Model fit does not predict accuracy in single-gene protein phylogenetics

Stephanie J. Spielman[1]

[1]Department of Biological Sciences, Rowan University, Glassboro, NJ, 08028, USA
*Corresponding author: spielman@rowan.edu

Abstract

When reconstructing a phylogeny from molecular data, it is regarded as best practice to perform relative model selection to determine the most appropriate evolutionary model for the data at hand. This procedure ranks all available evolutionary models by a certain theoretic information criterion (e.g. Akaike Information Criterion, AIC), and the best-fitting model is then specified for phylogenetic inference. While it is often assumed that using better-fitting models results in more accurate inferences, there is in fact no guarantee that this assumption will hold. Indeed, recent studies have observed that the specific model employed for phylogenetic construction may not have substantial effects on the inferred tree topologies. In this study, we examine whether there is a systematic relationship between model fit and inference accuracy, in the specific context of topological accuracy in protein phylogenetics, using both simulation studies as well as analysis of natural sequence data. Notably, the simulations performed here employ a robust codon-level generative approach that is fully distinct from all protein-level inference models. We broadly find that phylogenies inferred across a range of models with vastly different fits to the data yield highly consistent topologies. We additionally find that well-fitting and poorly-fitted models alike have similar potential to infer strongly-supported but incorrect nodes, raising the possibility that all available models of protein evolution may be systematically misspecified. Moreover, we find that the GTR model, where amino-acid exchangeabilities are treated as free parameters, performs very similarly to models with fixed substitution rates and appears reasonable to use in protein phylogenetics in spite of its tendency to overfit sequence data. In sum, this work builds on a growing body of literature finding that relative model selection does not necessarily guarantee increased accuracy in phylogenetic inference, and therefore may not be a critical step in analysis in spite of decades of tradition.

Keywords: phylogenetics, protein models, relative model selection, maximum likelihood

Introduction

When analyzing sequence data in evolutionarily-aware contexts, and in particular when inferring phylogenetic trees using modern statistical approaches, researchers must select an appropriate evolutionary model. The most common modeling framework for such applications follow a continuous-time Markov process, considering either nucleotides, codons, or amino-acids as states (Yang 2014; Arenas 2015). Since this framework's introduction, a wide array of model parameterizations have been developed, ranging in complexity from the simple equal-rates Jukes-Cantor (JC) model (Jukes and Cantor 1969) where substitution rates among all states are equal, to the most complex form where all substitution rates are distinct (Tavare 1984). Additional levels of complexity beyond a model's core substitution rates, such as incorporating among site rate variation, further increase the number of models from which practitioners can choose (Yang 2014).

To choose among dozens, if not hundreds, of available model formulations, the field has largely converged upon a strategy of *relative model selection*. For a given multiple sequence alignment, this approach systematically evaluates the statistical fit to the data for all possible models using various metrics, most commonly theoretic information criteria (Posada and Buckley 2004) (although see refs. Sullivan and Joyce (2005); Abadi et al. (2019) for other frameworks). Such criteria include, for example, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), which provide a measure of goodness-of-fit to the data while penalizing models with excessive parameters that could lead to overfitting (Sullivan and Joyce 2005). Once available models are ranked by a given criterion, the model with the best fit to the data is subsequently specified during phylogenetic inference.

First popularized by the seminal software `MODELTEST` over 20 years ago (Posada and Crandall 1998), many different computational approaches to performing model selection on different types of data (i.e., nucleotide, protein, binary, etc.) have been developed, with continued development as recently as 2019 (Darriba et al. 2011, 2012; Whelan et al. 2015; Kalyaanamoorthy et al. 2017; Darriba et al. 2019). Alongside this popularization has emerged a near-dogmatic mentality that employing the best-fitting model will increase the reliability, and potentially the accuracy, of inferences. Indeed, model selection has been described as "an essential stage in the pipeline of phylogenetic inference" (Arenas 2015) and is often viewed as a panacea to avoid model misspecification and biased inferences. As a result, it is trivial to find casual and misleading remarks about the role of model selection across biological research fields. For example, the popular online database for HIV sequences, HIV LANL (`https://www.hiv.lanl.gov/`), contains an analysis option "FindModel" to perform model selection on sequence data, leading with the header "Purpose: FindModel analyzes your alignment to see which phylogenetic model best describes your data; **this model can then be used to generate a better tree**" (emphasis added; `https://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html`).

In spite of this pervasive attitude, there is no guarantee that the best-fitting model will infer the most accurate phylogenies. Indeed, relative model selection is inherently unable to determine whether a given model is reasonable to use in the first place. To circumvent this drawback, many have advocated for a shift in focus towards absolute model selection methods, or similarly tests of model adequacy (Bollback 2002; Brown 2014; Brown and Thomson 2018). Such approaches, which include analysis of posterior predictive distributions or hy-

pothesis tests that ask whether the given inference model produces molecular properties that mirror those seen in the data at hand (Goldman 1993a,b; Ripplinger and Sullivan 2010; Gelman et al. 2013; Brown 2014; Duchne et al. 2015, 2016; Bollback 2002; Brown 2014; Duchne et al. 2015; Hhna S. 2017; Brown and Thomson 2018). Inference models that generate data with molecular features consistent with the natural sequences being analyzed, the inference model is considered adequate for the data at hand. Despite recommendations that model adequacy approaches should be used in conjunction with relative model selection, they have yet to see widespread adoption in large part due to their time-consuming, computationally-intensive nature (Duchne et al. 2015; Brown and Thomson 2018). As such, the "average" researcher performing phylogenetic reconstruction will most often rely on relative model selection to justify use of a selected model.

Several lines of recent research have questioned the reliability of relative model selection in phylogenetic contexts. For example, Spielman and Wilke (2015b) showed that, in the context of identifying selection pressures from sequence data using codon models, AIC and BIC strongly prefer models with systematic biases, and models that in fact mitigate inference biases have relatively poorer fit to the data. Keane et al. (2006) observed that selecting protein models based on *ad-hoc* assumptions of biological relevance to the data may not result in improved inferences. Similarly, Spielman and Kosakovsky Pond (2018) found that the model employed when estimating site-level evolutionary rates in protein alignments has little, if any, effect on the inferred rates, with the primary exception that the simple equal-rates Jukes Cantor (JC) model has unique and previously unrecognized power to identify rapidly-evolving sites.

In the context of phylogenetic inference specifically, several studies have been conducted to investigate the consequences of employing different model selection criteria on nucleotide data. Ripplinger and Sullivan (2008) showed that, while different criteria choose different models, resulting phylogenies are not significantly different from one another, with most differences occurring at poorly-supported nodes. Most recently, Abadi et al. (2019) echoed and extended this insight to show that phylogenies inferred with the most complex nucleotide-level model (GTR) did not differ significantly from the entirely uninformative JC model, even though JC generally shows poor fits to most datasets. In total, these studies have suggested that model selection itself may either be unnecessary or inadvertently lead to high confidence in biased results. A thorough examination of the practical ramifications of relative model selection is merited to reconcile these recent findings with the overarching sensibility that model selection is a fundamentally necessary component of phylogenetic analysis.

In this work, we explore whether there exists a systematic relationship between model fit and inference accuracy, such that inferences performed with better-fitting models consistently lead to more accurate results, and similarly inferences performed with poorly-fitting models consistently lead to less accurate results. We specifically study this question in the context of phylogenetic inference from protein data, using both simulations and natural sequences. Protein models are unique phenomenological compared to codon-level and nucleotide-level because amino acids themselves do not evolve. Rather, the underlying DNA sequences evolve, and nonsynonymous changes induce protein-level substitutions. From biological first principles, then, there is no way to describe the actual mechanism of evolution when only protein data is available.

