe natural selection. Likewise, it is often useful to know if specific sites in a gene are evolving differently in nature than expected from experiments in the lab. Algorithms implemented in `phydms` makes statistically rigorous identification of these sites possible.

The speed and ease of use of `phydms` makes these analyses practical for real datasets. As deep mutational scanning data become available for an increasing number of genes, `phydms` will facilitate comparison of the experimental measurements to selection in nature.

3.6 Acknowledgments

We thank Erick Matsen, Vladimir Minin, and Joe Felsenstein for helpful comments that aided in the planning and design of the software. We thank Hugh Haddox for assistance in testing the software.

Supplemental files

Descriptions of each file are shown below. Please see the publication for the actual files [59].

S1 File. This PDF contains details of the calculations of the likelihood and its gradient as implemented in `phydms`.

S2 File. This ZIP file contains the code, input data, and full results of the `phydms` analysis summarized in Table 2.

S3 File. This ZIP file contains the code, input data, and full results of the stringency parameter comparison with `phydms_logoplot` summarized in Figure 4.

S4 File. This ZIP file contains the code, input data, and full results of the multiple β-lactamase deep mutational scan comparison summarized in Table 3.

S5 File. This ZIP file contains the code, input data, and full results of the multiple HA deep mutational scan comparison summarized in Table 4.

S6 File. This ZIP file contains the code, input data, and full results of the `phydms --omegabysite` analysis summarized in Figure 5.

S7 File. This ZIP file contains the code, input data, and full results of the program runtime comparison summarized in Table 5.

S8 File. This ZIP file contains the code, input data, and full results of the alignment size comparison summarized in Table 6.

3.7 Software Availability

All data and code are available on GitHub at `https://github.com/jbloomlab/phydms`. Detailed documentation is at `http://jbloomlab.github.io/phydms`.

Chapter 4

MODELING SITE-SPECIFIC AMINO-ACID PREFERENCES DEEPENS PHYLOGENETIC ESTIMATES OF VIRAL SEQUENCE DIVERGENCE

A version of this chapter has been previously published as:

Sarah K. Hilton and Jesse D. Bloom. “Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence.” Virus Evolution. 4:vey033 (2018).

4.1 Abstract

Molecular phylogenetics is often used to estimate the time since the divergence of modern gene sequences. For highly diverged sequences, such phylogenetic techniques sometimes estimate surprisingly recent divergence times. In the case of viruses, independent evidence indicates that the estimates of deep divergence times from molecular phylogenetics are sometimes too recent. This discrepancy is caused in part by inadequate models of purifying selection leading to branch-length underestimation. Here we examine the effect on branch-length estimation of using models that incorporate experimental measurements of purifying selection. We find that models informed by experimentally measured site-specific amino-acid preferences estimate longer deep branches on phylogenies of influenza virus hemagglutinin. This lengthening of branches is due to more realistic stationary states of the models, and is mostly independent of the branch-length-extension from modeling site-to-site variation in amino-acid 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 all of these site-specific but time-homogeneous and site-independent models are limited by the fact that a protein's amino-acid preferences gradually shift as it evolves. Overall, our work underscores the importance of modeling site-specific amino-acid preferences when estimating deep divergence times—but also shows the inherent limitations of approaches that fail to account for how these preferences shift over time.

4.2 Introduction

Molecular phylogenetics is commonly used to estimate the historical timing of evolutionary events [161]. This is done by estimating branch lengths based on the inferred number of substitutions, and then converting these branch lengths into units of time under the assumption of a molecular clock [166, 28]. However, phylogenetic estimates of the divergence times of many viral lineages are clearly too recent [31, 60, 1]. For example, the

integration of filoviruses into their host genomes indicate that Ebola and Marburg virus diverged from their common ancestor 7 to 12 million years ago—but the estimate of this divergence time based on phylogenetic analyses of the viral sequences is only $\sim$10,000 years ago [20, 137]. Similarly, the phylogenetic estimate of when major simian immunodeficiency virus groups diverged is almost 100 times more recent than the estimate based on the geographic isolation of their host species [146, 150]. These examples, along with other similar discrepancies with measles virus [41], coronavirus [144], and hepatitis B virus [34, 61], indicate that phylogenetic methods have a systematic bias toward underestimation of deep branches.

This underestimation occurs in part because phylogenetic models do a poor job of describing the real natural selection on protein-coding genes. These genes evolve under purifying selection to maintain the structure and function of the proteins they encode. In general, these constraints are highly idiosyncratic among sites [32]. However, most phylogenetic models try to account for these constraints using relatively simple approaches such as allowing the rate of substitution to vary across sites according to some statistical distribution [156, 160]. These models of purifying selection are usually inadequate [30, 29], potentially causing branch lengths to be severely underestimated [145, 52].

More recent work has used mutation-selection models to better account for purifying selection [52, 159, 110, 135, 80]. These models explicitly incorporate the fact that different protein sites prefer different amino acids, and so can improve phylogenetic estimates when there are deep branches [95, 71, 73, 102, 141, 134]. However, these approaches require inferring the site-specific purifying selection from natural sequence data.

Even more recently, it has become possible to directly measure purifying selection on proteins using deep mutational scanning. This high-throughput approach involves experimentally measuring how each amino-acid mutation affects protein function in the lab [40]. The resulting experimental measurements of which amino acids are preferred at each protein site can be used to inform phylogenetic substitution models [11]. These experimentally informed codon models (ExpCMs) generally exhibit much better phylogenetic fit

than standard substitution models [26, 59, 51, 75].

Here we examine how ExpCMs and other models of purifying selection estimate branch lengths on a phylogenetic tree of influenza virus hemagglutinin (HA). We find that ExpCMs estimate longer deep branches, and show that this extension of branch length is mostly independent and additive with that achieved by the more conventional approach of modeling rate variation. We also show that ExpCMs estimate similar branch lengths to a mutation-selection model that infers the amino-acid preferences from the natural sequence data rather than using values obtained in experiments. However, all of these mutation-selection models are limited by their failure to account for another feature of purifying selection: the fact that a site's amino-acid preferences shift over time due to epistasis. Therefore, truly accurate analyses of deep phylogenies need to account for the fact that amino-acid preferences vary across time as well as across sites.

4.3 Results

4.3.1 Different ways substitution models account for purifying selection

Here we consider how purifying selection is handled by codon models, which are the most sophisticated of the three classes (nucleotide, codon, and amino acid) of phylogenetic substitution models in widespread use for protein-coding genes [5]. Standard codon models distinguish between two types of substitutions: synonymous and nonsynonymous. The relative rate of these substitutions is referred to as dN/dS or ω. In their simplest form, codon substitution models fit a single ω that represents the gene-wide average fixation rate of nonsynonymous mutations relative to synonymous ones. Here we will use such substitution models in the form proposed by Goldman (1994) [44]. When these models have a single gene-wide ω they are classified as M0 by Yang *et al.*, 2000 [160]. We will refer to M0 Goldman-Yang models simply as GY94 models (Equation 1). The gene-wide ω is usually < 1 [86], and crudely represents the fact that many amino-acid substitutions are under purifying selection.

A single gene-wide ω ignores the fact that purifying selection is heterogeneous across sites. The most common strategy to ameliorate this defect is to allow ω to vary among sites according to some statistical distribution [156, 160]. For instance, in the M5 variant of the GY94 model [160], ω follows a gamma distribution as shown in Figure 6A. We will denote this model as GY94+$\Gamma\omega$. A GY94+$\Gamma\omega$ captures the fact that the rate of nonsynonymous substitution can vary across sites. However, these models do not capture the fact that the same amino-acid mutation can have very different effects at different sites.

Mutation-selection models account for the fact that purifying selection depends idiosyncratically on the specific amino-acid mutation at each site [52, 159, 110, 135, 80]. Here we will consider mutation-selection models where the site-specific selection is assumed to act solely at the protein level (different codons for the same amino acid are treated as selectively equivalent). Such models 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 6B). These models capture the site-to-site variation in amino-acid composition that is an obvious features of real proteins, and usually better describe actual evolution than models with only rate variation as assessed by Bayesian or maximum-likelihood criteria [72, 73, 102, 141, 110, 11, 12, 59].

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 ω, a transition-transversion rate, and nine parameters describing the nucleotide composition of the stationary state. However, mutation-selection models must additionally specify 19 parameters defining the amino-acid preferences for *each* site (there are 20 amino acids whose preferences are constrained to sum to one). This corresponds to $19 \times L$ parameters for a protein of length L, or 9,500 parameters for a 500-residue protein. It is challenging to obtain values for these amino-acid preference parameters in a maximum-likelihood

framework without overfitting the data [105]. Here we will primarily use experimentally informed codon models (ExpCMs), which define the site-specific amino-acid preference parameters *a priori* from deep mutational scanning experiments so that they do not need to be fit from phylogenetic data [see Methods and 11, 59, 14]. Because the amino-acid preference parameters in an ExpCM are obtained from experiments, the number of ExpCM free parameters is similar to a non-site-specific substitution model. An alternative strategy to account for site-specific amino-acid preferences is to formally consider them as random effects across sites, rather than parameters, and infer them using Bayesian methods [72, 107]. This strategy is 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 an ExpCM can still incorporate an ω parameter, which now represents the relative rate of nonsynonymous to synonymous substitution *after* accounting for the constraints due to the site-specific amino-acid preferences [14, 108]. This ω parameter for an ExpCM can be drawn from a statistical distribution (e.g., a gamma distribution) just like for GY94-style models [107, 51]. We will denote such models as ExpCM+$\Gamma\omega$. Figure 6C shows the full spectrum of models that incorporate all combinations of gamma-distributed ω and site-specific stationary states.

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

Given a single branch, a substitution model transforms sequence divergence into branch length. Under a molecular-clock assumption, this branch length is proportional to time. The transformation from sequence divergence to branch length is trivial when the sequence identity is high. For instance, when 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 [166] will correctly infer the short branch length of $1/L$ substitutions per site. However, as substitutions accumulate it becomes 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 sequence divergence into branch length. Any time-homogenous substitution model predicts that after a very large number of substitutions, two related sequences will approach some asymptotic amino-acid sequence identity. For instance, if all 20 amino acids are equally likely in the stationary state, then this asymptotic sequence identity will be $\frac{1}{20} = 0.05$. If the substitution model underestimates the asymptotic sequence identity 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 7 shows how different substitution models predict amino-acid sequence identity to decrease as a function of branch length using 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, with an asymptotic sequence identity of 0.062. While this predicted sequence identity is higher than $\frac{1}{20} = 0.05$ due to redundant codon and nucleotide biases favoring certain amino acids, it is much lower than the pairwise identity of even the most diverged HAs in nature. While it is of course possible that the identity of HAs in nature would become even lower given more time, it seems biochemically improbable that it would ever become as low as 0.062. The reason is that like many proteins HA has a highly conserved structure and function that imposes constraints that cause many sites to sample only a small subset of the 20 amino acids among all known HA homologs [90].

Accounting for site-to-site dN/dS 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 7 shows that the GY94+$\Gamma\omega$ model takes longer to reach the asymptote than GY94, but that the asymptote is identical for both models. This fact holds true even if we use experimental measurements of HA's site-specific amino-acid preferences [27] to calculate a different ω_r value for each site using the method of [124] (see Equation 7). Specifically, this GY94+ω_r model predicts that different sites will approach the asymptote at different rates, but the asymptote is always the same (Figure 7). The

invariance of the asymptotic sequence identity under different schemes for modeling ω is a fundamental feature of the mathematics of this type of reversible substitution model. These models are reversible stochastic matrices, which can be decomposed into stationary states and symmetric exchangeability matrices [88]. The stationary state is invariant with respect to multiplication of the symmetric exchangeability matrix by any non-zero number. Different schemes for modeling ω only multiply elements of the symmetric exchangeability matrix. Therefore, no matter how "well" a model accounts for site-to-site variation in ω, it will always have the same stationary state as a simple GY94 model.

However, mutation-selection models such as ExpCMs have site-specific stationary states. They predict that different sites will have different asymptotic sequence identities (Figure 7)—a prediction that accords with the empirical observation that some sites are much more variable than others in alignments of highly diverged sequences. For instance, Figure 7 shows that at sites such as 183 and 305 in the H1 HA, an ExpCM but not a GY94-style model predicts that the identity will always be relatively high. When sites with highly constrained amino-acid preferences such as these are common, an ExpCM can estimate a long branch length at modest sequence identities that a GY94 model might attribute to a shorter branch.

4.3.3 Simulations demonstrate how failure to model site-specific amino-acid preferences 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 (Figure 8). Specifically, we simulated alignments of sequences along the HA phylogenetic tree using an ExpCM parameterized by the amino-acid preferences of H1 HA as experimentally measured by deep mutational scanning [27]. We then estimated the branch lengths from the simulated sequences using all the substitution models in Figure 6C, and compared these estimates to the actual branch lengths

used in the simulations. Note that these simulations closely parallel those performed by [52] and [145].

The models with a uniform stationary state underestimated the lengths of long branches on the phylogenetic tree of the simulated sequences (Figure 8). The GY94 model estimated branch lengths that are $\sim$60% of the true values for the longest branches. Accounting for site-to-site variation in ω did not fix the fundamental problem: the GY94+$\Gamma\omega$ did slightly better, but still substantially underestimated the longest branches. However, there was no systematic underestimation of long branches by the ExpCM and ExpCM+$\Gamma\omega$ models. The improved performance of the ExpCMs is due to their modeling of the site-specific amino-acid preferences: if we parameterize ExpCMs by amino-acid preferences that have been averaged across HA sites (and so are no longer site-specific), then they perform no better than GY94 models (Figure 8). Therefore, models with uniform stationary states underestimate the length of long branches in phylogenies of sequences that have evolved under strong site-specific amino-acid preferences.

4.3.4 Experimentally informed site-specific models estimate longer branches on real data.

The foregoing section shows the superiority of ExpCMs to GY94 models for estimating long branches on phylogenies simulated with ExpCMs. But how do these models perform on real data? Real genes do evolve under functional constraint, but these constraints are almost certainly more complex than what is modeled by an ExpCM. However, if ExpCMs do a substantially better job than GY94 models of capturing the true constraints, then 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 $>$68% amino-acid identity, but sequences in different subtypes have as little

as 38% identity. However, HA proteins from all subtypes have a highly conserved structure that performs a highly conserved function [49, 114]. We used `RAxML` [126] with a nucleotide substitution model (GTRCAT) to infer a phylogenetic tree for 92 HA sequences drawn from 15 of the 18 subtypes (we excluded bat influenza and one other rare subtype). For the rest of this paper, we fix the tree topology to this `RAxML`-inferred tree. Although the nucleotide model used with `RAxML` to infer this tree topology is probably less accurate than codon models, the modular subtype structure of the HA phylogeny (the tree is clearly divided into widely separated clades) means that most of the phylogenetic uncertainty lies in the length of the long branches separating the subtypes rather than in the tree topology itself. We note that other genes may have phylogenetic structures that are more prone to topological uncertainty. In such cases, the accuracy of the substitution model may be important for avoiding topological errors such as long-branch attraction [37, 71].

Deep mutational scanning has been used to measure the amino-acid preferences of all sites in two different HAs. One scan measured the preferences of an H1 HA [27] and the other measured the preferences of an H3 HA [75]. The amino-acid preferences measured for these two HAs are shown in 13 and 14. The H1 and H3 HAs have only $\sim$42% amino-acid identity. As described in Lee *et al.*, 2018 [75], the amino-acid preferences clearly differ between the H1 and H3 HA at a substantial number of sites (these differences are apparent in a simple visual comparison of 13 and 14; see site 33 as an example). Therefore, we also created a third set of amino-acid preferences by averaging the measurements for the H1 and H3 HAs, under the conjecture that these averaged preferences might better describe the "average" constraint on sites across the full HA tree (15). These three sets of HA amino-acid preferences define three different ExpCMs.

We fit the GY94 model and each of the three ExpCMs to the fixed HA tree topology estimated using `RAxML`, and also tested a version of each model with $\Gamma\omega$ rate variation. Table 7 shows that all ExpCMs fit the actual data much better than the GY94 models. The best fit was for the ExpCM informed by the average of the H1 and H3 deep mutational scans. For all models, incorporating $\Gamma\omega$ rate variation improved the fit, although even

ExpCMs without $\Gamma\omega$ greatly outperformed the GY94+$\Gamma\omega$ model (Table 7). As mentioned in the previous section, ω is generally < 1 when a single value is fit to all sites in a gene [86], and this is the case for all the models we tested (Table 7). However, the ExpCMs always fit an ω greater than the GY94 model, suggesting that the site-specific amino-acid preferences capture some of the purifying selection that the GY94 models can represent only via a small ω. Among the models with $\Gamma\omega$, the GY94+$\Gamma\omega$ model fits all four ω categories to values $\ll$1, but the ExpCM+$\Gamma\omega$ models fit one of the ω categories to a value > 1. This increase in ω values makes sense given the different interpretation of ω for each family of models. The ExpCM ω is the relative rate of fixation of nonsynonymous to synonymous mutations *after* accounting for the functional constraints described by the amino-acid preferences. This more realistic null model gives ExpCMs enhanced power to detect diversifying selection for amino-acid change [14, 108], which is known to occur at some sites in HA due to immune selection [9].

