Relative 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 molecular phylogeny, it is regarded as best practice to perform relative model selection to determine an appropriate evolutionary model for the data. This procedure ranks available evolutionary models by a certain information theoretic criterion, and the best-fitting model is then specified for phylogenetic inference. While many assume that better-fitting models yield more accurate results, recent studies have observed that the specific model employed for phylogenetic construction may not substantially affect inferred tree topologies. Here, we examine whether there is a systematic relationship between model fit and topological inference accuracy, in the specific context of single-gene protein phylogenetic inference, using both simulations and analysis of natural sequence data. Notably, all simulations were performed using site-heterogeneous mechanistic codon models that are fully distinct from all protein-level inference models used for phylogenetic analysis. This strategy allows us to investigate how phylogenetic inference will perform when the model is misspecified to the underlying data, as will be the case for any analysis of real sequences. We broadly find that phylogenies inferred across models with vastly different relative fits to the data produce 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 even in circumstances when it is not the relatively best-fitting model to the data. This work builds on a growing body of literature finding that relative model selection does not necessarily promote increased accuracy in phylogenetic inference. We conclude that, while relative model selection may not hinder phylogenetic analysis, in most circumstances, it may not offer specific predictable improvements and is not a reliable proxy for accuracy.

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 follows a time-reversible 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), generally referred to as the GTR model. 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 information theoretic criteria (Posada and Buckley 2004). Such criteria include, for example, Akaike Information Criterion (AIC) and Bayesian Information Criterion (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 and continue to be developed (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. Relative 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. While it is of course necessary to select a model of evolution for any analysis, whether selecting a model based on relative fit to the data will directly translate to inference improvements has not been rigorously evaluated. Even so, casual and potentially misleading remarks about the role of relative model selection abound 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 include analysis of posterior predictive distributions or hypothesis

tests that ask whether the given inference model produces molecular properties (e.g. number of invariant sites, mean GC-content for nucleotide-level data, entropy, etc.) that mirror those seen in the data at hand (Goldman 1993a,b; Ripplinger and Sullivan 2010; Gelman et al. 2013; Brown 2014; Duchêne et al. 2015, 2016; Bollback 2002; Brown 2014; Duchêne et al. 2015; Hohna et al. 2017; Brown and Thomson 2018). 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 (Duchêne et al. 2015; Brown and Thomson 2018). As such, the majority of researchers performing phylogenetic reconstruction will most often rely on relative model selection to justify using a given 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 systematically biased models, and models that mitigate bias have relatively poorer fits 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 time-reversible nucleotide-level model (`GTR+I+G`) 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 relative 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. We ask whether phylogenetic reconstruction performed with better-fitting models consistently leads to increased accuracy compared to poorly-fitting models. We specifically study this question in the context of phylogenetic inference from protein data, using both simulations and natural sequences. Protein models are uniquely phenomenological compared to codon-level and nucleotide-level models because nucleotides, not amino acids, are the fundamental unit of evolution (Liberles et al. 2013; Jones et al. 2018). As such, precise evolutionary quantities such as mutation rate cannot be directly applied to protein data. From biological first principles, then, there is no mechanistic way to describe the evolutionary process 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 $-\sum \pi_i Q_{ii} = 1$. 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 are most commonly applied. Early efforts to generate these models produced seminal matrices such as the Dayhoff model (Dayhoff et al. 1978), and statistical advances over time led to more robust models derived from substantially larger training datasets that were specifically intended for phylogenetic reconstruction. These include the commonly-used models `JTT` (Jones et al. 1992), `WAG` (Whelan and Goldman 2001), and `LG` (Le and Gascuel 2008), 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), for example. 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). It is possible to incorporate among site rate variation (ASRV), 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 ~ 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 effects of applying more widely-used single matrix protein exchangeability models.

In addition, we may expect that relative model selection as applied to protein models behaves somewhat distinctly compared to applications on nucleotide models. Relative 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 often have different numbers of parameters, all empirical protein exchangeability models contain the exact same number of fixed exchangeability parameters, and options for increasing model complexity most entail adding very few additional parameters to account for phenomena such as ASRV or proportion of invariant sites. As such, the primary differences among empirical protein models

emerge from different exchangeabilities and *not* from the complexity of substitution process itself. This study therefore also seeks to clarify the specific role that relative model selection should play in the context of proteins, since that many of the complexity concerns that exist for nucleotide models are not applicable.

Overall, we do not observe a strong, systematic relationship between relative model fit to the data and inference accuracy, specifically considering phylogenetic topology. Except for the most relatively poorly-fitting models, protein models with drastically different fits to a given dataset produce surprisingly consistent phylogenies. Our analysis additionally reveals 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 to its excessive number of free parameters to optimize which tends 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 recover the true phylogeny. We therefore suggest that this often-ignored model may in fact be more viable than is generally assumed.