The simplest protein model, under the general time reversible framework, is described by

a continuous-time Markov process with an instantaneous rate matrix, for the substitution amino acid i to j, $Q_{ij} = r_{ij}\pi_j$ scaled such that $-\sum_{i=1} \pi_i Q_{ii} = 1$. In this model, parameters r_{ij} describe the substitution rate, or exchangeabilities, between amino acids i and j, and π_j represents the stationary frequency of target amino acid j (Yang 2014; Arenas 2015). These exchangeabilities represent the average propensity of each type of amino acid substitutions. As there are 189 such free parameters, assuming symmetric exchangeabilities, these values are rarely estimated from a given alignment itself. Instead, empirically-derived models with fixed exchangeabilities that have been *a priori* derived from hundreds or thousands of training datasets, using an approach pioneered by Jones et al. (1992) which produced the standard JTT model, are most commonly employed (Arenas 2015). Such models include the commonly used WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008) general amino acid models, as well as certain specialist amino acid models like the chloroplast-sequence–derived cpREV model (Adachi et al. 2000) or Influenza-sequence–derived FLU model (Dang et al. 2010). Unlike exchangeability parameters, the π_j parameters are more often estimated from the alignment at hand, either optimized during phylogenetic reconstruction or directly obtained by counting the amino acids in the alignment, known as the $+F$ parameterization (Yang 2014).

While this modeling framework has become the default analysis choice for most users constructing trees from protein sequences, it ignores heterogeneous site-specific evolutionary constraints which are known to dominate protein evolution (Echave et al. 2016). While it is possible to incorporate among site rate variation, generally by scaling individual site rates according to a discrete Gamma distribution or similar (Yang 2014), these models assume the same evolutionary pattern governs each site in a given unpartitioned alignment. Other modeling approaches have been developed to more directly account for the pervasive heterogeneity in protein evolution, such as the Bayesian CAT model in `PhyloBayes` (Lartillot and Philippe 2004; Le et al. 2008) or mixture models which consider a distribution of individual matrices (Le et al. 2012; Arenas 2015). In spite of their known benefits, these models' computational complexity and resource requirements have somewhat limited their adoption as the standard modeling framework standard in protein phylogenetics. For example, while the Bayesian CAT model is well-suited for long, multi-gene alignments, the underlying MCMC sampler generally cannot accommodate more than $\sim$ 100 taxa, ultimately restricting the CAT model's utility to phylogenomic analyses on a small number of taxa (`http://megasun.bch.umontreal.ca/People/lartillot/www/phylobayes4.1.pdf`). Therefore, in this study, we focus on the more widely-used single matrix protein exchangeability models.

Overall, we do not observe a strong, systematic relationship between relative model fit to the data and inference accuracy. Except for the most severely misspecified models, protein models with drastically different fits to a given dataset produce surprisingly consistent phylogenies that, in simulations, generally are close estimates to the true tree. Our analysis additionally considers the merits of the protein-level GTR model, which optimizes all 189 free exchangeability parameters to the data. This framework is generally not considered usable on protein sequence data due its excessive number of free parameters to optimize, generally leading it to overfit most datasets. In fact, we find that GTR performs comparably to protein models with fixed exchangeability parameters, regardless of its relative fit to a given dataset, and in certain circumstances is uniquely able to identify the true phylogeny. We therefore suggest that this often-ignored model may in fact be more viable than is generally assumed.

Methods and Materials

Data Preparation

Simulations were performed along eight phylogenies obtained from the literature (Table 2) using the site-wise mutation–selection modeling framework (Halpern and Bruno 1998) in the Python library `pyvolve` (Spielman and Wilke 2015a). In this codon-level model, each site evolves according to a distinct set of fitness parameters, and mutation rates are shared across all sites. For all simulations, we assumed equal and symmetric mutation rates among nucleotides, and we assumed all synonymous codons shared the same fitness with their corresponding amino acid. We obtained site-specific fitness parameters for simulations using one of three sets amino-acid propensities measured experimentally using deep-mutational scanning (DMS), as described in Table 1. We simulated 20 replicates for each parameterization along each phylogeny, resulting in a total of 480 simulated alignments. Because simulations were performed at the codon level, simulated alignments were translated before subsequent analysis. To ensure that the input branch lengths along phylogenies used for simulation represented expected number of amino-acid substitutions, rather than nucleotide substitutions, all phylogeny branch lengths were scaled up by a factor of 3 during simulation.

All empirical protein datasets were collected from the obtained from the PANDIT database (Whelan 2006). 200 alignments (and corresponding PANDIT phylogenies) were randomly chosen from all PANDIT families where the "PANDIT-aa-restricted" set of sequences contained between 20–500 (inclusive) sequences with between 100–1000 sites (inclusive).

Phylogenetic inference and analysis

All model selection, phylogenetic inference, and topology tests were performed in `IQ-TREE v1.6.7`, unless otherwise stated (Nguyen et al. 2015). Model selection was performed on each alignment using the `ModelFinder` (Kalyaanamoorthy et al. 2017) algorithm in `IQ-TREE`, specifying the argument `-m TESTONLY` to match commonly-used programs like `ProtTest` (Darriba et al. 2011). Approximately unbiased (AU) topology tests (Shimodaira 2002) described were performed in `IQ-TREE` by specifying the argument `-zb 10000 -au` to perform 10,000 RELL replicates (Kishino et al. 1990). Calculation of Robinson-Foulds distance and other topological comparisons were performed using the Python library `dendropy v4.4.0` (Sukumaran and Holder 2010).

Statistical analysis and availability

All statistical analysis and visualization was performed in `R`, making use of the `ggplot2` visualization library and `tidyverse` analysis framework (R Core Team 2017; Wickham 2016, 2017). All data and code used is freely available from `https://github.com/spielmanlab/aa_phylo_fit_topology` and **will be deposited to Zenodo upon acceptance of the manuscript**.

Results and Discussion

Simulation Approach

To begin, we adopted a simulation-based approach to assess whether employing models of different fit induce systematic shifts in the accuracy of inferred phylogenetic topologies. While simulation is a powerful tool for studying the power and limitations of statistical methods, the procedure is often highly confounded: A model must be employed to simulate data (generative model), but of course a model must also be chosen to analyze the data (inference model). If the inference model is accurately specified, we can expect strong inference model performance, particularly when using a consistent method such as maximum-likelihood (ML) estimation (Self and Liang 1987). However, when the generative and inference models correspond closely, it is easy to become overconfident about a model's performance. Indeed, in real data analysis, any model used will be misspecified to a degree, some more than others.

Therefore, to ensure that insights gained from simulations are as applicable to empirical data analysis, it is key to examine how models perform when the model is misspecified to the data. Such approaches have previously been shown, in evolutionary sequence analysis, to reveal unrecognized performance behaviors or biases in inference methods which would go unnoticed if the data met all model assumptions (Holder et al. 2008; Spielman and Wilke 2015b; Spielman et al. 2016; Spielman and Wilke 2016; Jones et al. 2016, 2018). Therefore, we employ the codon-level mutation–selection model for all simulations (Halpern and Bruno 1998). The instantaneous rate matrix for this model at a given codon site k is specified as

$$q_{ij}^{k} = \begin{cases} \mu_{ij} \frac{S_{ij}^{k}}{1-\exp(S_{ji}^{k})} & \text{single nucleotide change} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \quad (1)$$

for a substitution from codon i to j, where μ_{ij} is the site-invariant nucleotide-level mutation rate, and S_{ij}^{k} is the scaled selection coefficient for site k, which represents the fitness difference between codons j and i at site k. Each site k in a given alignment is specified by a unique 61-length vector of codon fitness values (excluding the three stop codons). Therefore, while both empirical protein models and the mutation–selection model follow general time-reversible Markov framework, they employ entirely distinct focal parameters: The protein models used for phylogenetic inference consider a site-invariant matrix of phenomenological exchangeabilities among amino-acids, while the mutation–selection models used for simulation consider site-specific codon fitness profiles coupled with nucleotide-level mutation rates.

Table 1: Deep mutational scanning datasets used for simulation.

Name	Description	Number of Sites	Reference
NP	Influenza H1N1 nucleoprotein	497	Bloom (2014); Doud et al. (2015)
HA	Influenza H1N1 hemagglutinin	564	Thyagarajan and Bloom (2014)
HIV	HIV-1 Env Protein	661	Haddox et al. (2018)

To ensure that each simulation reflected evolutionary heterogeneity seen in real proteins, we obtained simulation parameters from three sets of site-level codon fitness values experi-

mentally determined using deep-mutational scanning (DMS) (Table 1), assuming the same fitness for synonymous codons. For each of these three parameterizations, we simulated 20 replicate alignments across eight different empirical phylogenies (Table 2) using the Python library `pyvolve` (Spielman and Wilke 2015a). Each simulated alignment therefore captured the site-specific fitness landscape of the respective protein whose DMS-derived parameters were employed. We translated all alignments to amino acids before all subsequent analysis.