Importantly, models that account for purifying selection via either $\Gamma\omega$ rate variation or site-specific amino-acid preferences do not just exhibit better fit—they also estimate longer branches on the HA tree. Figure 9 shows the branch lengths optimized by each model on a common scale. The tree's deepest branches are shortest when they are optimized by the GY94 model, which lacks both $\Gamma\omega$ and site-specific amino-acid preferences. Adding either $\Gamma\omega$ rate variation or site-specific amino-acid preferences increases the length of the deep branches. Specifically, the tree's diameter (the distance between the two most divergent tips) for the GY94+$\Gamma\omega$ model is 159% of the GY94 model tree diameter (8). The tree diameter is 122% and 135% of the GY94 model tree diameter for ExpCMs informed by H1 or H3 amino-acid preferences, respectively, and 160% of the GY94 model for the ExpCM informed by the average of the H1 and H3 preferences (8).

The deepening of branch lengths that results from the $\Gamma\omega$ and site-specific amino-acid preference approaches to modeling purifying selection are largely independent. This can be seen by examining the ExpCM+$\Gamma\omega$ models, which combine $\Gamma\omega$ rate variation with site-specific amino-acid preferences. As shown in Figure 9, these ExpCM+$\Gamma\omega$ models

estimate longer branches than models with just $\Gamma\omega$ rate variation (GY94+$\Gamma\omega$) or just site-specific amino-acid preferences (ExpCMs). The near independence of these effects is quantified in 8, which shows that 76% of the tree diameter extension of ExpCM(H1+H3 avg)+$\Gamma\omega$ versus can be explained by simply adding the extension from incorporating $\Gamma\omega$ (GY94+$\Gamma\omega$ versus GY94) to the extension from incorporating site-specific amino-acid preferences (compare ExpCM(H1+H3 avg) to GY94).

However, while adding $\Gamma\omega$ rate variation increases the length of deep branches in a roughly uniform fashion across the tree, the branch lengthening from adding site-specific amino-acid preferences is not uniform across the tree (Figure 9 and Figure 10). Instead, the increase in branch length is most pronounced on branches leading to the HA sequence that was used in the deep mutational scanning experiment that informed the ExpCM. For instance, the ExpCM informed by the H1 data most dramatically lengthens branches near the H1 clade of the tree, while the ExpCM informed by the H3 data has the largest effect on branches near the H3 clade. The ExpCM informed by the average of the H1 and H3 data has a more uniform effect across the tree, but still most strongly extends branches leading to either the H1 or H3 clade. Therefore, Figure 9 and Figure 10 show that ExpCMs estimate longer branches, but that the effect is shaped by the set of amino-acid preferences used to inform the model.

4.3.5 Shifting amino-acid preferences limit the benefits of models with site-specific stationary states for estimating long branch lengths.

The fact that an ExpCM leads to the most profound increase in branch length leading to the sequence used in the experiment can be rationalized in terms of existing knowledge about epistasis during protein evolution. Each ExpCM is informed by a single set of experimentally measured amino-acid preferences. But in reality, the effect of a mutation at one site in a protein can depend on the amino-acid identities of other sites in the protein [93, 46, 54, 139, 130]. This epistasis can lead to shifts in a protein's amino-acid

preferences over evolutionary time [96, 26, 118, 8, 51]. Because the deep mutational scanning experiments that inform our ExpCMs were each performed in the context of a single HA genetic background, their measurements do not account for the accumulation of epistatic shifts in amino-acid preferences as HA evolves. Therefore, an ExpCM is expected to most accurately describe the evolution of sequences closely related to the one used in the experiment.

We can observe how shifting amino-acid preferences degrade the accuracy of an ExpCM by fitting the model to trees containing increasingly diverged sequences. For both H1 and H3 HAs, we created three phylogenetic trees (17): a "low" divergence tree that contains sequences with $\geq$59% amino-acid identity to the HA used in the experiment, an "intermediate" divergence tree that contains sequences with $\geq$46% amino-acid identity to the HA in the experiment, and a "high" divergence tree that contains all HAs (which have as little as 38% identity to the HA in the experiment). Figure 11 shows the subtrees containing each of these sets of HA sequences. For each subtree, we examined the congruence between site-specific natural selection and the amino-acid preferences measured in the deep mutational scanning experiment using the ExpCM stringency parameter β [12, 59]. Values of β that are $>$1 indicate that natural selection prefers the same amino acids as the experiments but with a greater stringency, suggesting strong congruence between natural selection and the experimental preferences. In contrast, values of β that are $<$1 flatten the preferences, suggesting that they provide a relatively poor description of natural selection on the protein.

The value of β decreases as the divergence from the sequence used in the deep mutational scan increases Figure 11. This inverse relationship between β and overall divergence is seen for the ExpCMs informed by both the H1 and H3 experiments. As β decreases, the preferences "flatten" and so the ExpCM draws less information from the experiment. At the most extreme value of $\beta = 0$, the preferences would be perfectly uniform and look similar to the GY94 preferences in Figure 9. In reality, β never reaches a value this low, indicating the deep mutational scanning experiments remain somewhat in-

formative about real natural selection across the entire swath of HAs. However, Figure 11 shows that the amino-acid preferences clearly become less informative about natural selection as we move away from the experimental sequence on the tree. This shifting of amino-acid preferences helps explain why the ExpCM informed by the average of the H1 and H3 experiments performs best (Table 7, Figure 9, and Figure 10): averaging the measurements across these two HAs is a heuristic method of accounting for shifts in preferences during HA evolution.

The fact that amino-acid preferences shift as a protein evolves leaves us with an inherent tension: models with site-specific amino-acid preferences only become important for accurate branch-length estimation as sequences become increasingly diverged, but this same divergence degrades the accuracy of extrapolating the amino-acid preferences from any given experiment across the phylogenetic tree. Crucially, this problem is more fundamental than the inability of a single deep mutational scanning experiment to measure amino-acid preferences in more than one genetic background. If amino-acid preferences shift during evolution, there simply will not be any single set of time-homogeneous site-specific preferences that accurately describes evolution along the entirety of a phylogenetic tree that covers a wide span of sequences.

4.3.6 A model with amino-acid preferences estimated from natural sequences gives similar results to an ExpCM

The previous sections used ExpCMs, which are mutation-selection models that use site-specific amino-acid preferences that have been measured by experiments. However, there are other mathematically similar implementations of mutation-selection models that infer the amino-acid preferences directly from the natural sequence data. When these models are designed for use in phylogenetic inference, they are generally implemented in a Bayesian framework, which avoids the overfitting problems associated with trying to make maximum-likelihood estimates of the thousands of amino-acid preference parame-

ters [70]. (Note that the maximum-likelihood implementations of [135, 136] are designed for estimating the amino-acid preferences, *not* for phylogenetic inference.) The model most comparable to our ExpCMs is the codon mutation-selection model implemented in `PhyloBayes-MPI`, which we will refer to as pbMutSel [107]. In the pbMutSel model, the amino-acid preferences are modeled using Dirichlet processes rather than derived from experiments. However, like an ExpCM, a pbMutSel model still assumes a single set of time-homogeneous site-specific amino-acid preferences for the entire tree.

Comparing ExpCM and pbMutSel models can help determine the ultimate limits of mutation-selection models that assign each site a single set of amino-acid preferences. If the limitations of ExpCMs described above arise simply because the deep mutational scanning experiments do not correctly measure the "true" amino-acid preferences across the entirety of a highly diverged phylogenetic tree, then we would expect the pbMutSel models (which infer these preferences from the entire tree) to perform better. On the other hand, if the major limitation is that no single set of time-homogenous amino-acid preferences can fully describe evolution over the entire tree, then we would expect ExpCM and pbMutSel models to perform similarly.

We fit a pbMutSel model to the entire HA phylogenetic tree, and compared the results to those from analyzing the same tree with the best ExpCM, which is the ExpCM(H1+H3 avg)+$\Gamma\omega$ variant. This is a direct apples-to-apples comparison, since the pbMutSel model also draws ω from a gamma-distribution [107]. First, we compared the amino-acid preferences inferred by the pbMutSel model to the preferences measured in the experiments. Figure 12A shows that the preferences inferred by pbMutSel are quite similar to the (H1+ H3 avg) obtained by averaging the deep mutational scanning measurements for the H1 and H3 HAs. Notably, the amino-acid preferences from the pbMutSel model are more correlated with the (H1+ H3 avg) than the H1 and H3 measurements are with each other (Figure 12A). This strong correlation indicates that the ExpCM(H1+H3 avg)+$\Gamma\omega$ is unlikely to be much different than a pbMutSel model that is parameterized only using the natural sequence data.

We next compared the branch lengths estimated by using the ExpCM(H1+H3 avg)+$\Gamma\omega$ and pbMutSel models. As shown in Figure 12B, these two models estimated similar branch lengths across the entire HA phylogenetic tree. However, the estimates are not identical, and the tension between local and global accuracy of the amino-acid preferences is still apparent. Specifically, the branches leading to the H1 or H3 sequences used in the experiments were estimated to be slightly longer by the ExpCM, while some other branches were estimated to be slightly longer by the pbMutSel model. The relatively longer branches leading to the experimental sequences when using the ExpCM(H1+H3 avg)+$\Gamma\omega$ suggests that the "tree average" amino-acid preferences inferred by the pbMutSel model are not as accurate as the preferences from the deep mutational scanning for sequences close to those used in the experiments. However, for sequences distant from those used in the experiments, the "tree average" preferences inferred by the pbMutSel model appear to be slightly better than the experimental values. Therefore, while the ExpCM and pbMutSel models differ slightly in the extent to which they lengthen different branches, neither model can avoid the tension between the local and global accuracy of amino-acid preferences.

4.4 Discussion

We examined how estimates of deep branch lengths on phylogenetic trees are affected by accounting for the fact that proteins prefer specific amino acids at specific sites. We did this by comparing inferences from models informed by experimental measurements of site-specific amino-acid preferences with more conventional codon substitution models, as well as with models that infer the amino-acid preferences from the natural sequences. We found that models that account for site-specific amino-acid preferences estimated deeper long branches, regardless of whether these preferences are measured experimentally or inferred from the sequence alignment. Additionally, we showed that the extension in branch length from site-specific amino-acid preferences is mostly independent of

the extension that results from simply modeling rate variation.

Overall, our results underscore the importance of modeling purifying selection in a way that is more nuanced than simply allowing the substitution rate to vary across sites. Protein sites do not simply differ in their rates of substitution—different sites also prefer different amino acids. There are now two ways to account for this fact: using models informed by deep mutational scanning experiments, or using models that infer site-specific amino-acid preferences from the natural sequence alignment. Combining either type of model with rate variation increases the inferred length of deep branches relative to models that only incorporate rate variation. We expect that further improvements could be achieved by also incorporating other factors such as host-specific substitution rates [149] that are known to be important for modeling the evolution of viral genes such as HA.

However, assuming a single set of site-specific amino-acid preferences is still an imperfect way to model evolution over a highly diverged phylogenetic tree. In the case of the experimentally informed models, it is fairly obvious why this is true: the amino-acid preferences are measured in just one genetic background, and therefore provide only a single snapshot of preferences that shift over evolutionary time due to epistasis [96, 118, 8, 51, 130, 26]. As a result, experimentally measured amino-acid preferences are most accurate for sequences similar to the one used in the experiment, and so cause the largest increases in branch length in that region of the phylogenetic tree. However, this limitation is not unique to experimentally informed models, but is a general limitation of describing purifying selection using a single set of site-independent and time-homogenous amino-acid preferences. For instance, we showed that averaging experimental measurements on two protein homologs does a somewhat better job of capturing the “average” constraint across the tree, and performs similarly to approaches that infer the “average” preferences from natural sequence data [110, 107]. But even these “average” preferences exhibit a tradeoff between local and global accuracy for the inference of deep branch lengths.

So while modeling site-specific amino-acid preferences is a clear improvement over

most conventional models, the next step towards greater accuracy will require relaxing the assumption that these preferences are time homogeneous and site independent. Of course, many authors have pointed out the shortcomings of models that fail to account for the full site-interdependent complexity of purifying selection [109, 22, 96, 45]. However, the challenge is to overcome these shortcomings with models that are tractable for real phylogenetic questions. There are two main issues: first, the Felsenstein pruning algorithm [36] that is typically used to evaluate phylogenetic likelihoods breaks down when sites are no longer treated independently. Some alternative algorithms have been proposed [16, 106, 109, 22], but they are still in their infancy. Second, site-interdependent models require a realistic "fitness function" that describes the interactions among sites. It appears that typical structural modeling programs are insufficient for this purpose [106]. But hope comes from experimental progress in measuring actual site-interdependent constraints on proteins [92, 152, 131, 76], combined with new methods for using these measurements to parameterize fitness functions [116, 94]. Perhaps some day such truly realistic models might be useful for phylogenetic inference. Until that day, our work shows that modeling a single set of time homogenous amino-acid preferences provides at least some improvement.

4.5 Methods

4.5.1 Substitution models

All of the substitution models used in this paper have been described previously. However, here we briefly recap their exact mathematical implementations.

GY94 model

The GY94 model is M0 variant of the Goldman-Yang model described by [160]. Specifically, the substitution rate P_{xy} from codon x to codon y is

$$P_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ by more than one nucleotide,} \\ \Phi_y & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transversion,} \\ \omega\Phi_y & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transversion,} \\ \kappa\Phi_y & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transition,} \\ \omega\kappa\Phi_y & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y) \text{ and } x \text{ is converted to } y \text{ by a single-nucleotide transition,} \\ -\sum_{z \neq x} P_{xz} & \text{if } x = y, \end{cases}$$

(Equation 1)

where $\mathcal{A}(x)$ is the amino-acid encoded by codon x, κ is the transition-transversion rate, Φ_y is the equilibrium frequency of codon y, and ω is the relative rate of nonsynonymous and synonymous substitutions. We define the codon frequency parameters, Φ_y, using the "corrected F3X4" method from [97]. There are nine parameters describing the nucleotide frequencies at each codon site (the nucleotides are constrained to sum to one at each codon position), and these parameter values are calculated from the empirical alignment frequencies. The "corrected F3X4" method calculates the Φ_y values from these nucleotide frequencies but corrects for the exclusion of sequences with premature stop codons from the analysis.

The frequency p_x of codon x in the stationary state of a GY94 model is simply

$$p_x = \Phi_x. \qquad \text{(Equation 2)}$$

Overall, a GY94 model has 11 free parameters: κ, ω, and the 9 nucleotide frequency parameters used to define Φ_y.

Experimentally Informed Codon Model (ExpCM)

The ExpCM models used in this paper are the ones described in [14]. Briefly, the rate of substitution $P_{r,xy}$ of site r from codon x to y is

$$P_{r,xy} = Q_{xy} \times F_{r,xy} \qquad \text{(Equation 3)}$$

where Q_{xy} is proportional to the rate of mutation from x to y, $F_{r,xy}$ is proportional to the probability that this mutation fixes, and the diagonal elements P_{xx} are set by $P_{xx} = -\sum_{z \neq x} P_{xz}$.

The rate of mutation Q_{xy} is assumed to be uniform across sites, and takes an HKY85-like [55] form as

$$Q_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ by more than one nucleotide,} \\ \phi_w & \text{if } x \text{ can be converted to } y \text{ by a transversion of a nucleotide to } w, \\ \kappa \times \phi_w & \text{if } x \text{ can be converted to } y \text{ by a transition of a nucleotide to } w \end{cases} \qquad \text{(Equation 4)}$$

where ϕ_w is the nucleotide frequency of nucleotide w and κ is the transition-transversion rate.

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

$$F_{r,xy} = \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y), \\ \omega \times \frac{\ln\left[\left(\pi_{r,\mathcal{A}(y)}/\pi_{r,\mathcal{A}(x)}\right)^{\beta}\right]}{1-\left(\pi_{r,\mathcal{A}(x)}/\pi_{r,\mathcal{A}(y)}\right)^{\beta}} & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y), \end{cases} \qquad \text{(Equation 5)}$$

where β is the stringency parameter [12, 59] and ω is the relative rate of nonsynonymous to synonymous substitutions after accounting for the amino-acid preferences.