Table 2: Empirical trees used for simulation.

Name	Number of taxa	Tree length*	Reference
Green plants	360	24.67	Ruhfel et al. (2014)
Ray-finned fish	305	29.44	Hughes et al. (2018)
Placental Mammals	274	13.88	dos Reis et al. (2012)
Aves	200	5.21	Prum et al. (2015)
Lassa virus	179	6.62	Andersen et al. (2015)
Spiralia	103	25.01	Marltaz et al. (2019)
Opisthokonta	70	20.92	Ryan et al. (2013)
Yeast	23	9.45	Salichos and Rokas (2013)

*: Computed as the sum of branch lengths, measured in expected substitutions per site, from the phylogeny.

For each alignment, we employed `ModelFinder` (Kalyaanamoorthy et al. 2017) to determine relative fit of standard protein exchangeability models using Bayesian Information Criteria (BIC), specifying arguments to mimic behavior of the commonly-used `ProtTest` software (Darriba et al. 2011). Previous studies have shown that either most standard measures of fit when in phylogenetics [BIC, Akaike Information Criteria (AIC), and small-sample Akaike Information Criteria (AIC_c), and Deviance Information Criteria (DIC)] perform comparably (Abadi et al. 2019), or that BIC may me somewhat more robust than other options (Luo et al. 2010). We therefore focus on BIC alone in this study as the criterion for model selection.

For each alignment, we ranked all tested models by BIC and identified five models ranging in goodness-of-fit for subsequent phylogenetic inference. Specifically, we employed the five models whose BIC values most closely matched the five-number summary (minimum, first quartile, median, third quartile, and maximum) of the distribution of BIC values for all models evaluated for a given alignment. This procedure allowed us to select, for each alignment, a set of models with a consistent range of fits to the given alignment. As depicted in Figure 1, we term the best-fitting model "M1" and so on for subsequent ranks along the five-number summary. We then used `IQ-TREE` to infer a phylogeny under each of these five models, along with two additional models: The equal-rates Jukes Cantor [JC, aka Poisson; (Jukes and Cantor 1969)] model, as well as a GTR model where exchangeability parameters are optimized to the data during inference. Notably, these two modeling frameworks are generally unused in protein phylogenetics. The JC model is assumed to be overly simplistic to reasonably describe protein evolution, and the GTR framework is presumed too parameter-rich and likely to overfit the data (Yang 2014).

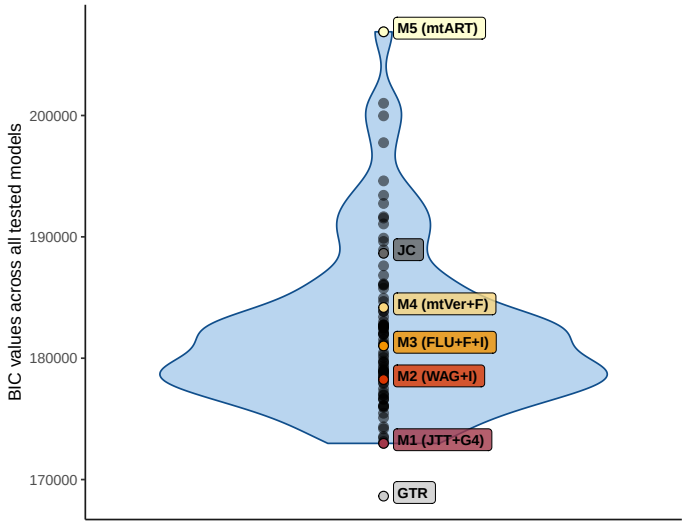

Figure 1: Approach to determining models for phylogenetic inference. This figure depicts the full distribution of BIC scores determined by `ModelFinder` for a single representative simulated alignment using HA-derived parameters along the Placental Mammals phylogeny. "M1" indicates the best-fitting model for this alignment, and "M5" indicates the worst-fitting model for this alignment. Models M2, M3, M4 indicate, respectively, models whose BIC score fell at the first, second (i.e. median), and third quartile of the full distribution of BIC values. The respective models corresponding to M1–5 are given in parentheses in the figure. We also show where the JC and GTR models, also used for phylogenetic reconstruction, fall within this overall distribution. For the vast majority of datasets, the JC model ranked between models M4 and M5, as shown for this representative alignment, and the GTR model was usually the best-fitting model.

Across all simulated alignments, the JC showed, as expected, consistently poor fit and always ranked between models M4 and M5 (Figure 1). However, JC was never actually the worst-fitting model, suggesting that entirely uninformative models are potentially suitable for analysis compared to models which inaccurately describe the evolutionary process. By contrast, the GTR model nearly always fit datasets better than did the respective M1 model. The only exceptions to this trend were certain alignments simulated along the Yeast phylogeny, which contained the fewest taxa (23) of all simulation trees. For roughly 50% of all simulations along the Yeast phylogeny, GTR was the best-fitting model, and M1 was the best-fitting model overall for the remaining 50%. The surprisingly high fit of the GTR model to most simulated datasets suggests that these simulations are rich in phylogenetic information.

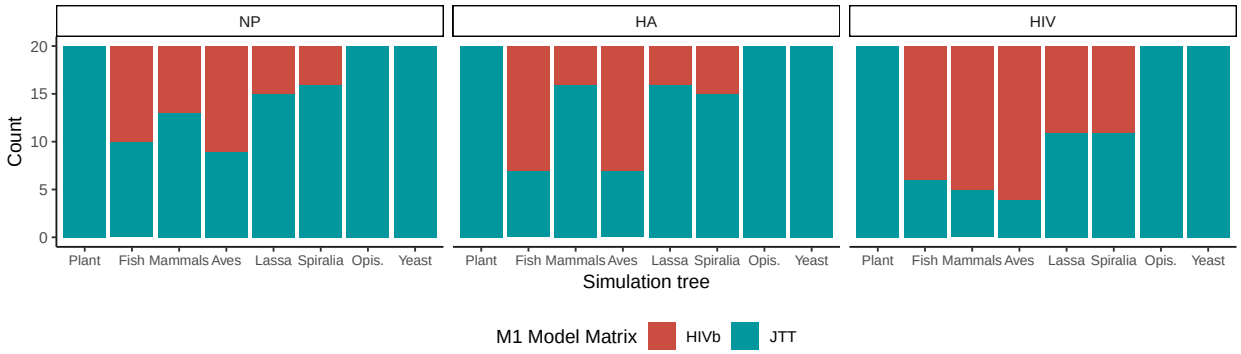

Figure 2: Best-fitting model (M1) matrix across simulations, where each column contains 20 simulation replicates. For visual clarity, the following abbreviations have been applied: Plant for Green plant, Fish for Ray-finned fish, Mammals for Placental mammals, Lassa for Lassa virus, and Opis. for Opisthokonta.

In general, we expect that the best-fitting model determined by model selection (M1) should be the model that best reflects evolutionary properties of data. If this is indeed the case, we further expect the M1 model to be broadly consistent among simulations with the same DMS-derived parameters. Considering only the selected model matrix (i.e. both WAG+I and WAG+F would be considered the same model matrix, WAG), we observed this expected trend Figure 2. Interestingly, the M1 model was consistent for all simulations, regardless of which DMS parameter set was used, with either JTT-based (Jones et al. 1992) or HIVb-based (Nickle et al. 2007) matrix emerging as the best-fitting model. As all employed DMS simulation parameters were derived from viral proteins, it is reasonable to observe that a virus-specialist model is often preferred. Even so, it is somewhat surprising that all simulations shared this preference for two specific model matrices, with simulations along three phylogenies (Green plant, Opsithokonta, and Yeast) consistently selecting a JTT model. Similarly, the model matrix mtArt, a model trained on arthropod-derived mitochondrial sequences (Abascal et al. 2006), was always the worst-fitting M5 model across all alignments by a substantial BIC margin. Contrasting with the M1 and M5 models, a wide range of model matrices corresponded to M2, M4 and M4 models, with between 3–13 model matrices observed at a given performance ranking (Figures S1–3).