The stationary state of an ExpCM is

$$p_{r,x} = \frac{\phi_{x_1}\phi_{x_2}\phi_{x_3}\left(\pi_{r,A(x)}\right)^{\beta}}{\sum_z \phi_{z_1}\phi_{z_2}\phi_{z_3}\left(\pi_{r,A(z)}\right)^{\beta}} \qquad \text{(Equation 6)}$$

where ϕ_{x_1}, ϕ_{x_2}, and ϕ_{x_3} are the nucleotides at position 1, 2, and 3 of codon x.

An ExpCM has five free parameters: κ, ω, and the three independent ϕ_x values. The amino-acid preferences $\pi_{r,a}$ are *not* free parameters since they are determined *a priori* by an experiment independent of the sequence alignment being analyzed.

$\Gamma\omega$ *rate variation*

The GY94+$\Gamma\omega$ is equivalent to the M5 model in [160] with ω drawn from $K = 4$ categories. The ExpCM+$\Gamma\omega$ similarly draws ω from a Γ distribution discretized into $K = 4$ bins. Each bin is equally weighted and ω takes on the mean value of the bin. Because the Γ distribution is defined by two parameters, adding $\Gamma\omega$ to a model with a single ω adds one free parameter. Therefore, the GY94+$\Gamma\omega$ model has 12 free parameters, and the ExpCM+$\Gamma\omega$ model has 6 free parameters.

GY94 with ω_r

In Figure 7, we describe GY94 models where each site r has its own ω_r value that is calculated from the amino-acid preferences using the relationship described by [124]. This relationship defines the expected rate of nonsynonymous to synonymous substitutions given the amino-acid preferences. We first fit an ExpCM to the "low divergence" H1 subtree (parameter values in 9), which allows us to calculate $P_{r,xy}$ (Equation 3), Q_{xy} (Equation 4), and $p_{r,x}$ (Equation 6). We then calculated ω_r using the equation of [124], normalizing by the gene-wide ω fit by the ExpCM:

$$\omega_r = \frac{\sum_x \sum_{y \in N_x} p_{r,x} \times \frac{P_{r,xy}}{\omega}}{\sum_x \sum_{y \in N_x} p_{r,x} \times Q_{xy}}, \qquad \text{(Equation 7)}$$

where N_x is the set of codons that are nonsynonymous to codon x and differ from codon x by only one nucleotide.

4.5.2 HA amino-acid preferences from deep mutational scanning experiments

We used amino-acid preferences measured in deep mutational scans of the A/WSN/1933 H1 HA [27] and the A/Perth/2009 H3 HA [75] to define the amino-acid preferences that inform the ExpCMs. We only used sites that can be unambiguously aligned in these H1 and H3 HAs. These alignable sites and their mapping to sequential numbering of the HA sequences used in the deep mutational scanning experiments are in 9. The experimentally measured amino-acid preferences masked to just include these alignable sites are in 10 and 11. For the average preference set, we took the pairwise average of the H1 and H3 preferences. The preference for every amino acid a at every site r in the average preference set is

$$\pi_{r,a,(\mathrm{H1+H3\ avg})} = \frac{\pi_{r,a,\mathrm{H1}} + \pi_{r,a,\mathrm{H3}}}{2} \qquad \text{(Equation 8)}$$

4.5.3 HA sequences and tree topology

We downloaded all full-length, coding sequences for 15 of the 18 influenza A virus HA subtypes from the Influenza Virus Resource Database [7] in June of 2017. We excluded rare subtypes 15, 17, and 18, which have few sequences in the database. We filtered and aligned the sequences using `phydms_prepalignment` [59]. Specifically, we used `phydms_prepalignment` with the flag `--minidentity 0.3` to remove sequences with ambiguous nucleotides, premature stops, or frameshift mutations as well as redundant sequences. We also removed all codon sites which that are not alignable between the H1 HA and H3 HA used in the deep mutational scanning experiments (these alignable sites are listed in 9). We subsampled the remaining sequences to five per subtype with ≤ 1 sequence per year per subtype. We also included a small number of sequences from the major human and equine influenza lineages to ensure representation of these

well-studied lineages. The resulting alignment contains 92 sequences, and is provided in 12.

We created four subalignments with "low" and "intermediate" divergence from either the H1 or the H3 deep mutational scanning reference sequence for the analysis in Figure 11. The "low divergence" alignments had $\geq 59\%$ amino-acid identity to the sequence used in the deep mutational scanning, and the "intermediate divergence" alignments had $\geq 46\%$ identity from the reference sequence (17).

We inferred the tree topology of each alignment using `RAxML` [126] and the GTRCAT model. We estimated the branch lengths of this fixed topology using each ExpCM and GY94 models with `phydms_comprehensive` [59].

4.5.4 Asymptotic amino-acid sequence identity

For the analysis in Figure 7, we fit models to the "low divergence" H1 subtree. This gave the parameter values in 9.

For each model, we calculated the expected amino-acid sequence identity for two sequences separated by a branch length of t as

$$\sum_a \sum_{x \in a} p_{r,x} \sum_{y \in a} \left[e^{t\mathbf{P_r}}\right]_{xy} \qquad \text{(Equation 9)}$$

where a ranges over all 20 amino acids, $x \in a$ indicates that x ranges over all codons that encode amino-acid a, $p_{r,x}$ is the stationary state of the model at site r and codon x (given by Equation 2 for GY94-family models, and Equation 6 for ExpCM-family models), and $\left[e^{t\mathbf{P_r}}\right]_{xy}$ is the value in row x and column y of the matrix obtained by exponentiating the product of t and the substitution matrix $\mathbf{P_r}$ for site r (defined by Equation 1 for GY94-family models and Equation 3 for ExpCM-family models).

4.5.5 Simulations

For Figure 8, we simulated sequences using `pyvolve` [123] along the full HA tree using an ExpCM defined by parameters fit to the "low divergence" H1 subtree (9). We performed 10 replicate simulations and estimated the branch lengths for each replicate using `phydms_comprehensive` [59].

4.5.6 pbMutSel inference with `PhyloBayes-MPI`.

For Figure 12, we fit a pbMutSel model to the full HA tree. We ran one chain for 5500 steps, saved every sample, and discarded the first 550 samples as a burnin. We used `PhyloBayes-MPI` program `readpb_mpi` to compute the majority-rule consensus tree and the posterior average site-specific amino-acid preferences. Convergence was assessed visually using Tracer [103] and by the correlation of amino-acid preferences inferred by two independent chains (r=0.996).

In order to make the branch lengths in Figure 12 comparable between the pbMutSel tree returned by `PhyloBayes-MPI` and the other trees returned by `phydms`, we we normalized the branch lengths on the pbMutSel consensus tree and the ExpCM(H1+H3 avg)+$\Gamma\omega$ by dividing each branch by the length from A/South Carolina/1/1918 and A/Solomon Islands/3/2006. These two H1 sequences are early and late representatives of the longest known human influenza lineage, and are of sufficiently high identity that different ExpCM and GY94 substitution models all estimate nearly identical branch lengths separating them.

4.5.7 Software versions and computer code

All code used for the analyses in this paper is available at `https://github.com/jbloomlab/divergence_timing_manuscript`. The external computer programs that we used were

- `phydms` [59] version 2.2.2 (available at `github.com/jbloomlab/phydms`) to fit the Ex-

pCM and GY94 models.

- `pyvolve` [123] version 0.8.7 (available at `https://github.com/sjspielman/pyvolve`) to simulate the sequences.
- `PhyloBayes-MPI` [107] version 1.8 (available at `https://github.com/bayesiancook/pbmpi`) to fit the pbMutSel model.
- `RAxML` [126] version 8.2.11 (available at `https://github.com/stamatak/standard-RAxML`) to infer tree topology.
- We used `ggplot2` [147], `ggtree` [163], and `ggseqlogo` [140] for visualization of the results.
- `snakemake` [69] version 3.11.2 (available at `https://snakemake.readthedocs.io/en/stable/`) to run the pipelines.

Supplemental files

Descriptions of each file are shown below. Please see the publication for the actual files [58].

S9 File. List of alignable sites between H1 HA and H3 HA. This files provides a conversion between the numbering scheme we use in the paper (sequential numbering of just the alignable sites) to sequential numbering of the H1 HA reference sequence A/Wilson Smith/1933 and the H3 HA reference sequence A/Perth/2009.

S10 File. Amino acid preferences measured by the deep mutational scanning of the H1 HA strain A/WSN/1933 [27]. This file only contains measurements for the alignable sites between H1 and H3 HAs. Conversion from this numbering scheme to sequential numbering of A/WSN/1933 is in 9.

S11 File. Amino acid preferences measured by the deep mutational scanning of the H3 HA strain A/Perth/2009 [75]. This file only contains measurements for the alignable sites between H1 and H3 HAs. Conversion from this numbering scheme to sequential numbering of A/Perth/2009 is in in 9.

S12 File. The HA sequences for the full HA tree.. The sequences in this alignment contain only the alignable sites between H1 and H3 HAs. Conversion from this numbering scheme to sequential numbering of A/Perth/2009 is in in 9.

Acknowledgments

We thank Erick Matsen and Trevor Bedford for helpful comments about the project and manuscript. SKH is supported in part by training grant T32AI083203 from the NIAID of the National Institutes of Health. This work was supported by the NIAID and NIGMS of the NIH under grant numbers R01AI127893 and R35GM126911. JDB is an Investigator of the Howard Hughes Medical Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. All code and data used for the analyses in this paper is available at `https://github.com/jbloomlab/divergence_timing_manuscript`.

Chapter 5
CONCLUSION

My graduate research has focused on developing computational tools and methods that integrate two existing methods, phylogenetics and deep mutational scanning, to understand the effect of site-specific amino-acid constraint on protein evolution. These two methods, one computational and one experimental, have complementary strengths and weaknesses. Phylogenetics provides methods to study natural sequences, which are subjected to natural selective pressures, in a principled manner; however, these methods are constrained to the genetic sequences we have sampled. Deep mutational scanning allows for the unbiased measurement of all single amino-acid changes to a protein, but the assay occurs in an artificial laboratory setting. Here, I have described my work leveraging each method's strengths, the realism of the natural sequences with the completeness of the deep mutational scan, to form a better understanding of the constraint on a given protein.

Deep mutational scanning is a flexible, experimental framework for measuring the functional effect of single-amino-acid changes. Since the first publication [40], deep mutational scanning has been used to study dozens of unique proteins [33]. These studies serve a diverse range of goals, such as protein engineering [151], basic evolutionary principles [130], and interpreting clinical variation [128, 42]. However, a common challenge for all of these applications is data analysis. Deep mutational scanning produces large datasets which are often analyzed by iteratively looking at site-level summary metrics, individual mutation measurements, and the 3-D protein structure. Furthermore, it is often important to contextualize the results with external datasets, such as previous experiments, natural sequence data, or clinical data. Currently, these analyses rely on several

independent software pipelines and a high level of computational expertise, which could make them prohibitive to some groups.

In Chapter 2, I developed `dms-view`, a web-based visualization tool for deep mutational scanning experiments to address these concerns. Using `dms-view`, users can select sites of interest to view individual mutation measurements and the sites on the 3-D protein structure. `dms-view` tracks the input data and user selections in the URL, making it possible to save specific views of interactively generated visualizations to share with collaborators. I designed `dms-view` to be agnostic to the exact experimental setup to match the flexibility of the deep mutational scan itself. Rather than take in raw data from a deep mutational scan and compute standard metrics, `dms-view` displays user-calculated metrics tailored to their experiment. For example, at the mutation level, a user could view deep mutational scanning amino-acid preferences or the natural frequencies in nature.

In the future, I want to continue developing `dms-view` to expand its use within the deep mutational scanning community. One way to expand `dms-view`'s audience is to integrate it better with existing deep mutational scanning analysis infrastructure. MaveDB is a centralized deep mutational scanning database [33], which facilitates meta-analysis studies or comparison of related deep mutational scans. I would work with MaveDB developers to make `dms-view` a default option for visualizing existing and new deep mutational scanning datasets. Having an easy, interactive visualization for each deep mutational scan in the database will help others quickly browse and explore the available datasets without having to download the raw data and build a new visualization pipeline.

While tools like `dms-view` allow for qualitative comparisons of deep mutation scans and natural amino-acid frequencies, more sophisticated methods are needed to make this comparison in a way that accounts for sequence sampling and shared evolutionary history. Molecular phylogenetic algorithms enable the calculation of the statistical likelihood of an alignment of naturally occurring gene sequences given a phylogenetic tree and a substitution model [36, 35]. Using Experimentally Informed Codon Models (ExpCMs) [11], which describe the selection on amino-acid changes using the results of a

deep mutational scan, I can link and compare the selection in the laboratory to selection in nature.

In Chapter 3, I implemented ExpCMs in a new python package, `phydms`. `phydms` runs approximately ~ 100 fold faster than the previous implementation of ExpCMs. I wrote several command line tools to assist with and streamline analyses with ExpCMs. A common analysis is to ask whether an ExpCM defined by a particular deep mutational scan is a better descriptor of natural sequence evolution than another ExpCM or a standard codon model. The command line tool `phydms_prepalignment` curates sequences for phylogenetic analysis and `phydms_comprehensive` facilitates the comparison by fitting the different models of interest and summarizing key results and parameters. Since its publication, `phydms` has been used to analyze deep mutational scanning data for influenza [75, 121, 62] and HIV [51] proteins.

Post publication, I have continued to develop `phydms` and ExpCMs. For example, I implemented a standard method to account for rate variation across sites. In the original version of `phydms`, I implemented the ExpCM with a gene-wide parameter called ω or the ratio of nonsynonymous to synonymous mutations. Now, ω can follow a Γ-distribution rather than being fit to a single gene-wide value. I found that this additional rate variation was critical for ExpCMs for the HIV protein Env [51]. I hypothesize that the Γ-distributed ω helps account for the fact that while Env is under strong immune pressure in nature [4, 104, 143], immune selection is absent in the deep mutational scan.

A natural future direction for `phydms` and ExpCMs is the development of additional methods to test for sites under diversifying selection. Identifying sites under diversifying selection is of special interest for viral proteins because they could indicate sites of virus-host interactions. Current phylogenetic methods to detect site-specific diversifying selection focus on identifying sites with an elevated ratio of nonsynonymous to synonymous mutations [99]. Previous work to detect positive selection with ExpCMs extended these approaches [14].

However, I propose an alternative strategy which tests whether or not experimentally-

defined sites of immune selection are relevant to natural sequence evolution. There are decades of work with influenza using experimental methods to identify sites targeted by the immune system, from early working selecting escape variants from monoclonal antibodies [162] to more recent deep mutational scanning approaches with polyclonal human sera [74]. While varied in design, these experiments all result in a set of sites hypothesized to be targeted by the immune system in nature.

I propose testing the relevance of these experimentally-defined sites for natural sequence evolution using ExpCMs. This test follows the same philosophy that underlies ExpCMs, investigating whether precise, controlled experimental methods are good descriptors of natural sequence evolution. I would replace the gene-wide ω with the expression $\omega_1 (1 + \omega_2 \delta_r)$, where δ_r is a site-specific variable representing the experimental hypothesis. δ_r could be qualitative indictor variable (1 if an immune site else 0) or a continuous variable taking on the "strength" of selection measured by the deep mutational scanning-like approach. The test would then ask if the ExpCM with the experimentally-defined sites of diversifying selection (ω_2 is free) is a better model than an ExpCM without these sites ($\omega_2 = 0$). We could assess significance using either a likelihood ratio test [63] or a randomization approach. This strategy is not specific to ExpCMs but could be applied to other codon models which contains an ω-like term. Overall, this test would provide another principled way to compare the results of an experimental assay investigating protein evolution to the selective pressures the protein faces in nature and complement existing methods trying to infer sites under positive selection from natural sequences.

In addition to providing a way to evaluate the relevance of a deep mutational scan to natural sequence evolution, the site-specific nature of ExpCMs affects phylogenetic inference itself. In Chapter 4, I investigated the effect of modeling site-specific constraint with deep mutational scanning data on the estimation of branch lengths. A long-standing observation in phylogenetics is that long branches in phylogenetic trees are consistently underestimated [31, 60, 1]. Halpern and Bruno [52] hypothesized that site-specific substitution models could alleviate this problem by modeling purifying selection more accurately.