There are several possible explanations for the overall similarity among M1 model matrices. First, while simulations across DMS parameterizations accounted for realistic differences in protein-level selection, other simulation parameters could have induced overly-similar properties across alignments. Specifically, all simulations assumed symmetric and equal mutation rates among nucleotides, the same fitness among synonymous codons (no codon usage bias), and no indels (insertions/deletions). Alternatively, the similarity among M1 model matrices may reflect inherent biases in experimentally-derived DMS data itself. While DMS can recover local evolutionary constraints acting on each position in a protein, pooling all sites together may obscure protein-specific evolutionary signal, giving the appearance that entirely distinct proteins have more comparable evolutionary patterns (Ramsey et al. 2011). Indeed, it has been suggested that DMS-derived fitnesses may not always reflect true evolutionary constraints observed in nature due to the controlled laboratory conditions in which they are obtained Haddox et al. (2016). Finally, it is possible that JTT and HIVb happen to possess unique evolutionary information that generally represent protein evolution, in spite

of the strong biological differences between the training datasets for each of these models.

Inferred tree topologies are consistent, although distant from the true tree, regardless of model fit

To assess the relationship between model fit and phylogenetic topological accuracy, we first calculated the Robinson-Foulds (RF) distance between each inferred tree and the true simulation tree. To ensure consistent comparisons across simulation conditions, all RF distances were normalized by the maximum possible RF for the given phylogeny. Throughout, we use the acronym "nRF" to refer to normalized Robinson-Foulds distance. Figure 3 displays the nRF distance between each set of inferred phylogenies compared to the true simulation phylogeny across all simulations.

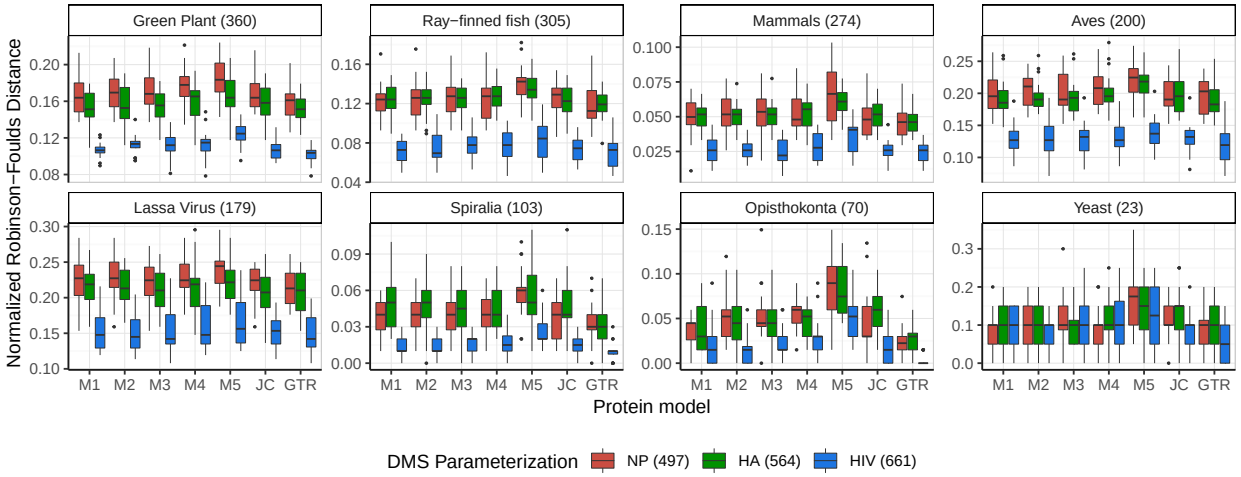

Figure 3: Normalized Robinson-Foulds distance between each inferred tree, across protein models, and the respective true tree used in simulation. Each panel represents simulations along a given tree, with the number of tips in the tree indicated in parentheses. Each boxplot represents simulations derived from a given set of DMS parameters, where the number of sites is given in parentheses in the legend.

If accuracy tracks model fit in protein phylogenetics, we should observe that nRF increases as model goodness-of-fit decreases, with the best-fitting model (here, M1 and/or GTR) inferring the most accurate tree. Our results did not convey this trend: nRF was remarkably consistent across inference models, with slight elevations apparent for M5 models under certain simulation trees, most notably Opisthokonta. Instead, the dominant trend in Figure 3 is that nRF decreased as the number of sites in the dataset increased, with HIV simulations showing much smaller nRF values compared to NP and HA simulations.

We fit a mixed effects linear model to determine the specific influence of protein model fit on nRF in these simulations, specifying nRF as the response, the protein model (M1–M5, JC, and GTR) as a fixed effect, and the simulation tree and DMS parameterization each as random effects. We performed a Tukey test to evaluate pairwise differences in mean nRF among protein models. There was no resulting significant difference in any nRF comparison among M1, M2, M3, M4, and JC models (all $P > 0.3$, except M1–M4 comparison where $P = 0.03$). This broad consistency suggests that most protein models, so long as they are not

severely misspecified, will yield phylogenies of comparable distance to the true tree. We did, on the other hand, observe significant differences for GTR-inferred and M5-inferred trees: i) the GTR model had significantly smaller nRF compared to other models (all $P < 0.007$), and ii) the M5 model had a significantly higher nRF compared to all other models (all $P < 0.001$). That said, all effect sizes for significant comparisons were exceedingly small, ranging from 0.007–0.029, representing the GTR–M1 and GTR–M5 comparisons, respectively. The latter coefficient implies that nRF increases, on average, by a mere average $\sim 3\%$ from the best-fitting to worst-fitting model.

All models infer similar amounts of strongly-supported but incorrect splits

Although model fit did not substantially affect nRF in any simulations, very few inferences actually converged on the exact true tree (RF distance of 0). Out of the total 3360 tree inferences conducted (480 simulated alignments $\times$ seven protein inference models), only 130 inferences across 59 distinct simulations achieved RF distance of 0 (Table S1). All of these inferences corresponded either the Spiralia, Opthisthokonta, or Yeast phylogenies, which were the three phylogenies with the fewest number of taxa (Table 2). Most notably, all models, including M5, were able to reach the true tree for at least one replicate, but the GTR model most frequently yielded the true true (44/130 times).

However, RF distance is a notoriously conservative metric that considers only presence or absence of nodes without considering their uncertainty, i.e. the level of support for inferred nodes under a given inference model. If differing splits are poorly supported, RF will overstate the distance between trees being evaluated. By contrast, differing splits with strong support represent more problematic deviations from the true tree.

We therefore evaluated bootstrap support for each inferred phylogeny using the ultra-fast bootstrap approximation (`UFBoot2`) implemented in `IQ-TREE` (Minh et al. 2013; Hoang et al. 2017). Because `UFBoot2` is a less biased measure compared to the standard nonparametric bootstrap, it necessitates a somewhat different interpretation such that nodes with $\geq 95\%$ support are considered highly reliable (Minh et al. 2013; Hoang et al. 2017). In addition, this threshold of 95% for identifying trusted supported splits corresponds to a false positive rate of 5%.

For each inferred tree, across all models, we evaluated whether each inferred node was accurate (present in the true tree) and whether each inferred node was strongly supported (`UFBoot2` ≥ 0.95) under the given model. We evaluated the proportion of false positive nodes (supported nodes not present in the true tree) as well as accurately inferred nodes (supported nodes that are present in the true tree combined with unsupported nodes not present in the true tree). These proportions, for HA simulations, are shown in Figures 4, with corresponding results for NP and HIV simulations in Figures S4-5.