To address this hypothesis, I evaluated whether or not site-specific ExpCMs estimate longer branches than a site-uniform codon model. I found that while ExpCMs do estimate longer branches on the influenza virus tree than the site-uniform model, this effect is limited by intraprotein epistasis. Because I had ExpCMs from two diverged homologs, I could observe that branch length extension was more pronounced near the wildtype sequence of the deep mutational scan. While modeling site-specific amino-acid preferences is important for branch length estimation, there is not a single set of evolutionary-stable stationary-state frequencies across an entire tree.

This work suggests that for a more accurate model of protein evolution, we need to relax the assumption that there is a single description of site-specific constraint across the entire tree. One imperfect step forward for ExpCMs is to allow the site-specific amino-acid preferences to "degrade" as sequences become more evolutionary diverged from the wildtype sequence of the deep mutational scan. Following the strategy of Smith *et al.* (2015) [120] for branch-specific rate variation, I could define the site-specific amino-acid preferences as a combination of a deep mutational scan and the amino-acid preferences from a site-uniform model and allow the proportion to vary along different branches of the tree. In this way, the site-specific amino-acid preferences would not be constrained to be uniform across the entire tree.

However, the strategy above does not address the underlying mechanistic weakness in current phylogenetic models; site-specific amino-acid preferences are site intradependent. Others have also investigated the effects of misspecification under the incorrect assumption of site independence [109, 22, 96, 45] but relaxing the assumption requires an algorithmic alternative to the Felsenstein pruning algorithm. Recent work has shown promise for alternative algorithms [16, 106, 109, 22] and experimental [92, 152, 131, 76] and statistical [116, 94] methods for epistasis-aware fitness functions. Additional work remains to address concerns of computational tractability and biological-realism, respectively. My work in Chapter Four shows that ExpCMs defined by deep mutational scans from two diverged homologs imperfectly capture the effect of shifting preferences and

therefore could be used as standard to evaluate new site-intradependent methods.

Overall, I have worked on understanding site-specific constrain on proteins using both molecular phylogenetics and experimental functional data. I have shown that leveraging this complimentary approaches can create a more comprehensive understanding of the natural selective pressures on a protein. While my work has focused mainly on influenza virus evolution, these are general techniques and tools that are applicable to the evolution of many, diverse proteins.

BIBLIOGRAPHY

[1] Pakorn Aiewsakun and Aris Katzourakis. Time-dependent rate phenomenon in viruses. *Journal of Virology*, 90(16):7184–7195, 2016.

[2] Chiara A Airoldi, Sara Bergonzi, and Brendan Davies. Single amino acid change alters the ability to specify male or female organ identity. *Proceedings of the National Academy of Sciences*, 107(44):18898–18902, 2010.

[3] Laith Q Al-Mawsawi, Nicholas C Wu, C Anders Olson, Vivian Cai Shi, Hangfei Qi, Xiaojuan Zheng, Ting-Ting Wu, and Ren Sun. High-throughput profiling of point mutations across the hiv-1 genome. *Retrovirology*, 11(1):124, 2014.

[4] J Albert, B Abrahamsson, K Nagy, E Aurelius, Hans Gaines, G Nystrom, and EM Fenyii. Rapid development of isolate-specific neutralizing antibodies and consequent emergence of virus variants which resist neutralization by autologous sera. *AIDS*, 4(107-112):22J, 1990.

[5] Miguel Arenas. Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 6:319, 2015.

[6] Orr Ashenberg, Jai Padmakumar, Michael B Doud, and Jesse D Bloom. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS pathogens*, 13(3):e1006288, 2017.

[7] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the National Center for Biotechnology Information. *Journal of Virology*, 82(2):596–601, 2008.

[8] Georgii A Bazykin. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biology Letters*, 11(10):20150315, 2015.

[9] T. Bedford, M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, A. J. Hay, J. W. McCauley, C. A. Russell, D. J. Smith, and A. Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, 2014.

[10] Samir Bhatt, Edward C Holmes, and Oliver G Pybus. The genomic rate of molecular adaptation of the human influenza a virus. *Molecular biology and evolution*, 28(9):2443–2451, 2011.

[11] Jesse D Bloom. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution*, 31:1956–1978, 2014.

[12] Jesse D Bloom. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol. Biol. Evol.*, 31:2753–2769, 2014.

[13] Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC bioinformatics*, 16(1):168, 2015.

[14] Jesse D Bloom. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*, 12:1, 2017.

[15] Maciej F Boni, Yang Zhou, Jeffery K Taubenberger, and Edward C Holmes. Homologous recombination is very rare or absent in human influenza a virus. *Journal of virology*, 82(10):4807–4811, 2008.

[16] Andrew J Bordner and Hans D Mittelmann. A new formulation of protein evolutionary models that account for structural constraints. *Molecular Biology and Evolution*, 31(3):736–749, 2013.

[17] Lisa Brenan, Aleksandr Andreev, Ofir Cohen, Sasha Pantel, Atanas Kamburov, Davide Cacchiarelli, Nicole S Persky, Cong Zhu, Mukta Bagul, Eva M Goetz, T S Mikkelsen, F Piccioni, D E Root, and C M Johannessen. Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell reports*, 17(4):1171–1183, 2016.

[18] David Bryant, Nicolas Galtier, and Marie Anne Poursat. Likelihood calculation in molecular phylogenetics. *Mathematics of evolution and phylogeny*, pages 33–62, 2005.

[19] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[20] Serena A Carroll, Jonathan S Towner, Tara K Sealy, Laura K McMullan, Marina L Khristova, Felicity J Burt, Robert Swanepoel, Pierre E Rollin, and Stuart T Nichol. Molecular evolution of viruses of the family filoviridae based on 97 whole-genome sequences. *Journal of Virology*, 87(5):2608–2616, 2013.

[21] Benjamin S Chambers, Kaela Parkhouse, Ted M Ross, Kevin Alby, and Scott E Hensley. Identification of hemagglutinin residues responsible for h3n2 antigenic drift during the 2014–2015 influenza season. *Cell reports*, 12(1):1–6, 2015.

[22] Sang Chul Choi, Asger Hobolth, Douglas M Robinson, Hirohisa Kishino, and Jeffrey L Thorne. Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular Biology and Evolution*, 24(8):1769–1782, 2007.

[23] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

[24] Antony M Dean and Joseph W Thornton. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Reviews Genetics*, 8(9):675–688, 2007.

[25] Adam S Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D Bloom. An antigenic atlas of hiv-1 escape from broadly neutralizing antibodies distinguishes functional and structural epitopes. *Immunity*, 50(2):520–532, 2019.

[26] Michael B Doud, Orr Ashenberg, and Jesse D Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.*, 32:2944–2960, 2015.

[27] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses*, 8:155, 2016.

[28] Alexei J Drummond, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88, 2006.

[29] David A Duchêne, Sebastian Duchêne, Edward C Holmes, and Simon YW Ho. Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Molecular Biology and Evolution*, 32(11):2986–2995, 2015.

[30] Sebastián Duchêne, Francesca Di Giallonardo, and Edward C Holmes. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Molecular Biology and Evolution*, 33(1):255–267, 2015.

[31] Sebastián Duchêne, Edward C Holmes, and Simon YW Ho. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B*, 281(1786):20140732, 2014.

[32] Julian Echave, Stephanie J Spielman, and Claus O Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 2016.

[33] Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome biology*, 20(1):1–11, 2019.

[34] Mario Ali Fares and Edward C Holmes. A revised evolutionary history of hepatitis b virus (HBV). *Journal of Molecular Evolution*, 54(6):807–814, 2002.

[35] J. Felsenstein. Maximum likelihood and minimum-step methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22:240–249, 1973.

[36] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[37] Joseph Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410, 1978.

[38] Jason D Fernandes, Tyler B Faust, Nicolas B Strauli, Cynthia Smith, David C Crosby, Robert L Nakamura, Ryan D Hernandez, and Alan D Frankel. Functional segregation of overlapping genes in HIV. *Cell*, 167(7):1762–1773, 2016.

[39] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.*, 31:1581–1592, 2014.

[40] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, 2014.

[41] Yuki Furuse, Akira Suzuki, and Hitoshi Oshitani. Origin of measles virus: divergence from rinderpest virus between the 11th and 12th centuries. *Virology journal*, 7(1):52, 2010.

[42] Hannah Gelman, Jennifer N Dines, Jonathan Berg, Alice H Berger, Sarah Brnich, Fuki M Hisama, Richard G James, Alan F Rubin, Jay Shendure, Brian Shirts, et al. Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Medicine*, 11(1):85, 2019.

[43] Philip E Gill, Walter Murray, and Margaret H Wright. *Practical optimization*. Academic Press, Cambridge, Massachusetts, 1982.

[44] Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

[45] Richard A Goldstein and David D Pollock. Sequence entropy of folding and the absolute rate of amino acid substitutions. *Nature ecology & evolution*, 1(12):1923–1930, 2017.

[46] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, 2013.

[47] Katelyn M Gostic, Monique Ambrose, Michael Worobey, and James O Lloyd-Smith. Potent protection against h5n1 and h7n9 influenza via childhood hemagglutinin imprinting. *Science*, 354(6313):722–726, 2016.

[48] Laurent Guéguen, Sylvain Gaillard, Bastien Boussau, Manolo Gouy, Mathieu Groussin, Nicolas C Rochette, Thomas Bigot, David Fournier, Fanny Pouyet, Vincent Cahais, et al. Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol*, 30(8):1745–1750, 2013.

[49] Ya Ha, David J Stevens, John J Skehel, and Don C Wiley. H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. *The EMBO journal*, 21(5):865–875, 2002.

[50] Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathogens*, 12(12):e1006114, 2016.

[51] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife*, 7:e34420, 2018.

[52] Aaron L Halpern and William J Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917, 1998.

[53] Michael J Harms and Joseph W Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14(8):559–571, 2013.

[54] Michael J Harms and Joseph W Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203–207, 2014.

[55] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[56] Alan J Hay, Victoria Gregory, Alan R Douglas, and Yi Pu Lin. The evolution of human influenza viruses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1416):1861–1870, 2001.

[57] Frederick Hayden. Developing new antiviral agents for influenza treatment: what does the future hold? *Clinical Infectious Diseases*, 48(Supplement_1):S3–S13, 2009.

[58] Sarah K Hilton and Jesse D Bloom. Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence. *Virus evolution*, 4(2):vey033, 2018.

[59] Sarah K Hilton, Michael B Doud, and Jesse D Bloom. phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*, 5:e3657, 2017.

[60] Simon YW Ho, Sebastián Duchêne, Martyna Molak, and Beth Shapiro. Time-dependent estimates of molecular evolutionary rates: evidence and causes. *Molecular Ecology*, 24(24):6007–6012, 2015.

[61] Edward C Holmes. Molecular clocks and the puzzle of RNA virus origins. *Journal of Virology*, 77(7):3893–3897, 2003.

[62] Nancy Hom, Lauren Gentles, Jesse D Bloom, and Kelly K Lee. Deep mutational scan of the highly conserved influenza a virus m1 matrix protein reveals substantial intrinsic mutational tolerance. *Journal of virology*, 93(13):e00161–19, 2019.

[63] John P Huelsenbeck and Keith A Crandall. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28(1):437–466, 1997.

[64] Amanda E Jetzt, Hong Yu, George J Klarmann, Yacov Ron, Bradley D Preston, and Joseph P Dougherty. High rate of recombination throughout the human immunodeficiency virus type 1 genome. *Journal of virology*, 74(3):1234–1240, 2000.

[65] Aashiq H Kachroo, Jon M Laurent, Christopher M Yellman, Austin G Meyer, Claus O Wilke, and Edward M Marcotte. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, 348(6237):921–925, 2015.

[66] J. D. Kalbeisch and J. F. Lawless. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871, 1985.

[67] Toby Kenney, Hong Gu, et al. Hessian calculation for phylogenetic likelihood based on the pruning algorithm and its applications. *Statistical applications in genetics and molecular biology*, 11(4):1–46, 2012.

[68] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature Methods*, 12(3):203–206, 2015.

[69] Johannes Köster and Sven Rahmann. Snakemake——a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[70] Nicolas Lartillot. The Bayesian Kitchen: overcoming the fear of over-paramerization. `http://bayesiancook.blogspot.com/2014/01/the-myth-of-over-parameterization.html`, 2014. Last accessed: March-12-2018.

[71] Nicolas Lartillot, Henner Brinkmann, and Hervé Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7(Suppl 1):S4, 2007.

[72] Nicolas Lartillot and Hervé Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004.

[73] Si Quang Le, Nicolas Lartillot, and Olivier Gascuel. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B*, 363(1512):3965–3976, 2008.

[74] Juhye M Lee, Rachel Eguia, Seth J Zost, Saket Choudhary, Patrick C Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron C Hurt, Seema S Lakdawala, et al. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife*, 8, 2019.

[75] Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences*, DOI 10.1073/pnas.1806133115, 2018.

[76] Chuan Li, Wenfeng Qian, Calum J Maclean, and Jianzhi Zhang. The fitness landscape of a trna gene. *Science*, 352(6287):837–840, 2016.

[77] Amit R Majithia, Ben Tsuda, Maura Agostini, Keerthana Gnanapradeepan, Robert Rice, Gina Peloso, Kashyap A Patel, Xiaolan Zhang, Marjoleine F Broekema, Nick Patterson, M Duby, T Sharpe, E Kalkhoven, E D Rosen, I Barraso, S Ellard, UK Monogenic Diabetes Consortium, S Kathiresan, Myocardial Infarction Genetics Consortium, S O'Rahilly, UK Congenital Lipodystrophy Consortium, K Chatterjee, J C Florez, T Mikkelsen, D B Savage, and D Altshuler. Prospective functional classification of all possible missense variants in PPARG. *Nature genetics*, 48(12):1570–1575, 2016.

[78] Kenneth A Matreyek, Lea M Starita, Jason J Stephany, Beth Martin, Melissa A Chiasson, Vanessa E Gray, Martin Kircher, Arineh Khechaduri, Jennifer N Dines, Ronald J Hause, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature genetics*, 50(6):874–882, 2018.

[79] David Mavor, Kyle Barlow, Samuel Thompson, Benjamin A Barad, Alain R Bonny, Clinton L Cario, Garrett Gaskins, Zairan Liu, Laura Deming, Seth D Axen, Elena Caceres, Weilin Chen, Adolfo Cuesta, Rachel E Gate, Evan M Green, Kaitlin R Hulce, Weiyue Ji, Lillian R Kenner, Bruk Mensa, Leanna S Morinishi, Steven M Moss, Marco Mravic, Ryan K Muir, Stefan Niekamp, Chimno I Nnadi, Eugene Palovcak, Erin M Poss, Tyler D Ross, Eugenia C Salcedo, Stephanie K See, Meena Subramaniam, Allison W Wong, Jennifer Li, Kurt S Thorn, Shane O Conchuir, Benjamin P Roscoe, Eric D Chow, Joseph L DeRisi, Tanja Kortemme, Daniel N Bolon, and James S Fraser. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, 5, 2016.

[80] David M McCandlish and Arlin Stoltzfus. Modeling evolution using the probability of fixation: history and implications. *The Quarterly Review of Biology*, 89(3):225–252, 2014.

[81] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and

Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138, 2012.

[82] Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):e112, 2014.

[83] Ignacio Mena, Martha I Nelson, Francisco Quezada-Monroy, Jayeeta Dutta, Refugio Cortes-Fernández, J Horacio Lara-Puente, Felipa Castro-Peralta, Luis F Cunha, Nídia S Trovão, Bernardo Lozano-Dubernard, et al. Origins of the 2009 h1n1 influenza pandemic in swine in mexico. *Elife*, 5:e16777, 2016.

[84] Parul Mishra, Julia M Flynn, Tyler N Starr, and Daniel NA Bolon. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell reports*, 15(3):588–598, 2016.

[85] José Luis Morales and Jorge Nocedal. Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". *ACM Transactions on Mathematical Software (TOMS)*, 38(1):7, 2011.

[86] Ben Murrell, Steven Weaver, Martin D Smith, Joel O Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P Martin, Davey M Smith, et al. Gene-wide identification of episodic selection. *Molecular Biology and Evolution*, 32(5):1365–1371, 2015.

[87] Spencer V Muse and Brandon S Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, 11(5):715–724, 1994.

[88] Rasmus Nielsen. *Statistical methods in molecular evolution*. Springer, 2006.

[89] Rasmus Nielsen and Ziheng Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, 1998.

[90] E Nobusawa, T Aoyama, H Kato, Y Suzuki, Y Tateno, and K Nakajima. Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology*, 182(2):475–485, 1991.

[91] Eri Nobusawa and Katsuhiko Sato. Comparison of the mutation rates of human influenza a and b viruses. *Journal of virology*, 80(7):3675–3678, 2006.

[92] C Anders Olson, Nicholas C Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014.

[93] E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, and J. W. Thornton. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317:1544–1548, 2007.

[94] Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, DOI 10.1073/pnas.1804015115, 2018.

[95] Hervé Philippe and Jacqueline Laurent. How good are deep phylogenetic trees? *Current Opinion in Genetics & Development*, 8(6):616–623, 1998.

[96] David D Pollock, Grant Thiltgen, and Richard A Goldstein. Amino acid coevolution induces an evolutionary stokes shift. *Proc. Natl. Acad. Sci. USA*, 109(21):E1352–E1359, 2012.

[97] Sergei Kosakovsky Pond, Wayne Delport, Spencer V Muse, and Konrad Scheffler. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*, 5(7):e11230, 2010.

[98] Sergei L Pond, Simon DW Frost, and Spencer V Muse. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, 2005.

[99] Sergei L Kosakovsky Pond and Simon DW Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22(5):1208–1222, 2005.

[100] David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004.

[101] Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. *PLoS pathogens*, 10(4), 2014.

[102] Si Le Quang, Olivier Gascuel, and Nicolas Lartillot. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323, 2008.

[103] Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. Posterior summarisation in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 2018.

[104] Douglas D Richman, Terri Wrin, Susan J Little, and Christos J Petropoulos. Rapid evolution of the neutralizing antibody response to hiv type 1 infection. *Proceedings of the National Academy of Sciences*, 100(7):4144–4149, 2003.

[105] Nicolas Rodrigue. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*, 193(2):557–564, 2013.

[106] Nicolas Rodrigue, Claudia L Kleinman, Hervé Philippe, and Nicolas Lartillot. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Molecular Biology and Evolution*, 26(7):1663–1676, 2009.

[107] Nicolas Rodrigue and Nicolas Lartillot. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7):1020–1021, 2014.

[108] Nicolas Rodrigue and Nicolas Lartillot. Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. *Molecular Biology and Evolution*, 34(1):204–214, 2017.

[109] Nicolas Rodrigue, Nicolas Lartillot, David Bryant, and HervE Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347(2):207–217, 2005.

[110] Nicolas Rodrigue, Hervé Philippe, and Nicolas Lartillot. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10):4629–4634, 2010.

[111] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.*, 2013.

[112] Alan F Rubin, Hannah Gelman, Nathan Lucas, Sandra M Bajjalieh, Anthony T Papenfuss, Terence P Speed, and Douglas M Fowler. A statistical framework for analyzing deep mutational scanning data. *Genome biology*, 18(1):150, 2017.

[113] Stuart Rudikoff, Angela M Giusti, Wendy D Cook, and Matthew D Scharff. Single amino acid substitution altering antigen-binding specificity. *Proceedings of the National Academy of Sciences*, 79(6):1979–1983, 1982.

[114] RJ Russell, SJ Gamblin, LF Haire, DJ Stevens, B Xiao, Y Ha, and JJ Skehel. H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes. *Virology*, 325(2):287–296, 2004.

[115] Alexias Safi, Kelley A Wallace, and Laura N Rusche. Evolution of new function through a single amino acid change in the yeast repressor sum1p. *Molecular and cellular biology*, 28(8):2567–2578, 2008.

[116] Zachary R Sailer and Michael J Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017.

[117] Rafael Sanjuán, Miguel R Nebot, Nicola Chirico, Louis M Mansky, and Robert Belshaw. Viral mutation rates. *Journal of virology*, 84(19):9733–9748, 2010.

[118] Premal Shah, David M McCandlish, and Joshua B Plotkin. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*, 112(25):E3226–E3235, 2015.

[119] Gavin JD Smith, Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J Lycett, Michael Worobey, Oliver G Pybus, Siu Kit Ma, Chung Lam Cheung, Jayna Raghwani, Samir Bhatt, et al. Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. *Nature*, 459(7250):1122–1125, 2009.

[120] Martin D Smith, Joel O Wertheim, Steven Weaver, Ben Murrell, Konrad Scheffler, and Sergei L Kosakovsky Pond. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular biology and evolution*, 32(5):1342–1353, 2015.

[121] YQ Shirleen Soh, Louise H Moncla, Rachel Eguia, Trevor Bedford, and Jesse D Bloom. Comprehensive mapping of adaptation of the avian influenza polymerase protein pb2 to humans. *Elife*, 8:e45079, 2019.

[122] Marion Sourisseau, Daniel JP Lawrence, Megan C Schwarz, Carina H Storrs, Ethan C Veit, Jesse D Bloom, and Matthew J Evans. Deep mutational scanning comprehensively maps how zika envelope protein mutations affect viral growth and antibody escape. *Journal of virology*, 93(23), 2019.

[123] Stephanie J Spielman and Claus O Wilke. Pyvolve: a flexible Python module for simulating sequences along phylogenies. *PLoS One*, 10(9):e0139047, 2015.

[124] Stephanie J Spielman and Claus O Wilke. The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, 32(4):1097–1108, 2015.

[125] Stephanie J Spielman and Claus O Wilke. Extensively parameterized mutation–selection models reliably capture site-specific selective constraint. *Molecular biology and evolution*, 33(11):2990–3002, 2016.

[126] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[127] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[128] Lea M Starita, Nadav Ahituv, Maitreya J Dunham, Jacob O Kitzman, Frederick P Roth, Georg Seelig, Jay Shendure, and Douglas M Fowler. Variant interpretation: functional assays to the rescue. *The American Journal of Human Genetics*, 101(3):315–325, 2017.

[129] Lea M Starita, David L Young, Muhtadi Islam, Jacob O Kitzman, Justin Gullingsrud, Ronald J Hause, Douglas M Fowler, Jeffrey D Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of brca1 ring domain variants. *Genetics*, 200(2):413–422, 2015.

[130] Tyler N Starr, Julia M Flynn, Parul Mishra, Daniel NA Bolon, and Joseph W Thornton. Pervasive contingency and entrenchment in a billion years of hsp90 evolution. *Proceedings of the National Academy of Sciences*, 115(17):4453–4458, 2018.

[131] Barrett Steinberg and Marc Ostermeier. Shifting fitness and epistatic landscapes reflect trade-offs along an evolutionary pathway. *Journal of Molecular Biology*, 428(13):2730–2743, 2016.

[132] Michael A Stiffler, Doeke R. Hekstra1, and Rama Ranganathan. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell*, 160(5):882–892, 2015.

[133] Paula Suárez-López and Juan Ortín. An estimation of the nucleotide substitution rate at defined positions in the influenza virus haemagglutinin gene. *Journal of general virology*, 75(2):389–393, 1994.

[134] Edward Susko, Léa Lincker, and Andrew J Roger. Accelerated estimation of frequency classes in site-heterogeneous profile mixture models. *Molecular Biology and Evolution*, 35(5):1266–1283, 2018.

[135] Asif U Tamuri, Mario dos Reis, and Richard A Goldstein. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(3):1101–1115, 2012.

[136] Asif U Tamuri, Nick Goldman, and Mario dos Reis. A penalized likelihood method for estimating the distribution of selection coefficients from phylogenetic data. *Genetics*, pages genetics–114, 2014.

[137] Derek J Taylor, Matthew J Ballinger, Jack J Zhan, Laura E Hanzly, and Jeremy A Bruenn. Evidence that Ebolaviruses and Cuevaviruses have been diverging from Marburgviruses since the Miocene. *PeerJ*, 2:e556, 2014.

[138] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, 2014.

[139] Danielle M Tufts, Chandrasekhar Natarajan, Inge G Revsbech, Joana Projecto-Garcia, Federico G Hoffmann, Roy E Weber, Angela Fago, Hideaki Moriyama, and Jay F Storz. Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Molecular Biology and Evolution*, 32(2):287–298, 2014.

[140] Omar Wagih. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, 33(22):3645–3647, 2017.

[141] Huai-Chun Wang, Karen Li, Edward Susko, and Andrew J Roger. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evolutionary Biology*, 8(1):331, 2008.

[142] Robert G Webster, William J Bean, Owen T Gorman, Thomas M Chambers, and Yoshihiro Kawaoka. Evolution and ecology of influenza a viruses. *Microbiology and molecular biology reviews*, 56(1):152–179, 1992.

[143] Xiping Wei, Julie M Decker, Shuyi Wang, Huxiong Hui, John C Kappes, Xiaoyun Wu, Jesus F Salazar-Gonzalez, Maria G Salazar, J Michael Kilby, Michael S Saag, et al. Antibody neutralization and escape by hiv-1. *Nature*, 422(6929):307–312, 2003.

[144] Joel O Wertheim, Daniel KW Chu, Joseph SM Peiris, Sergei L Kosakovsky Pond, and Leo LM Poon. A case for the ancient origin of coronaviruses. *Journal of Virology*, 87(12):7039–7045, 2013.

[145] Joel O Wertheim and Sergei L Kosakovsky Pond. Purifying selection can obscure the ancient age of viral lineages. *Molecular Biology and Evolution*, 28(12):3355–3365, 2011.

[146] Joel O Wertheim and Michael Worobey. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Computational Biology*, 5(5):e1000377, 2009.

[147] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

[148] Don C Wiley and John J Skehel. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual review of biochemistry*, 56(1):365–394, 1987.

[149] Michael Worobey, Guan-Zhu Han, and Andrew Rambaut. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254–257, 2014.

[150] Michael Worobey, Paul Telfer, Sandrine Souquière, Meredith Hunter, Clint A Coleman, Michael J Metzger, Patricia Reed, Maria Makuwa, Gail Hearn, Shaya Honarvar, et al. Island biogeography reveals the deep history of siv. *Science*, 329(5998):1487–1487, 2010.

[151] Emily E Wrenbeck, Matthew S Faber, and Timothy A Whitehead. Deep sequencing methods for protein engineering and design. *Current opinion in structural biology*, 45:36–44, 2017.

[152] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016.

[153] Nicholas C Wu, Yushen Du, Shuai Le, Arthur P Young, Tian-Hao Zhang, Yuanyuan Wang, Jian Zhou, Janice M Yoshizawa, Ling Dong, Xinmin Li, et al. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza a virus m segment. *BMC genomics*, 17(1):46, 2016.

[154] Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS genetics*, 11(7), 2015.

[155] Nicholas C Wu, Jia Xie, Tianqing Zheng, Corwin M Nycholat, Geramie Grande, James C Paulson, Richard A Lerner, and Ian A Wilson. Diversity of functionally permissive sequences in the receptor-binding site of influenza hemagglutinin. *Cell host & microbe*, 21(6):742–753, 2017.

[156] Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3):306–314, 1994.

[157] Ziheng Yang. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution*, 51(5):423–432, 2000.

[158] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[159] Ziheng Yang and Rasmus Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3):568–579, 2008.

[160] Ziheng Yang, Rasmus Nielsen, Nick Goldman, and Anne-Mette Krabbe Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449, 2000.

[161] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303, 2012.

[162] JW Yewdell, RG Webster, and WU Gerhard. Antigenic variation in three distinct determinants of an influenza type a haemagglutinin molecule. *Nature*, 279(5710):246–248, 1979.

[163] Guangchuang Yu, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees

with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017.

[164] XW Zhang, YL Yap, and A Danchin. Testing the hypothesis of a recombinant origin of the sars-associated coronavirus. *Archives of virology*, 150(1):1–20, 2005.

[165] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

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

LIST OF FIGURES

Figure Number Page

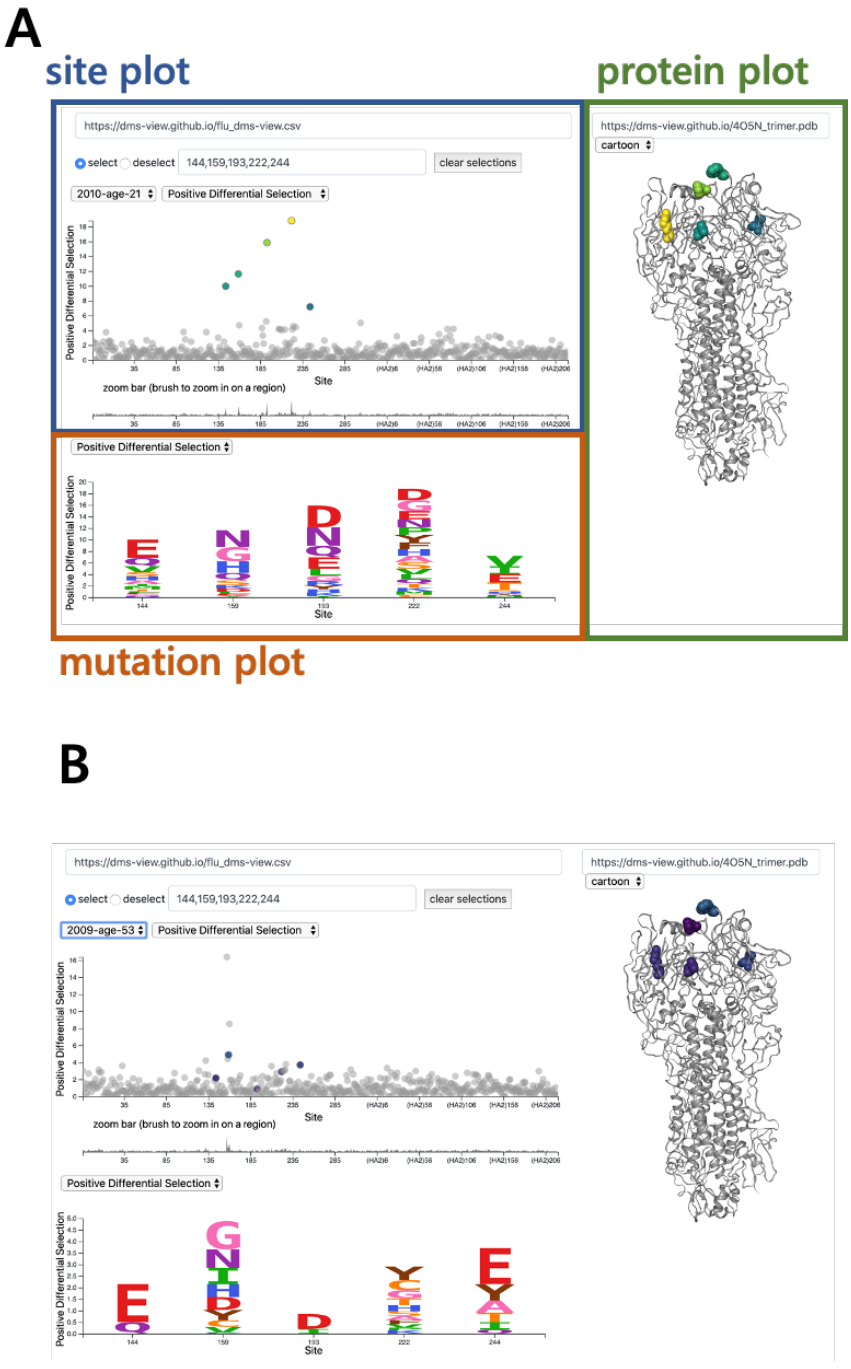

Figure 1: **Using `dms-view` to analyze a DMS. For further exploration of this dataset, please visit `https://dms-view.github.io`.** (A) The `dms-view` data section has three panels: the site plot, the mutation plot, and the protein structure plot. The interactive features for selecting sites and navigating are in the site plot panel. Here we show the five most highly targeted sites by the human sera "2010-Age-21" from the study by Lee *et al.* [74]. All five sites fall in the "globular head" of influenza virus HA. (B)The same five sites targeted by the sera in panel **A** but now plotted with the data from a different human sera, "2009-age-53". Using `dms-view` to compare, we see that different sites on HA are targeted by the different sera.