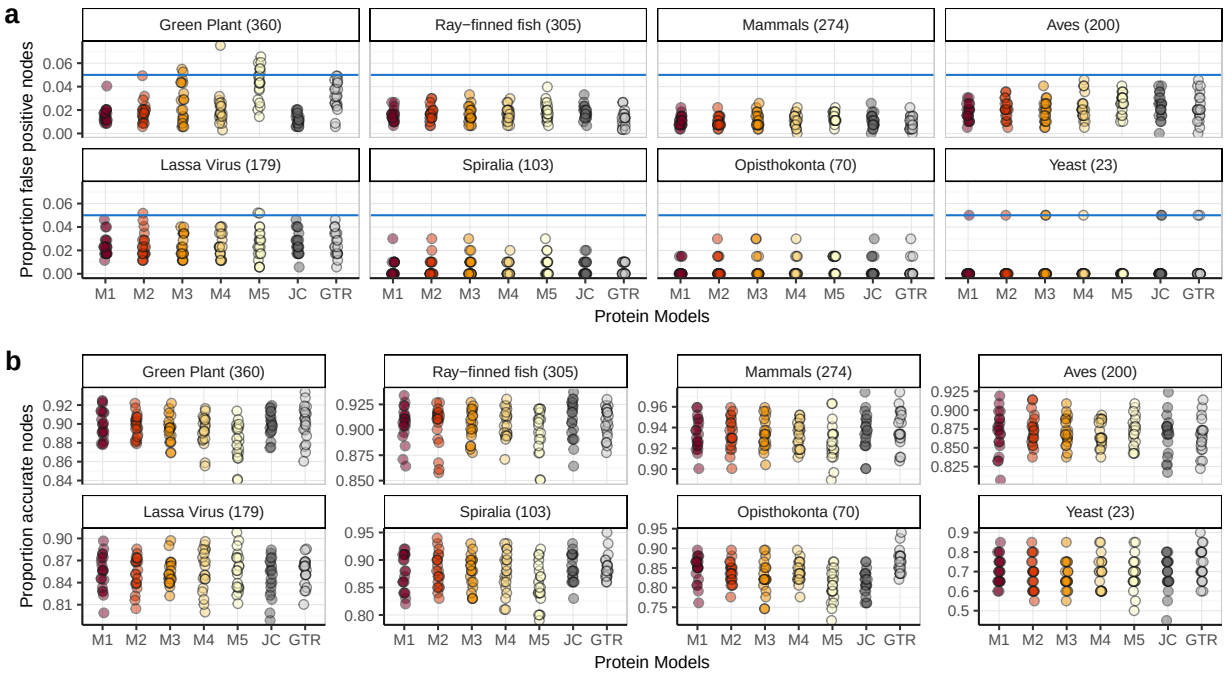

Figure 4: a) Proportion of false positive nodes in tree inferences, for HA simulations, using 95% `UFBoot2` as a threshold. The horizontal line in each panel is the $y = 0.95$ line, representing the expected false positive rate. b) Proportion of accurately-classified nodes in tree inferences, for HA simulations, using 95% `UFBoot2` as a threshold. Complementary results for NP and HIV simulations, shown in Figures S5-6, are broadly consistent to those shown here.

Results for this analysis agreed with those from nRF analysis: Protein models ranging in fit to the data yielded similar levels of support. The proportion of false positive nodes were remarkably similar across all conditions and were generally well-bounded below 5% with very few exceptions (Figures 4a and S4). Similarly, accuracy was fairly high and consistent across protein models and simulation conditions ((Figures 4b and S5).

We again analyzed this data with two linear models, considering either the proportion of false positive nodes or accuracy as the response, both with protein model as a fixed effect and DMS parameterization and tree as random effects, using a Tukey test to perform pairwise comparisons across protein models. The vast majority of comparisons were not significant (all $P > 0.06$), except for comparisons with M5 and GTR inferences. The M5 model yielded significantly more false positive nodes and was significantly less accuracy compared to all other models (both $P < 0.001$), but again effect sizes were extremely small. The largest effect size for false positive nodes came from the M5–M1 comparison, with M5 showing an average 0.6% increase in the number of false positives. Similarly, the largest effect size for accuracy came from the the M5–GTR comparison, with M5 showing an average 2.6% decrease in accuracy. In addition, GTR model produced significantly fewer false positive nodes and significantly higher accuracy compared to all other models (all $P < 0.004$), again with modest effect sizes of at most 1% fewer false positive nodes and at most 1.6% increase in accuracy (except in the M5–GTR comparison described above).

In total, consistent with nRF analysis, protein models ranging in fit to the data performed highly comparably, with the worst-fitting M5 model only producing marginally worse results

than other models. That no significant differences were observed among M1, M2, M3, M4, and JC models provides further evidence that model fit does not have substantial bearing on phylogenetic inference from protein data, so long as the model's fit is not severely poor. In addition, these results demonstrate that, any model regardless of fit has similar potential to detect a small proportion of highly-supported but incorrect nodes. This insight has important consequences for fundamental questions in phylogenetics. In particular, one reason it is desirable to avoid misspecified models is their presumed potential to yield supported but incorrect splits, or conversely correct splits which appear unsupported by the model Sullivan and Joyce (2005). For example, recent studies aiming to disentangle fundamental relationships among mammals (Philippe et al. 2011; Moran et al. 2015; Tarver et al. 2016) and metazoans (Pisani et al. 2015; Ryan et al. 2013) have suggested that one reason resolved phylogenies have been elusive is that many studies used employed inadequate models which tend to yield strong yet inconsistent support. Our results imply that *all* models are likely to support incorrect nodes, albeit in well-bounded manner, but no model - even the strongly misspecified M5 model - is substantially overrepresented for such nodes.

We note that these results also demonstrate the robustness of the `UFBoot2` bootstrap measurement to model misspecification, one of this measure's stated goals (Hoang et al. 2017). Indeed, in this circumstance, all models employed are misspecified to the simulated data, and support measures from `UFBoot2` suggest that all inferred trees show comparable levels of difference from the true tree.

Most inferred trees fall in the confidence set of trees under the M1 model

We next performed a series of AU (approximately unbiased) tests of tree topology to assess whether the observed topological differences represented significant deviations from the true phylogeny (Shimodaira 2002). For each alignment, we performed an AU test to compare the alignment's eight associated topologies: seven inferred trees and the true tree. As there exists no standard protein exchangeability model that will be correctly specified to data, we specified the given alignment's M1 model for each test, thereby asking whether any trees fell outside the confidence set of the M1 model. This approach mirrors a practical scenario where the best-fitting model determined by relative model selection would be assumed to be as accurately specified as possible.

Across all simulated alignments, we identified exceedingly few instances where any inferred phylogeny fell outside the M1 confidence set, at a threshold of $P < 0.01$ (Table S2). All trees inferred with models M2, M3, M4, and JC fell inside the respective M1 confidence set of trees (all $P \geq 0.044$). By contrast, for 19 simulations (4% of total), the M5 tree uniquely fell outside the M1 confidence set, and for a single simulation replicate the GTR tree uniquely fell outside the M1 confidence set. Most importantly, the true tree was in the M1 confidence set for all but 31 simulations (6.5% of total), most commonly HA simulations. None of these simulations were overlapping; for any given significant AU test, only one topology fell outside the M1 confidence set.

In sum, nRF comparisons, `UFBoot2` evaluation, and results from AU tests on simulated data provide consistent evidence that the exact choice of protein model does not have sub-

stantial bearing on the inferred phylogenetic topology, even if the model is a relatively poor fit to the data. We additionally emphasize the surprisingly good performance of the JC and GTR models in this simulation study. Neither of these models is commonly employed in single-gene protein phylogenetics, due to their established tendencdies to underfit and overfit data, respectively. While we indeed observed that JC underfit the data (Figure 1), GTR in fact emerged as the best-fitting model for the majority of simulations, except for roughly half of simulations along the smaller yeast phylogeny. The high informativeness of our simulations may therefore have overstated the power of the GTR model in protein data, relative to its performance on natural sequence data.

Analysis of natural sequence data reveals similar levels of consistency across models

We next analyzed how protein model fit affects phylogenetic inference for a set of natural sequence alignments. Because this analysis cannot truly assess inference accuracy (as the true phylogeny is unknown), we instead ask whether protein models ranging in goodness-of-fit to the data yield consistent or significantly different topologies. Furthermore, although the JC and GTR models performed well on simulated data, it is possible that these results were an artifact of the relative simplicity of simulated data compared to the complexity of natural sequence data. Examining how these protein models perform on real data is therefore crucial to properly contextualize their strong performance in simulations.

We randomly selected 200 protein alignments from the PANDIT database, considering only those with 20–500 (inclusive) sequences and 100–1000 sites (inclusive). As with simulated data, we determined the five protein models which most closely matched the BIC quartiles, and we used each model, as well as JC and GTR, to infer a phylogeny. The exact protein models identified at the BIC quartiles showed substantially more variety compared to selected models for simulated data (Figure S6), although over 75% of datasets selected an LG-based model Figure 5a. This strong bias towards LG likely reflects that the LG model itself was trained using PFAM alignments, the source for the PANDIT database (Whelan 2006; Le and Gascuel 2008). By contrast, similar to results from simulated data, the vast majority of M5 models were either mtArt or mtMam, a model trained on mammalian mitochondrial alignments (Yang et al. 1998). The general poor fit of certain mitochondrial models for both simulated and natural sequence data here likely reflects the highly unique nature of the data on which these models were originally trained.