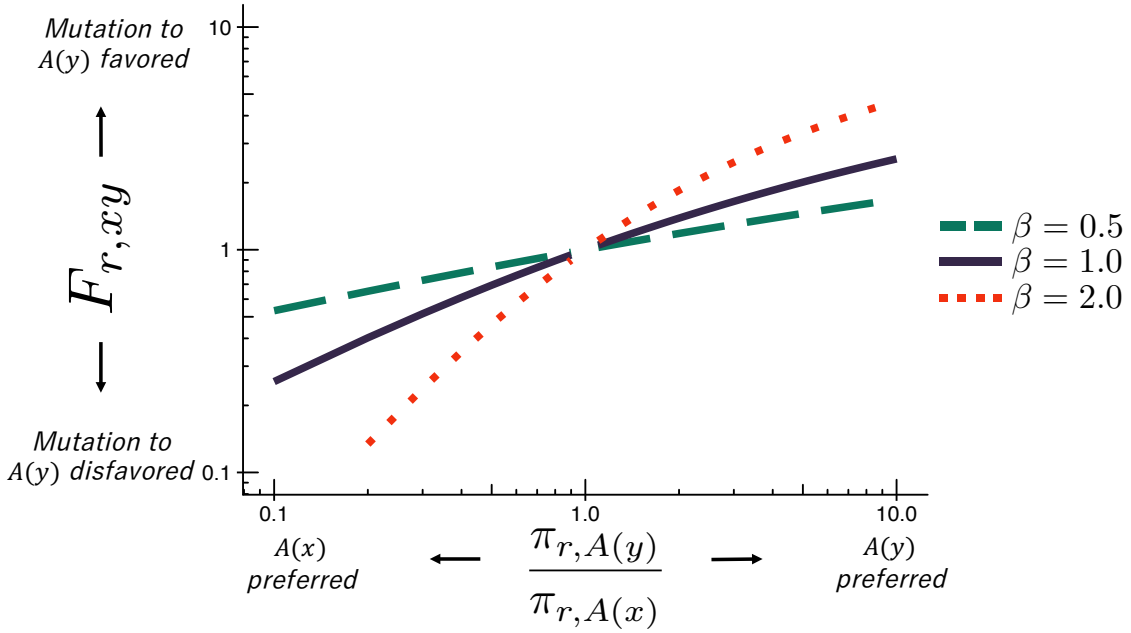

Figure 2: **The ExpCM fixation term $\mathbf{F_{r,xy}}$.** In an ExpCM, the rate of fixation of a mutation from codon x to codon y depends on the experimentally measured preferences of the amino acids $A\left(x\right)$ and $A\left(y\right)$ encoded by these codons. Mutations to preferred amino acids, with $\frac{\pi_{r,A(y)}}{\pi_{r,A(x)}} > 1$, result in a larger $F_{r,xy}$, and so are anticipated to fix more often. The value of $F_{r,xy}$ is modulated by re-scaling the preferences by a stringency parameter $\beta \neq 1$ to reflect differences in selection between the lab and nature. When $\beta > 1$, the selection for preferred amino acids is exaggerated. When $\beta < 1$, the selection for preferred amino acids is attenuated.

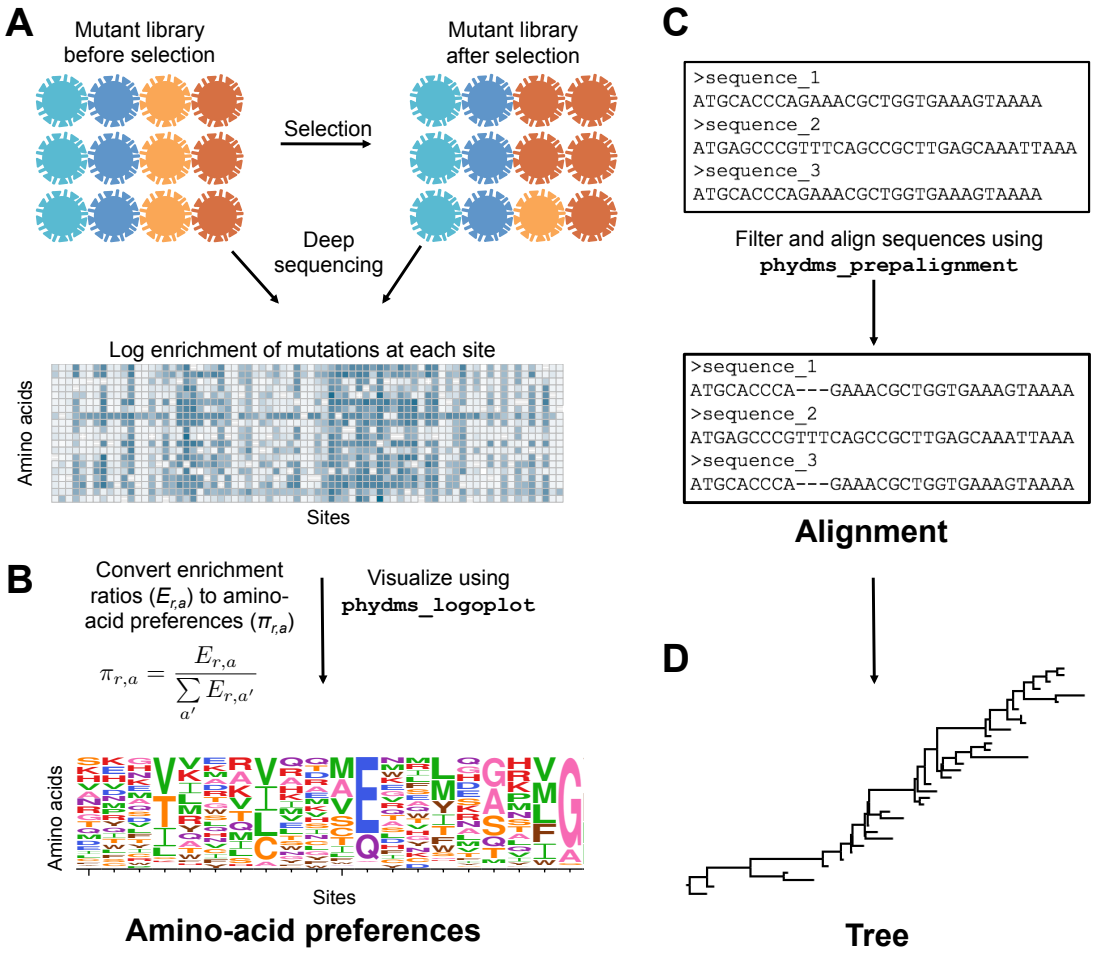

Figure 3: **Workflow for preparing input data to `phydms`.** Analysis with `phydms` requires amino-acid preferences measured by deep mutational scanning, a codon-level alignment of naturally occurring sequences, and a phylogenetic tree topology. **(A)** Deep mutational scanning involves performing a functional selection on a library of mutant genes, and using deep sequencing to quantify the enrichment or depletion of each mutation (relative to wildtype) after selection. **(B)** The amino-acid preferences used by the ExpCM can be calculated by normalizing the enrichment ratios for mutations to sum to one at each site. **(C)** We created a filtered, codon-level alignment of naturally occurring sequences using `phydms_prepalignment`. **(D)** We used `phydms_comprehensive` to automatically generate a tree topology from the filtered alignment using `RAxML`.

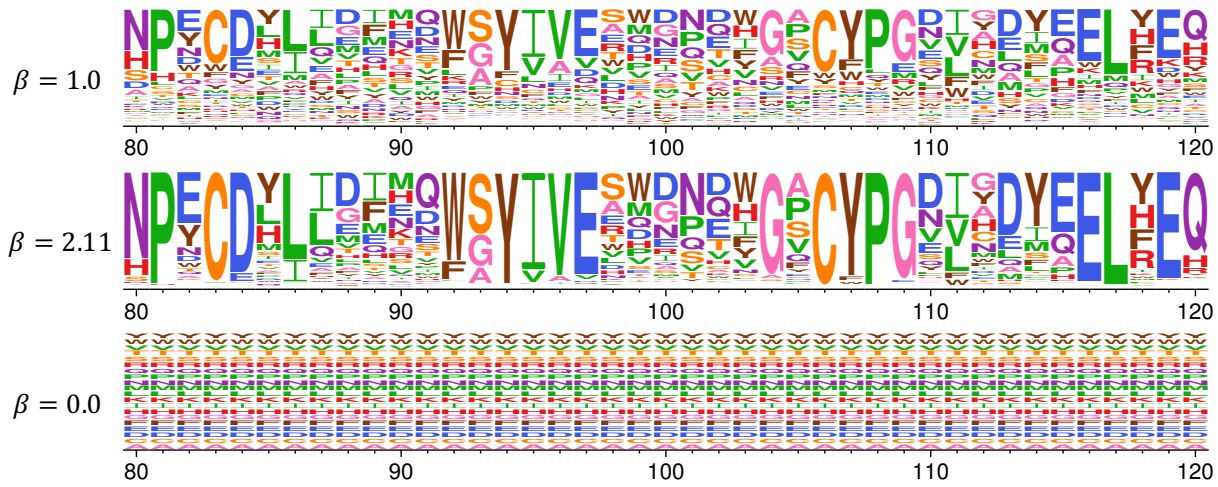

Figure 4: **Re-scaling of amino-acid preferences to reflect the stringency of selection in nature.** Analysis with `phydms` optimizes a stringency parameter β that relates the stringency of selection for preferred amino acids in the deep mutational scanning experiment to that in nature. When $\beta = 1$, the favored amino-acids are preferred in nature with the same stringency as during the experimental selections in the lab. When $\beta > 1$, selection in nature prefers the same amino acids as selection in lab but with greater stringency. When $\beta < 1$, selection in nature has less preference than the experiments for mutations favored in the lab, and when $\beta = 0$ then all site-specific information is lost. The actual optimized stringency parameter for HA reported in Table 2 is $\beta = 2.11$. We generated the logoplots shown above from the input data in Supplemental file 3 with the following commands:

```
phydms_logoplot HA_Doud_1.pdf --prefs HA_Doud_prefs_short.csv
phydms_logoplot HA_Doud_2_11.pdf --prefs HA_Doud_prefs_short.csv --stringency 2.11
phydms_logoplot HA_Doud_0.pdf --prefs HA_Doud_prefs_short.csv --stringency 0
```

Figure 5: **Identifying sites of diversifying selection.** The `phydms` option `--omegabysite` fits a site-specific value for ω_r, which gives the relative rate of non-synonymous to synonymous substitutions at site r after accounting for the selection due to the amino-acid preferences. This figure shows the results of such an analysis for HA. The overlay bar represents the strength of evidence for ω_r being greater (red) or less (blue) than one. Because this approach accounts for the constraints due to the amino-acid preferences, it can identify sites evolving faster than expected even if their absolute relative rates of nonysnonymous to synonymous substitutions do not significantly differ from one . The logoplot in this figure uses the stringency parameter value of $\beta = 2.11$, and was generated by running the following command on the data in Supplemental file 3:

```
phydms_logoplot results/omegabysite.pdf --prefs HA_Doud_prefs.csv --omegabysite
results/omegabysite.txt --stringency 2.11 --minP 0.001
```

In this figure, the HA sequence is numbered sequentially beginning with 1 for the first site with deep mutational scanning data, which is the second residue in the protein.

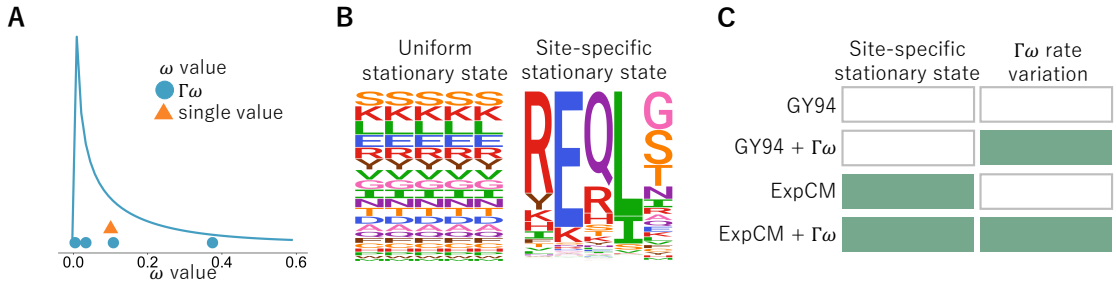

Figure 6: **Different ways codon models account for purifying selection.** (A) The dN/dS parameter, ω, can be defined as one gene-wide average (orange triangle) or allowed to vary according to some statistical distribution (blue line). For computational tractability, the distribution is discretized into K bins and ω takes on the mean of each bin (blue circles) [156, 160]. A gamma distribution (denoted by Γ) with $K = 4$ bins is shown here. (B) A substitution model's 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 Goldman-Yang [GY94; 44, 160] and experimentally informed codon model [ExpCM; 59] families with and without gamma-distributed ω to represent all possible combinations.

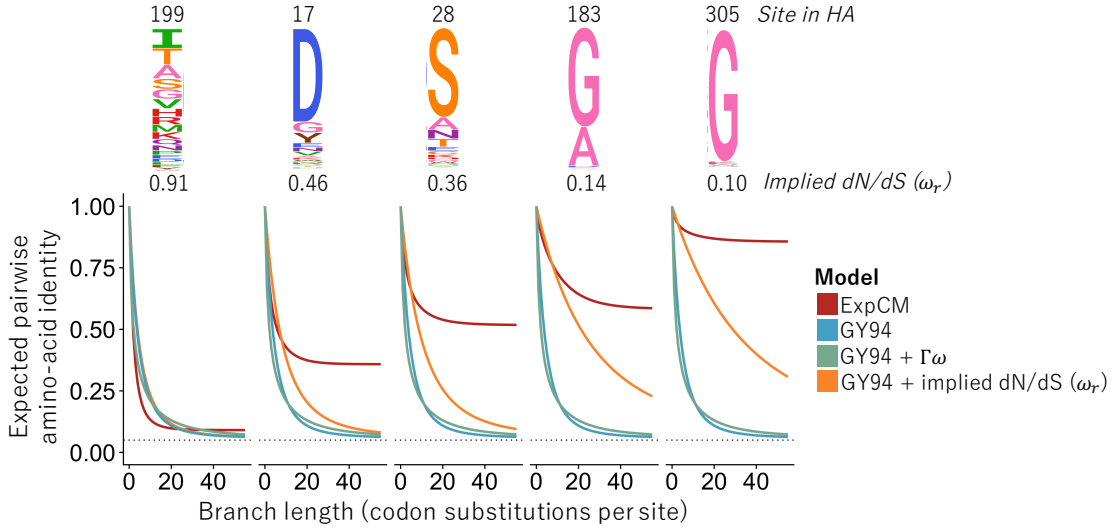

Figure 7: **Effect of stationary state and $\Gamma\omega$ rate variation on predicted asymptotic sequence divergence.** The logo plots at top show the amino-acid preferences for some sites in an H1 influenza hemagglutinin protein as experimentally measured by [27]. The graphs show the expected amino-acid identity at that site for two sequences separated by a branch of the indicated length (Equation 9). For the GY94 model, the graphs are identical for all sites since this model does not have site-specific parameters; the same is true for GY94+$\Gamma\omega$. The graphs do differ among sites if we calculate a different ω_r for each site r in the GY94 model using the amino-acid preferences [Equation 7; 124]. However, all GY94 models, including the one with site-specific ω_r values, approach the same asymptote since they all have the same stationary state. The ExpCM has different asymptotes for different sites since it accounts for how amino-acid preferences lead to site-specific stationary states.

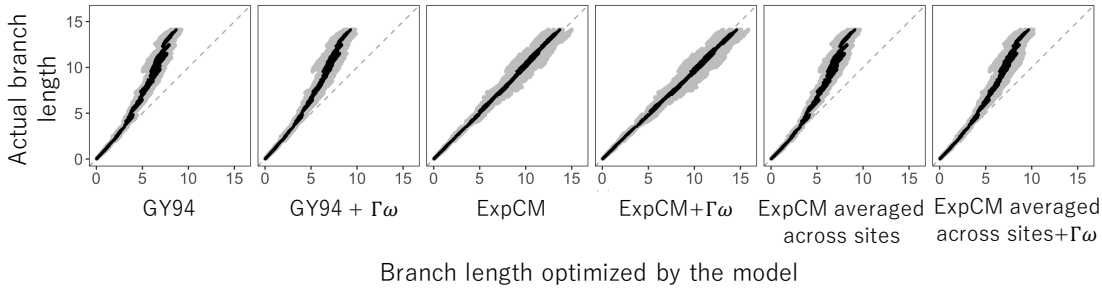

Figure 8: **Branch lengths inferred on data simulated under a model with site-specific amino-acid preferences.** We simulated alignments along a HA phylogenetic tree using an ExpCM parameterized by the actual site-specific amino-acid preferences for an H1 HA [27]. We then inferred the branch lengths of this tree from 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 what would be seen if the inferred branch lengths exactly matched those used in the simulations.

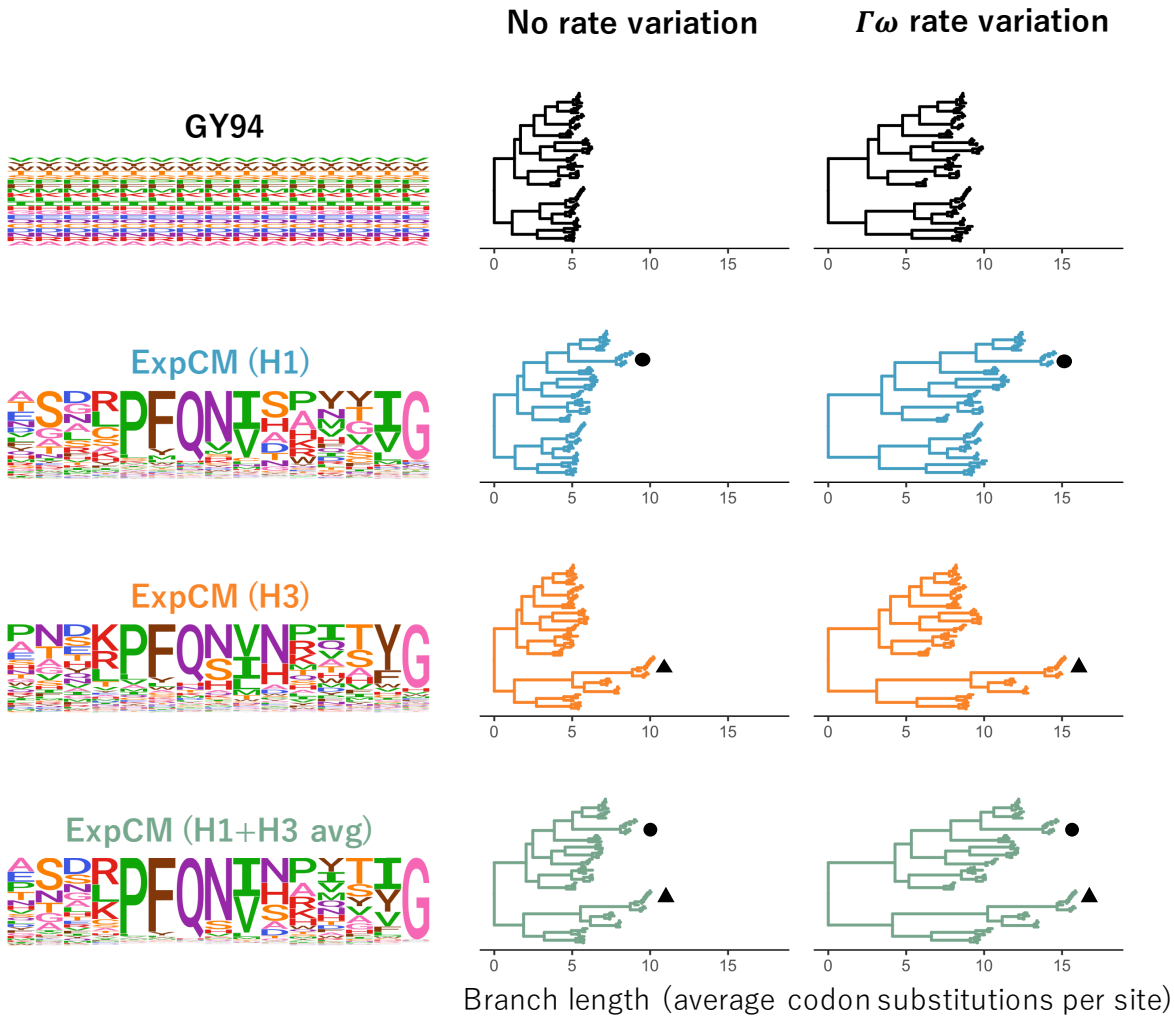

Figure 9: **Effect of site-specific amino-acid preferences and $\Gamma\omega$ rate variation on HA branch length estimation.** The branch lengths of the HA tree were optimized using the indicated ExpCM or GY94 model. The amino-acid preferences defining the model (ExpCM) or implied by the model (GY94) are shown as logo plots for 15 sites in HA; the full set of experimentally measured amino-acid preferences are in 13, 14, and 15. The ExpCMs use amino-acid preferences measured in deep mutational scanning of an H1 HA [27], an H3 HA [75], or the average of the measurements for these two HAs. Circle denotes the H1 clade and triangle denotes the H3 clade. The root of each tree is placed where it would fall if the tree was midpoint rooted using the branch lengths inferred by `RAxML` using the GTRCAT model. This figure enables qualitative visualization of the trees; for a quantitative comparison of branch lengths optimized by different models, see Figure 10.

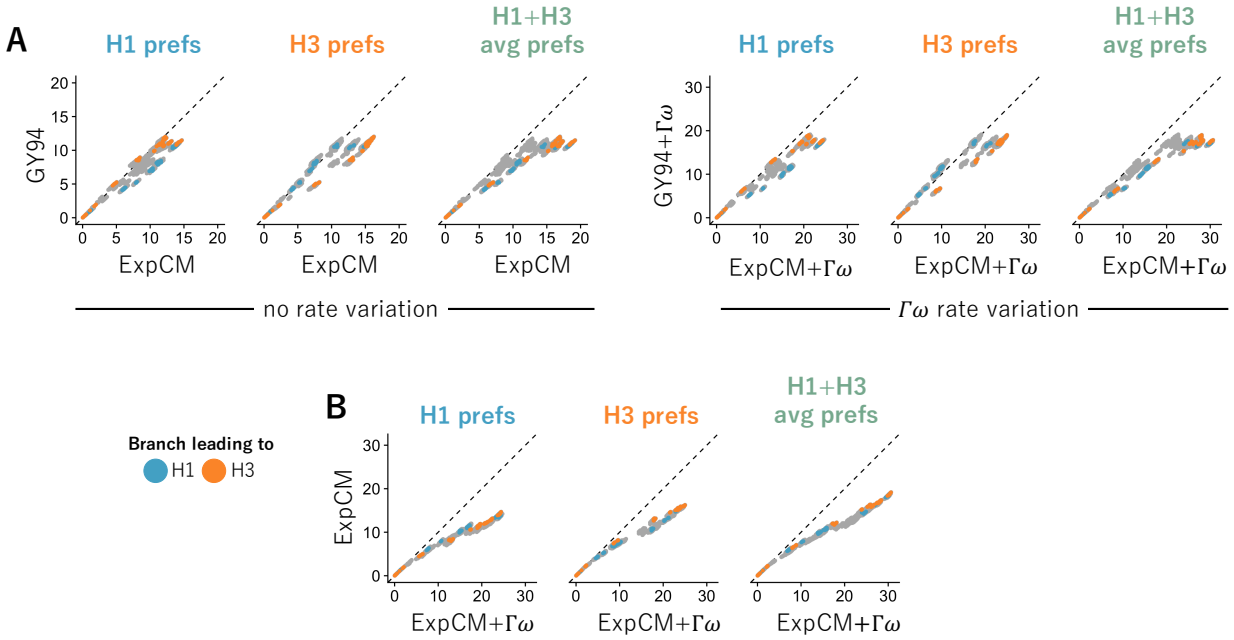

Figure 10: **Modeling site-specific amino-acid preferences using deep mutational scanning experiments extends branch lengths, especially for branches leading to the HA used in the experiment.** The points indicate the total length of branches separating all pairs of tips on the HA phylogenetic tree when the tree is optimized under the indicated model. Blue and orange denote branches that lead to the H1 and H3 HAs used in the deep mutational scanning. The amino-acid preference set defining the ExpCM is labeled above each each plot. (A) ExpCMs defined by amino-acid preferences from any of the deep mutational scanning experiments estimate generally longer branches than the GY94 model, with the increase particularly pronounced for branches leading to the HA used in the experiment. (B) The addition of $\Gamma\omega$ rate variation further extends branch lengths, without any apparent bias towards the HAs used in the experiment. Note that this figure shows the same data as Figure 9 in a different form.

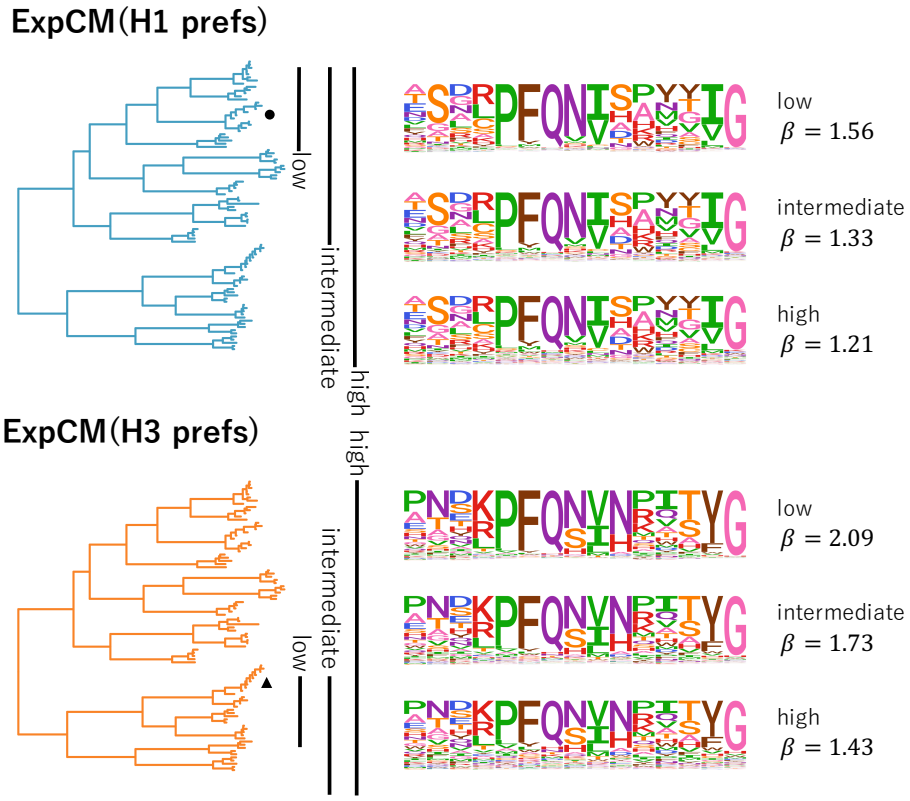

Figure 11: **The congruence between natural selection and the deep mutational scanning measurements decreases with sequence divergence.** We fit an ExpCM informed by the H1 or H3 deep mutational scanning experiments to trees spanning sequences with low, intermediate, and high divergence from the sequence used in the experiment. The ExpCM stringency parameter (β) is a measure of the congruence between natural selection and the experimental measurements [12, 59]. Larger values of β indicate that natural selection prefers the same amino acids as the experiments but with greater stringency. As divergence increases between the HA used in the experiment and the other sequences in the tree, the β value decreases and the amino-acid preference "flatten." Therefore, the preferences measured in each experiment are progressively less congruent with natural selection as we include increasingly diverged sequences.

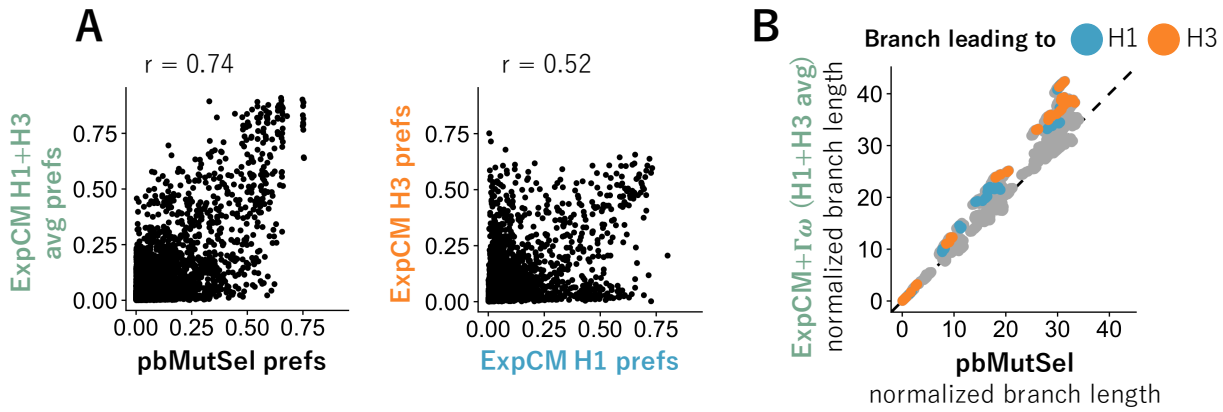

Figure 12: **Models inferred from natural sequences have similar stationary states to models defined by experimental preferences and estimate similar branch lengths.** We fit an ExpCM(H1+H3 avg)+$\Gamma\omega$ and a pbMutSel to the full HA tree in Figure 9. The pbMutSel amino-acid preferences are inferred from the natural HA sequences, while the ExpCM amino-acid preferences are experimentally measured and then rescaled by the stringency parameter in Table 7. (A) The pbMutSel preferences are more correlated with the re-scaled average of the H1 and H3 deep mutational scanning preferences than the individual re-scaled H1 and H3 deep mutational scanning preferences are to each other (Pearson's r: 0.74 versus 0.52). (B) The ExpCM(H1+H3 avg)+$\Gamma\omega$ and pbMutSel models estimated similar branch lengths when fit to the entire HA tree. Points denote branch lengths between all pairs of tips on the tree. Blue and orange denote branches that lead to the H1 and H3 deep mutational scanning reference sequences respectively. The `phydms` program implementing ExpCMs and the `PhyloBayes-MPI` program implementing pbMutSel models give branch lengths in different units, so to facilitate direct comparison between the models, we have normalized all branch lengths returned by each program by the length of the branches separating the earliest (A/South Carolina/1918) and latest (A/Solomon Islands/2006) seasonal human H1 sequences on the tree.

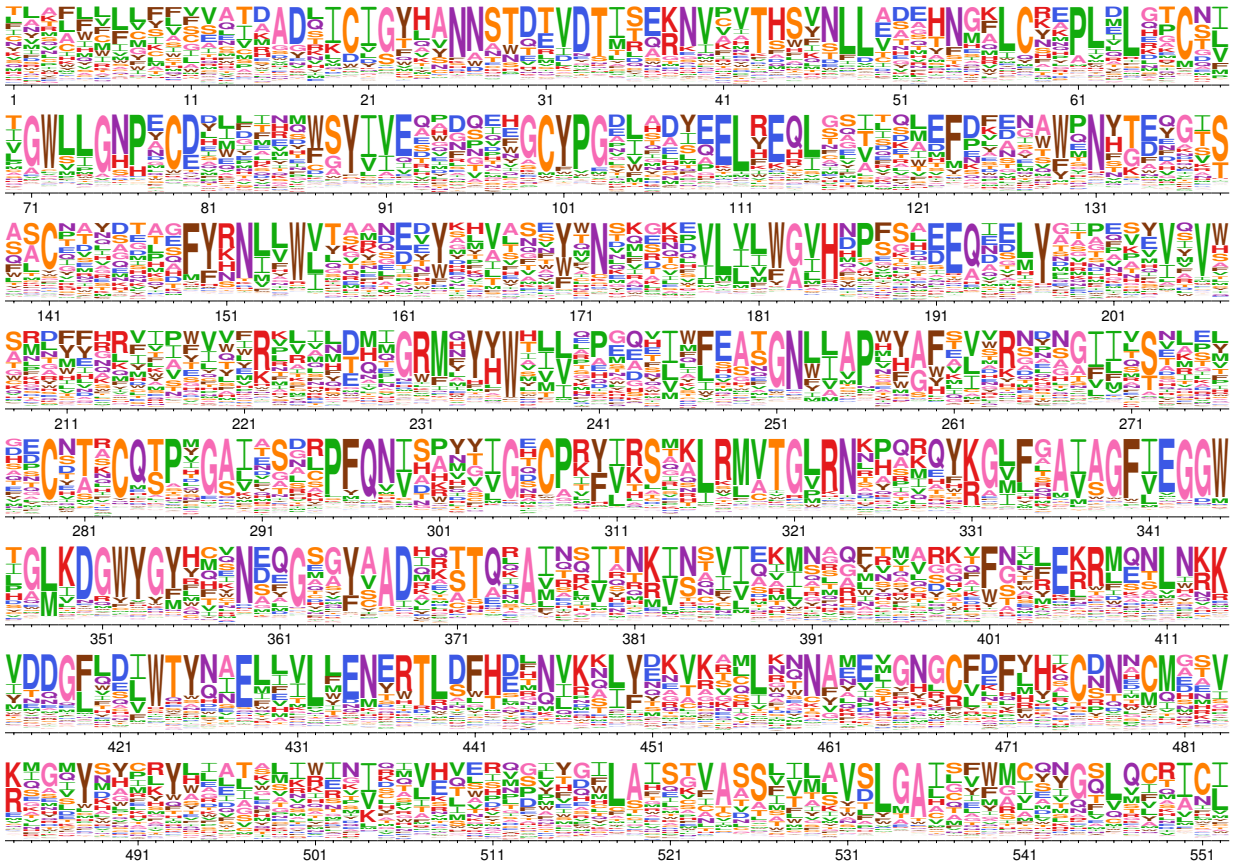

Figure 13: **H1 HA amino-acid preferences measured by deep mutational scanning.** Each column represents a site in the HA protein, and the height of each letter is proportional to the preference for the amino acid measured by [27] and then re-scaled by the stringency parameter in Table 7. The plot only shows sites that are alignable between the H1 and H3 HAs, and these alignable sites are numbered sequentially starting from 1. The conversion between the numbering scheme in this figure and sequential numbering of the H1 HA reference sequence is in 9.