Starkly contrasting with simulated alignments, the GTR model overfit virtually all natural alignments, emerging as the best-fitting model for only a single PANDIT alignment (Figure 5b). Even so, the GTR model generally was a better fit to each dataset than were JC and M5 models. As the PANDIT alignments analyzed here were substantially more sparse compared simulated alignments, with percent of gaps and ambiguous amino acids ranging from 19%–79% across alignments, the relatively poorer fit of the GTR model is not unreasonable. We tested whether certain features of the alignments, including percent of ambiguous characters, number of taxa, length of alignment, and/or treelength (sum of inferred branch lengths) could explain the relative rank of the GTR model among the seven models used for each alignment. Step-wise linear model selection using R^2 showed that the best

model to explain GTR rank was `GTR_rank` $\sim$ `number of taxa*treelength`, with $R^2 = 0.61$ (Figure 5c). As expected, then, the GTR model tends to be a better relative fit for more informative alignments and will be a poor fit to more uninformative alignments.

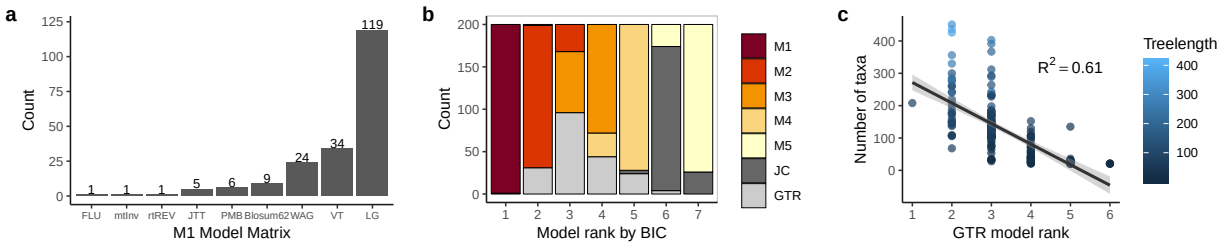

Figure 5: Model selection results on 200 PANDIT alignments. a) Best-fitting model (M1) matrix across PANDIT alignments. b) Relative model rank for models used for inference on PANDIT datasets, conveying the relative fits of JC and GTR models. c) Scatterplot showing key features in PANDIT datasets which predict the GTR model rank of the seven models examined for each alignment. Each point represents a PANDIT alignment, and treelength represents the sum of branch lengths each alignment's respective GTR-inferred phylogeny.

We examined to what extent trees inferred across protein models were consistent with one another using two separate analyses: i) an all-to-all comparison of Robinson-Foulds distances (Figure 6), and ii) AU tests for each inferred set of trees to assess whether they fell in the M1 confidence set (Figure 7). In Figure 6, we show the distributions of nRF distances across each pair of models, with the median value shown for each distribution. Overall, the mean nRF between M1 and M2 trees was significantly lower than all other comparison distributions ($P < 0.01$). There was virtually other no difference in average nRF for most comparisons among M1, M2, M3, M4, and GTR models. As such, while there are clear topological differences among these models, no single protein model of these five stood out has yielding substantially different topologies. These results are highly consistent with those from simulations and again suggest that relative model fit does not systematically affect inferred tree topologies.

By contrast, nRF comparisons with M5 and JC models were much higher, indicating that these two protein models tended to infer distinct topologies from M1–M4 and GTR models. We did not observe a significant difference in mean nRF between the M1–JC and M1–M5 comparisons, suggesting that M5 and JC yield trees with similar levels of deviation from M1. Interestingly, the mean nRF for the M5–JC comparison was significantly larger than were all other comparisons ($P < 0.001$). Therefore, while trees inferred with JC and M5 models were fairly distant from M1 trees, they were even farther from one another, indicating a qualitative difference between JC and M5 trees.

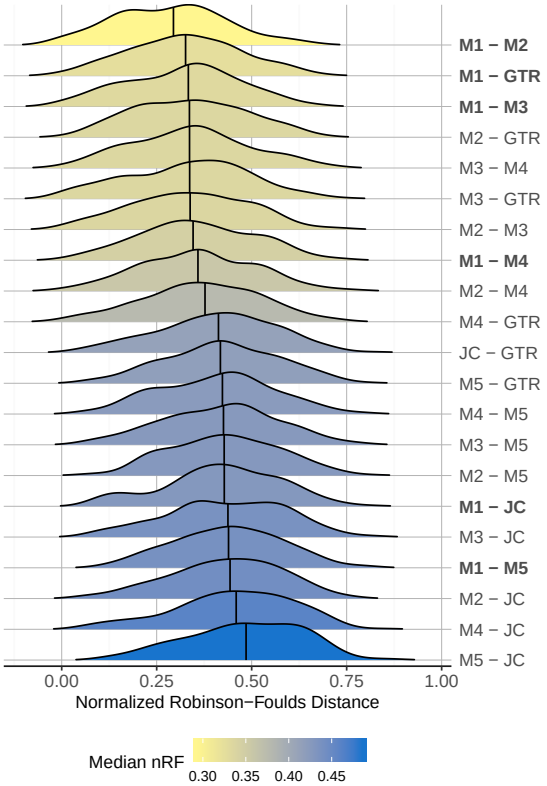

Figure 6: Distributions of nRF differences between trees inferred with each pair of models. Distributions are colored by the median nRF and arranged in increasing order of nRF. The vertical line through each distribution represents its median value, and rows containing M1 comparisons are bolded for clarity.

Further analysis with AU tests revealed that, for each alignment, most trees indeed fell in the M1 confidence set of trees (Figure 7a). For the 200 alignments examined, 199 (99.5%) M2 trees, 194 (97%) M3 trees, and 198 (99%) GTR trees fell in the M1 confidence set of trees ($P > 0.01$). While M2 and M3 models underfit the data compared to M1, and conversely GTR models nearly alwaysd overfit the data compared to M1, trees inferred with these three protein models were statistically consistent with those inferred with M1 models. By contrast, the M4 models deviated from the M1 confidence set somewhat more frequently, with only 183 (91.5%) of inferences consistent with the M1 model. Finally, at least 1/3 of inferred trees under M5 and JC each fell outside the M1 confidence set of trees for their respective alignments. We further asked whether the deviating trees inferred with M4, M5, and JC models represented the same or difference PANDIT alignments, as depicted with the Venn Diagram in Figure 7b. There was relatively little overlap between which M4 and JC/M5 trees fell outside the M1 confidence set, but there was much more overlap between M5 and JC such that. Even so, there were many instances where only the JC tree (from 19 alignments) or the M5 tree (from 37 alignments) uniquely differed from the M1 tree, again suggesting that JC and M5 inferred qualitatively distinct phylogenies.

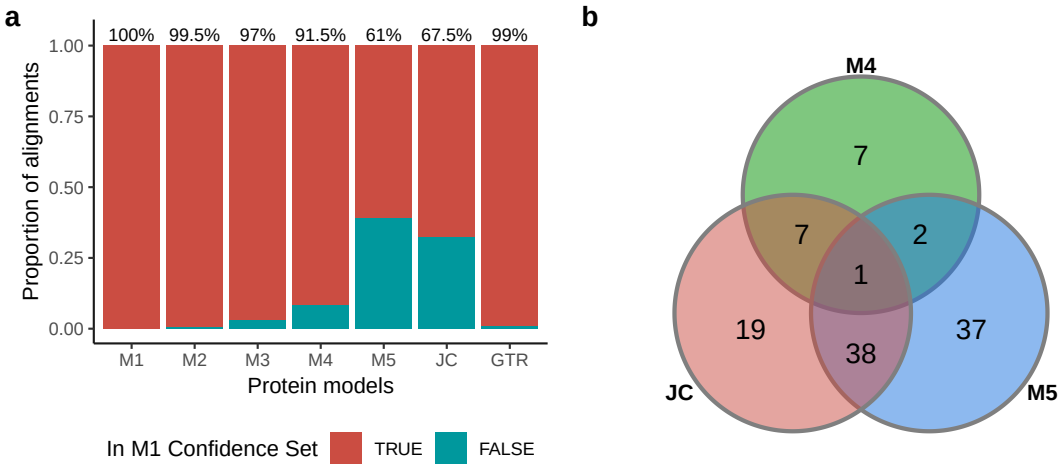

Figure 7: a) Proportion of PANDIT alignments whose the inferred tree, for each protein model, fell inside the M1 confidence set, as assessed with AU tests. The percentage shown at the top of each bar indicates the percent of trees in the M1 confidence set. b) Venn diagram depicting the number of PANDIT alignments whose trees inferred with M4, M5, and/or JC fell outside the respective M1 confidence set.

Unlike simulation results, where all JC trees fell inside the M1 confidence set, many JC trees built from natural sequence data had significantly different topologies. We therefore suggest that further work is necessary to truly understand the performance of this simplistic yet potentially effective model. Indeed, based on Figure 6, it appears that the equal-rates JC model infers unique topologies compared to any protein model with unequal exchangeabilities. While it is impossible to know which model(s), if any, converged upon the true phylogeny, the patterns observed from PANDIT data analysis imply that, so long as model fit is not exceptionally poor, the specific model is unlikely to strongly mislead or bias phylogenetic inference on protein data.

Conclusions

We have investigated whether relative model fit has a systematic affect on inference accuracy in single-gene protein phylogenetic inference. From both simulated and natural sequence data, we find that that inferred topologies are highly robust to the given protein model, so long as the model's fit to the data is not egregiously poor. We emphasize that the GTR model, which is often unused in protein phylogenetics due to its very large amount of free parameters to optimize, may be a very reasonable candidate model for phylogenetic inference even if/when it overfits the data.

This study focused exclusively on the practical ramifications of relative model selection when a single protein exchangeability models is applied to single-gene, non-partitioned data. Notably, we did not consider more complicated scenarios, such as the analysis of multiple concatenated genes in a partitioned analysis (Lanfear et al. 2017; Kainer and Lanfear 2015) and/or the use of more complex mixture models, such as the CAT model (Lartillot and Philippe 2004; Si Quang et al. 2008) or approaches that consider several exchangeability matrices proportioned across sites (Le et al. 2008; Huelsenbeck et al. 2008; Le et al. 2012).

As mixture models have been shown to fit many datasets better than single exchangeability models, in particular for saturated or highly heterogeneous data (Le et al. 2008; Si Quang et al. 2008; Arenas 2015), future work should investigate whether the improvement in fit these complex models confer corresponds to qualitatively different phylogenetic inferences.

In addition, we did not consider phylogenetic reconstruction from nucleotide data. However, a recent study by Abadi et al. (2019) used simulation to demonstrate that the GTR model and JC model as applied to nucleotide alignments do not produce systematically different phylogenetic topologies. Our results suggest that this phenomenon may also extend to protein-level data, ultimately providing increasing evidence that relative model selection is not strictly necessary in phylogenetic inference, in spite of decades of tradition. That said, while our results were very similar to those from Abadi et al. (2019), we did observe stronger differences between GTR and JC models as applied to natural sequence data, indicating that JC may not be as robust as simulations indicate. Either way, the finding that most reasonable models of sequence evolution will yield comparable phylogenetic topologies with similar levels of support stands.

The presumed relationship between model fit and inference accuracy likely emerged from the need for researchers to justify using a particular model; in the absence of a fully mechanistic model that accounts for the underlying evolutionary process, high goodness-of-fit to the data does provide a reasonable justification for using a particular model. Indeed, it is fundamentally impossible to employ a correctly specified model; no evolutionary model, regardless of its complexity or inclusion of parameters intended to capture fundamental biological processes (i.e. transition/transversion mutational biases), can possibly mimic the true generative process which gave rise to the data, that is actual biological evolution.

We additionally note that model selection was first applied in phylogenetics with an eye towards identifying the model with the most suitable level of complexity for the data, in the context of nucleotide-level models (Posada and Crandall 1998). While nucleotide models differ primarily in how many parameters are used to describe the substitution process, this circumstance does not cleanly apply to empirical protein exchangeability models. Unlike nucleotide models, all protein models contain the exact same number of fixed exchangeability parameters, and options for increasing model complexity most commonly entail adding very few additional parameters to account among site rate variation, proportion of invariant sites, and/or customized stationary frequencies. As such, the primary differences among protein models emerge from different relative substitution rates and *not* from the complexity of substitution process itself. It is therefore unclear what exact role relative model selection plays in protein phylogenetics, relative to nucleotide phylogenetics, especially given the broadly consistent performance of models with wide ranges in fit to the data. Future efforts may seek to investigate what aspects of protein models drive fit to the data, as model complexity is less likely to play as substantial a role as it plays in evaluating nucleotide models.

In sum, results presented here contribute to a growing body of evidence that the practical ramifications of model selection in phylogenetics may be vastly overstated. A key unanswered question in many of these findings is *why* there are such substantial differences in model fit even when these models perform extremely similarly. One possible explanation is that, while observed patterns in the data may more closely match some models than others, all available protein models may be similarly distant from describing the evolutionary process which in fact gave rise to the data. As such, while different models may better capture certain features

of the data, none of them may have sufficient ability to capture the generative process of biological evolution. This insight may pave the way for the development of categorically novel modeling frameworks; if new protein exchangeability models are developed with improved fit to data, but these new models do not yield meaningful consequences for inferences, the benefit to "building a better mouse trap" is modest at best.

References

Abadi, S., D. Azouri, T. Pupko, and I. Mayrose. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nature Communications 10:934.

Abascal, F., D. Posada, and R. Zardoya. 2006. MtArt: A new model of amino acid replacement for arthropoda. Molecular Biology and Evolution 24:1–5.

Adachi, J., P. J. Waddell, W. Martin, and H. M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. J. Mol. Evol. 50:348–358.

Andersen, K. G., B. J. Shapiro, C. B. Matranga, R. Sealfon, A. E. Lin, L. M. Moses, O. A. Folarin, A. Goba, I. Odia, P. E. Ehiane, M. Momoh, E. M. England, S. Winnicki, L. M. Branco, S. K. Gire, E. Phelan, R. Tariyal, R. Tewhey, O. Omoniwa, M. Fullah, R. Fonnie, M. Fonnie, L. Kanneh, S. Jalloh, M. Gbakie, S. Saffa, K. Karbo, A. D. Gladden, J. Qu, M. Stremlau, M. Nekoui, H. K. Finucane, S. Tabrizi, J. J. Vitti, B. Birren, M. Fitzgerald, C. McCowan, A. Ireland, A. M. Berlin, J. Bochicchio, B. Tazon-Vega, N. J. Lennon, E. M. Ryan, Z. Bjornson, D. A. Milner, A. K. Lukens, N. Broodie, M. Rowland, M. Heinrich, M. Akdag, J. S. Schieffelin, D. Levy, H. Akpan, D. G. Bausch, K. Rubins, J. B. McCormick, E. S. Lander, S. Gnther, L. Hensley, S. Okogbenin, S. F. Schaffner, P. O. Okokhere, S. H. Khan, D. S. Grant, G. O. Akpede, D. A. Asogun, A. Gnirke, J. Z. Levin, C. T. Happi, R. F. Garry, and P. C. Sabeti. 2015. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell 162:738–750.

Arenas, M. 2015. Trends in substitution models of molecular evolution. Frontiers in Genetics 6:319.

Bloom, J. D. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol. Biol. Evol. 31:1956–1978.

Bollback, J. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. Biology 19:1171–1180.

Brown, J. M. 2014. Predictive approaches to assessing the fit of evolutionary models. Syst Biol 63:289–92.

Brown, J. M. and R. C. Thomson. 2018. Evaluating Model Performance in Evolutionary Biology. Annual Review of Ecology, Evolution, and Systematics 49:95–114.

Dang, C. C., Q. S. Le, O. Gascuel, and V. S. Le. 2010. FLU, an amino acid substitution model for influenza proteins. BMC Evolutionary Biology 10:99.

Darriba, D., D. Posada, A. M. Kozlov, A. Stamatakis, B. Morel, and T. Flouri. 2019. ModelTest-NG: a new and scalable tool for the selection of dna and protein evolutionary models. bioRxiv .

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2011. Prottest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9:772.

dos Reis, M., Inoue Jun, Hasegawa Masami, Asher, Robert J., Donoghue, Philip C. J., and Yang, Ziheng. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proceedings of the Royal Society B: Biological Sciences 279:3491–3500.

Doud, M. B., O. Ashenberg, and J. D. Bloom. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. Mol. Biol. Evol. 32:2944–2960.