Figure 14: **H3 HA amino-acid preferences measured by deep mutational scanning.** Similar to 13 but shows the re-scaled preferences for the H3 HA as measured by [75].

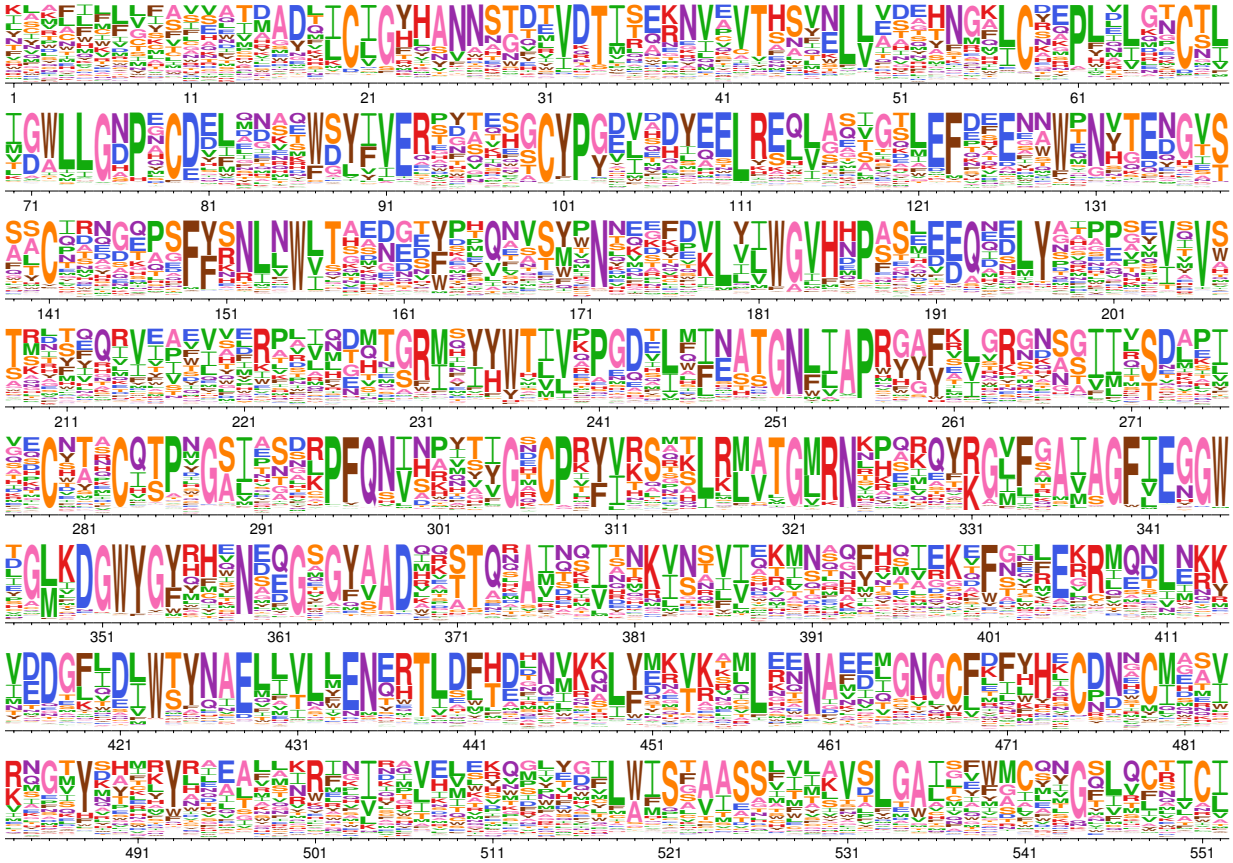

Figure 15: **Average of H1 HA and H3 HA amino-acid preferences measured by deep mutational scanning.** Similar to 13 but shows the re-scaled average of the preferences for the H1 and H3 HAs.

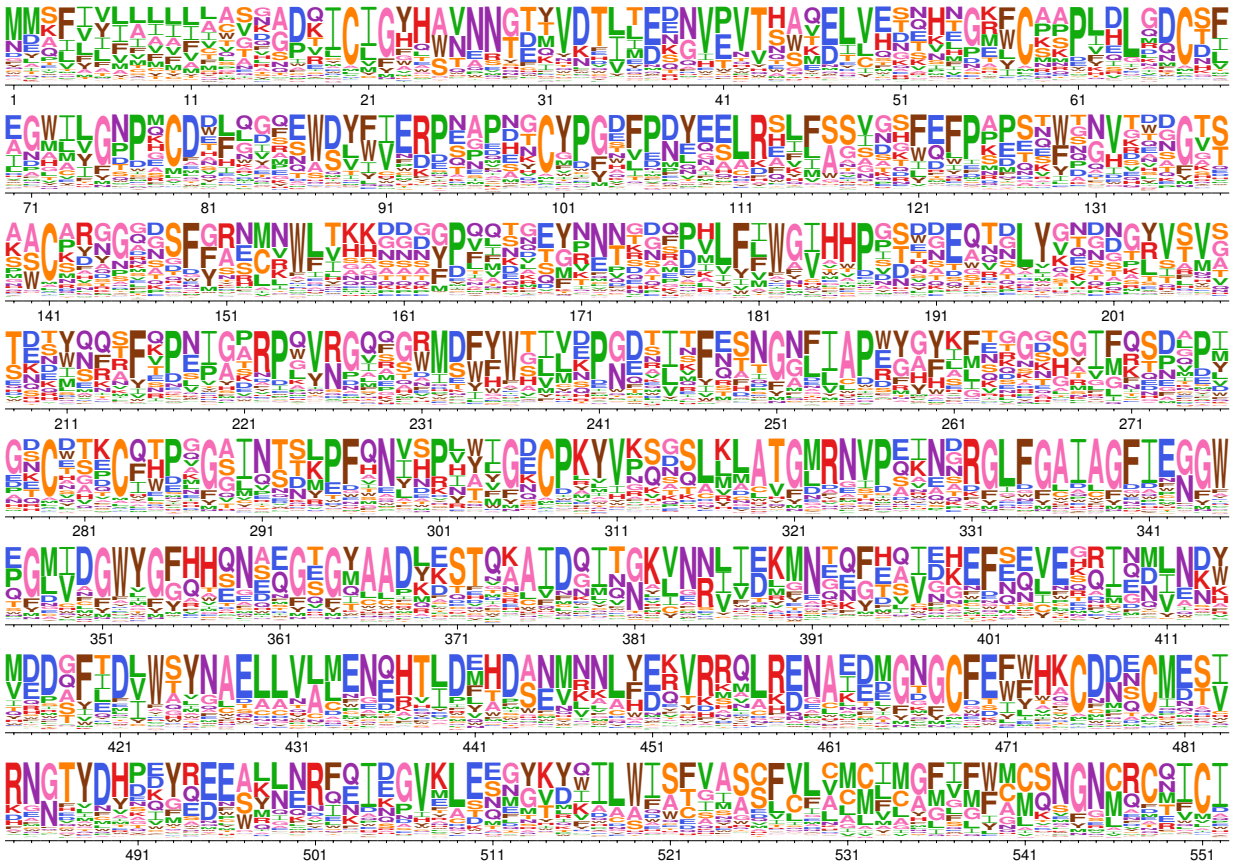

Figure 16: **Amino-acid preferences inferred by the pbMutSel model.** Similar to 13, but shows the preferences inferred by fitting the pbMutSel model to the full HA tree.

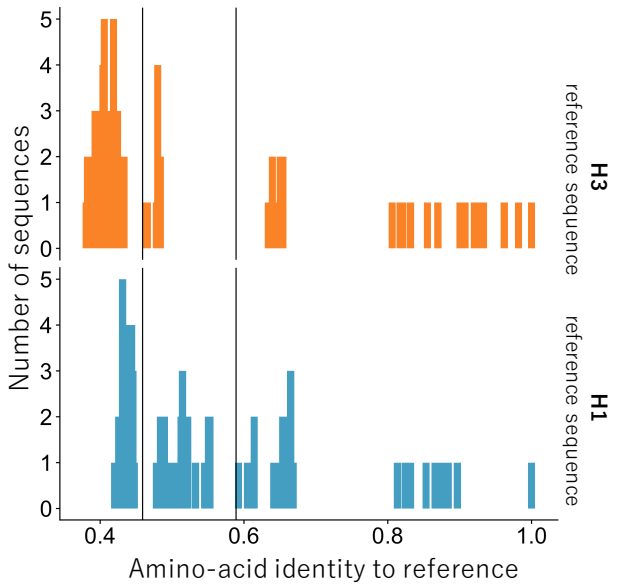

Figure 17: **Overall divergence for subtrees.** We created two subalignments for each HA used in the deep mutational scanning experiments. The "low divergence" alignments had $\geq$59% amino-acid identity to either the H1 or H3 reference sequence. The "intermediate divergence" alignments had $\geq$46% amino-acid identity to the reference sequences.

LIST OF TABLES

Table Number Page

Table 1: **Alignments and deep mutational scanning (DMS) studies for HA and β-lactamase.**

gene	DMS studies	residues in protein	residues with DMS data	sequences in alignment
HA	[27], [138]	565	564	34
β-lactamase	[132], [39]	285	263	50

Table 2: **Fitting of an ExpCM informed by the HA preferences from [27] to natural sequences using** `phydms_comprehensive`**.** Full code, data, and results are in Supplemental file 2.

model	$\triangle$AIC	log likelihood	number of parameters	parameter values
ExpCM	0.0	-4877.7	6	β=2.11, κ=5.14, ω=0.52
ExpCM averaged preferences	2090.6	-5922.9	6	β=0.68, κ=5.36, ω=0.22
YNGKP_M5	2113.5	-5928.4	12	α_ω=0.30, β_ω=1.42, κ=4.68
YNGKP_M0	2219.6	-5982.5	11	κ=4.61, ω=0.20

Table 3: **Comparison of multiple β-lactamase deep mutational scanning results using** `phydms_comprehensive`**.** Full code, data, and results are in Supplemental file 4.

model	ΔAIC	log likelihood	number of parameters	parameter values
ExpCM Stiffler preferences	0.0	-2581.3	6	β=1.31, κ=2.67, ω=0.72
ExpCM Firnberg preferences	96.2	-2629.4	6	β=2.42, κ=2.60, ω=0.63
YNGKP_M5	739.2	-2944.9	12	α_ω=0.30, β_ω=0.49, κ=3.02
YNGKP_M0	841.0	-2996.8	11	κ=2.39, ω=0.28

Table 4: **Comparison of multiple HA deep mutational scanning results using** `phydms_comprehensive`. Full code, data, and results are in Supplemental file 5.

model	$\triangle$AIC	log likelihood	number of parameters	parameter values
ExpCM Doud preferences	0.0	-4877.7	6	β=2.11, κ=5.14, ω=0.52
ExpCM Thyagarajan preferences	44.2	-4899.7	6	β=1.72, κ=4.94, ω=0.55
YNGKP_M5	2113.5	-5928.4	12	α_ω=0.30, β_ω=1.42, κ=4.68
YNGKP_M0	2219.6	-5982.5	11	κ=4.61, ω=0.20

Table 5: **Comparison of `phydms` to alternative software for optimizing a tree of 34 HA sequences.** HyPhy and `Bio++` use models that fit ϕ, whereas by default `phydms` determines ϕ_w empirically. Log likelihoods are not expected to be identical across software. Full code, data, and results are in Supplemental file 7.

software	runtime (minutes)	log likelihood	β	ω
`phydms`, scale branches	7.8	-4877.9	2.11	0.52
`phydms`, default settings	10.5	-4877.7	2.11	0.52
`phydms`, fit ϕ values	23.2	-4876.5	2.11	0.53
`phydms`, no gradient	52.8	-4894.0	2.13	0.57
`Bio++` via old `phydms`	962.6	-4880.6	2.09	0.53
`HyPhy` via `phyloExpCM`	2102.0	-4908.4	2.11	0.57

Table 6: **Comparison of parameter values and runtimes for HA alignments of different sizes using default `phydms` settings.** The alignments are different than those used for the other HA analyses in this paper thus explaining the slightly different parameter values. The alignments, full code, data, and results are in Supplemental file 8.

sequences in alignment	runtime (minutes)	β	ω
34	14.5	1.97	0.42
62	37.2	1.92	0.45
85	41.0	1.87	0.48
104	51.2	1.87	0.49

Table 7: **Fitting of substitution models to the HA phylogenetic tree. All ExpCMs describe the evolution of HA better than the GY94 models, as evaluated by the Akaike information criteria [ΔAIC, 100]. The models fit here are the same ones in Figure 9. The ω value for each of the $K = 4$ bins is shown for the models with $\Gamma\omega$ rate variation. All ExpCMs fit a stringency parameter > 1.**

Model	ΔAIC	Log Likelihood	ω	Stringency parameter (β)
ExpCM (H1+H3 avg) + $\Gamma\omega$	0	-51083	0.19, 0.50, 0.91, 1.86	1.69
ExpCM (H1+H3 avg)	1063	-51616	0.14	1.77
ExpCM (H1) + $\Gamma\omega$	1321	-51744	0.12 , 0.42, 0.89, 2.13	1.11
ExpCM (H3) + $\Gamma\omega$	1777	-51972	0.10, 0.36, 0.76, 1.84	1.28
ExpCM (H1)	2670	-52419	0.12	1.21
ExpCM (H3)	3377	-52773	0.12	1.43
GY94 + $\Gamma\omega$	4817	-53487	0.00, 0.03, 0.08, 0.24	-
GY94	7892	-55025	0.07	-

Table 8: **Branch length extension as measured by tree diameter. We calculated the tree diameter, the distance between the two most divergent tips, for the trees in Figure 9. For each tree, the diameter is reported as a raw value and as a percentage of the GY94 model tree, the smallest of the eight trees.**

Model	Tree diameter (average codon substitutions per site)	Percentage of GY94 tree diameter
GY94	12.04	100%
ExpCM(H1)	14.70	122%
ExpCM(H3)	16.28	135%
ExpCM(H1+H3 avg)	19.21	160%
GY94 + $\Gamma\omega$	19.15	159%
ExpCM(H1) + $\Gamma\omega$	24.75	206%
ExpCM(H3) + $\Gamma\omega$	25.03	208%
ExpCM(H1+H3 avg) + $\Gamma\omega$	30.78	256%

Table 9: **Model parameters fit to a low divergence tree. We fit GY94 models and an ExpCM defined by H1 deep mutational scanning preferences to the "low divergence from H1" tree in Figure 11. We used these model parameters calculate the expected pairwise sequence identity in Figure 7 and simulate the sequences in Figure 8.**

Model	Parameters
GY94	$\kappa = 3.17$, $\omega = 0.10$, $\phi_{1,A} = 0.32$, $\phi_{1,C} = 0.14$, $\phi_{1,G} = 0.28$, $\phi_{2,A} = 0.38$, $\phi_{2,C} = 0.18$, $\phi_{2G} = 0.20$, $\phi_{3,A} = 0.36$, $\phi_{3,C} = 0.19$, $\phi_{3,G} = 0.21$
GY94 + $\Gamma\omega$	$\alpha_\omega = 0.51$, $\beta_\omega = 3.92$, $\kappa = 3.49$, $\phi_{1,A} = 0.32$, $\phi_{1,C} = 0.14$, $\phi_{1,G} = 0.28$, $\phi_{2,A} = 0.38$, $\phi_{2,C} = 0.18$, $\phi_{2,G} = 0.20$, $\phi_{3,A} = 0.36$, $\phi_{3,C} = 0.19$, $\phi_{3,G} = 0.21$
ExpCM(H1)	$\beta = 1.56$, $\kappa = 3.64$, $\omega = 0.24$, $\phi_A = 0.378$, $\phi_C = 0.17$, $\phi_G = 0.23$