Duchne, D. A., S. Duchne, E. C. Holmes, and S. Y. W. Ho. 2015. Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations. Molecular Biology and Evolution 32:2986–2995.

Duchne, S., F. Di Giallonardo, and E. C. Holmes. 2016. Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales. Molecular Biology and Evolution 33:255–267.

Echave, J., S. J. Spielman, and C. O. Wilke. 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17:109–121.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. 2013. Bayesian Data Analysis. Third ed. Chapman and Hall/CRC.

Goldman, N. 1993a. Simple diagnostic statistical tests of models for DNA substitution. Journal of Molecular Evolution 37:650–661.

Goldman, N. 1993b. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36:182–198.

Haddox, H. K., A. S. Dingens, and J. D. Bloom. 2016. Experimental estimation of the effects of all amino-acid mutations to hivs envelope protein on viral replication in cell culture. PLOS Pathogens 12.

Haddox, H. K., A. S. Dingens, S. K. Hilton, J. Overbaugh, and J. D. Bloom. 2018. Mapping mutational effects along the evolutionary landscape of HIV envelope. eLife 7.

Halpern, A. L. and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2017. UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution 35:518–522.

Holder, M. T., D. J. Zwickl, and C. Dessimoz. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Phil. Trans. R. Soc. B 363:4013–4021.

Huelsenbeck, J. P., P. Joyce, C. Lakner, and F. Ronquist. 2008. Bayesian Analysis of Amino Acid Substitution Models. Philosophical Transactions: Biological Sciences 363:3941–3953.

Hughes, L. C., G. Ort, Y. Huang, Y. Sun, C. C. Baldwin, A. W. Thompson, D. Arcila, R. Betancur-R., C. Li, L. Becker, N. Bellora, X. Zhao, X. Li, M. Wang, C. Fang, B. Xie, Z. Zhou, H. Huang, S. Chen, B. Venkatesh, and Q. Shi. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences 115:6249–6254.

Hhna S., M. G. T. R. B. J., Coghill L.M. 2017. p^3: Phylogenetic posterior prediction in revbayes. Molecular biology and evolution 35:1028–1034.

Jones, C. T., N. Youssef, E. Susko, and J. P. Bielawski. 2016. Shifting balance on a static mutation–selection landscape: A novel scenario of positive selection. Molecular Biology and Evolution 34:391–407.

Jones, C. T., N. Youssef, E. Susko, and J. P. Bielawski. 2018. Phenomenological load on model parameters can lead to false biological conclusions. Molecular Biology and Evolution 35:1473–1488.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.) iii ed. Academic Press, New York.

Kainer, D. and R. Lanfear. 2015. The Effects of Partitioning on Phylogenetic Inference. Molecular Biology and Evolution 32:1611–1627.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14:587–589.

Keane, T. M., C. J. Creevey, M. M. Pentony, T. J. Naughton, and J. O. McInerney. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology 6:29.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution 31:151–160.

Lanfear, R., P. B. Frandsen, A. M. Wright, T. Senfeld, and B. Calcott. 2017. Partitionfinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molecular Biology and Evolution 34:772–773.

Lartillot, N. and H. Philippe. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution 21:1095–1109.

Le, S. Q., C. C. Dang, and O. Gascuel. 2012. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. Molecular Biology and Evolution Pages 2921–2936.

Le, S. Q. and O. Gascuel. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Le, S. Q., N. Lartillot, and O. Gascuel. 2008. Phylogenetic mixture models for proteins. Philosophical Transactions of the Royal Society B: Biological Sciences 363:3965–3976.

Luo, A., H. Qiao, Y. Zhang, W. Shi, S. Y. Ho, W. Xu, A. Zhang, and C. Zhu. 2010. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. BMC Evolutionary Biology 10:242.

Marltaz, F., K. T. C. A. Peijnenburg, T. Goto, N. Satoh, and D. S. Rokhsar. 2019. A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. Current Biology 29:312–318.e3.

Minh, B. Q., M. A. T. Nguyen, and A. von Haeseler. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution 30:1188–1195.

Moran, R. J., C. C. Morgan, and M. J. O'Connell. 2015. A Guide to Phylogenetic Reconstruction Using Heterogeneous ModelsA Case Study from the Root of the Placental Mammal Tree. Computation 3:177–196.

Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32:268–274.

Nickle, D. C., L. Heath, M. A. Jensen, P. B. Gilbert, J. I. Mullins, and S. L. Kosakovsky Pond. 2007. HIV-specific probabilistic models of protein evolution. PLOS ONE 2:e503.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wrheide, and D. Baurain. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9:e1000602.

Pisani, D., W. Pett, M. Dohrmann, R. Feuda, O. Rota-Stabelli, H. Philippe, N. Lartillot, and G. Wrheide. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proceedings of the National Academy of Sciences 112:15402–15407.

Posada, D. and T. R. Buckley. 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Systematic Biology 53.

Posada, D. and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.

Prum, R. O., J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. M. Lemmon, and A. R. Lemmon. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

Ramsey, D. C., M. P. Scherrer, T. Zhou, and C. O. Wilke. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188:479–488.

Ripplinger, J. and J. Sullivan. 2008. Does Choice in Model Selection Affect Maximum Likelihood Analysis? Systematic Biology 57:76–85.

Ripplinger, J. and J. Sullivan. 2010. Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. Molecular Biology and Evolution 27:2790–2803.

Ruhfel, B. R., M. A. Gitzendanner, P. S. Soltis, D. E. Soltis, and J. G. Burleigh. 2014. From algae to angiospermsinferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evolutionary Biology 14:23.

Ryan, J. F., K. Pang, C. E. Schnitzler, A.-D. Nguyen, R. T. Moreland, D. K. Simmons, B. J. Koch, W. R. Francis, P. Havlak, NISC Comparative Sequencing Program, S. A. Smith, N. H. Putnam, S. H. D. Haddock, C. W. Dunn, T. G. Wolfsberg, J. C. Mullikin, M. Q. Martindale, and A. D. Baxevanis. 2013. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science 342:1242592–1242592.

Salichos, L. and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Self, S. G. and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82:605–610.

Shimodaira, H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology 51:492–508.

Si Quang, L., O. Gascuel, and N. Lartillot. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317–2323.

Spielman, S. J. and S. L. Kosakovsky Pond. 2018. Relative Evolutionary Rates in Proteins Are Largely Insensitive to the Substitution Model. Molecular Biology and Evolution 35:2307–2317.

Spielman, S. J., S. Wan, and C. O. Wilke. 2016. A comparison of one-rate and two-rate inference frameworks for site-specific dn/ds estimation. Genetics 204:499–511.

Spielman, S. J. and C. O. Wilke. 2015a. Pyvolve: A flexible python module for simulating sequences along phylogenies. PLOS ONE 10:e0139047.

Spielman, S. J. and C. O. Wilke. 2015b. The relationship between dN/dS and scaled selection coefficients. Mol. Biol. Evol. 32:1097–1108.

Spielman, S. J. and C. O. Wilke. 2016. Extensively parameterized mutation–selection models reliably capture site-specific selective constraint. Molecular Biology and Evolution 33:2990–3002.

Sukumaran, J. and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Sullivan, J. and P. Joyce. 2005. Model Selection in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 36:445–466.

Tarver, J. E., M. dos Reis, S. Mirarab, R. J. Moran, S. Parker, J. E. OReilly, B. L. King, M. J. OConnell, R. J. Asher, T. Warnow, K. J. Peterson, P. C. J. Donoghue, and D. Pisani. 2016. The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. Genome Biology and Evolution 8:330–344.

Tavare, S. 1984. Lines of descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

Thyagarajan, B. and J. D. Bloom. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. eLife 3:e03300.

Whelan, S. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Research 34:D327–D331.

Whelan, S., J. E. Allen, B. P. Blackburne, and D. Talavera. 2015. ModelOMatic: Fast and Automated Model Selection between RY, Nucleotide, Amino Acid, and Codon Substitution Models. Systematic Biology 64:42–55.

Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18:691–699.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1.

Yang, N., R. Nielsen, and M. Hasegawa. 1998. Models of Amino Acid Substitution and Applications to Mitochondrial Protein Evolution. Mol. Biol. Evol. 15:1600–1611.

Yang, Z. 2014. Molecular Evolution: A Statistical Approach. Oxford University Press.