Methods and Materials

Simulation approach

We employ the site-wise 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

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

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, e.g. $S_{ij}^k = F_j - F_i$ where F_j is the fitness of codon j. Each site k in a given alignment is therefore specified by a unique 61-length vector of codon fitness values (excluding the three stop codons).

All simulations were conducted using the Python library `pyvolve` (Spielman and Wilke 2015a). We obtained site-specific codon fitness parameters used in simulations from four deep-mutational scanning (DMS) experiments (Table 1). DMS systematically determines the relative amino-acid preferences P for each site k in a real protein of length L (Firnberg et al. 2014). We converted these amino-acid preferences to to fitness values as $F_a = log(P_a)$, where F_a is the fitness of amino acid a and P_a is the experimentally-determined preference for amino acid a (Sella and Hirsh 2005). We subsequently converted all amino-acid fitnesses to codon fitness values, assuming equal fitness among synonymous codons at each site k. Each simulation therefore employed L experimentally-informed 61-length vectors describing codon fitnesses at each site k. Thus, the length of each simulation exactly matched the length of the originating protein itself; e.g. all simulations using LAC parameters were 262 codons in length and so on (Table 1). Simulations further assumed site-invariant equal mutation rates among all nucleotides.

We simulated 20 alignment replicates for set of DMS parameters along eight empirical phylogenies obtained from the literature (Table 2), resulting in a total of 640 simulated alignments. Because the MutSel model scales branch lengths to equal the expected number of neutral codon substitutions per unit time (Tamuri et al. 2012; Spielman and Wilke 2015a), all input phylogeny branch lengths were scaled up by a factor of three so that branch lengths could better approximate the number of amino-acid substitutions. Simulated alignments were then translated to amino-acids before subsequent analyses. In total, this simulation scheme ensured realistic levels of ASRV and heterotachy.

Model selection and phylogenetic inference

For each protein alignment, we employed `ModelFinder` in `IQ-TREE v1.6.8` (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017) to determine the relative fit of standard protein exchangeability models using Bayesian Information Criteria (BIC). Because 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 focus on BIC alone

in this study as the criterion for model selection.

We specified `ModelFinder` arguments that mimic behavior of the commonly-used `ProtTest` software (Darriba et al. 2011). `ModelFinder` examines a set of 21 commonly used protein models, as well as their respective parameterizations $+I$ (proportion of uninformative sites), $+F$ (using observed amino-acid frequencies), and $+G$ (four-category discrete gamma-distributed among-site rate variation, i.e. ASRV), totalling 168 examined parameterizations. We then ranked all evaluated models by their BIC score and identified five models ranging in goodness-of-fit for subsequent phylogenetic inference. Specifically, we identified the five models whose BIC scores most closely matched the five-number summary (minimum, first quartile, median, third quartile, and maximum) of the full distribution of BIC scores of all models evaluated for the given alignment. We refer to the best-fitting model `m1`, the second-best fitting model `m2`, and and so on for subsequent ranks along the five-number summary.

We then used `IQ-TREE v1.6.8` (Nguyen et al. 2015) to infer a phylogeny using each of these five models, along with two additional models which were not considered in `ModelFinder`: The equal-rates `JC` (Jukes and Cantor 1969) model, as well as the `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 and likely to underfit the data, and the `GTR` framework is presumed too parameter-rich and likely to overfit the data (Yang 2014).

Calculation of Robinson-Foulds distance and other topological comparisons were performed using the Python library `dendropy v4.4.0` (Sukumaran and Holder 2010). Approximately unbiased (AU) topology tests (Shimodaira 2002) were performed in `IQ-TREE` by specifying the argument `-zb 10000 -au` to perform 10,000 RELL replicates (Kishino et al. 1990).

Empirical data analysis

All empirical protein datasets were collected 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). Model selection and phylogenetic inference was performed as described above.

Statistical analysis and availability

All statistical analysis and visualization was performed in `R` (R Core Team 2017), making use of the `tidyverse` visualization and analysis framework (Wickham 2016; Wickham et al. 2019). Linear modeling was conducted using the `lme4` package (Bates et al. 2015), with corrections for multiple comparisons performed with `multcomp` (Hothorn et al. 2008). Significance throughout was assessed using a threshold of $\alpha = 0.01$. 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

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 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 not confounded by this logical gap, 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). To

this end, we simulated alignments according to the site-wise codon-level mutation–selection (MutSel) model (Halpern and Bruno 1998; Spielman and Wilke 2015a), rather than using an empirical protein model variant. While both empirical protein models and MutSel follow a time-reversible Markov framework that both satisty detailed balance, they employ entirely distinct focal parameters: protein models used for phylogenetic inference consider a site-invariant matrix of phenomenological exchangeabilities among amino-acids, while the MutSel models used for simulation consider site-wise codon fitness parameters coupled with site-invariant nucleotide-level mutation rates. All simulations were translated to amino-acids before subsequent analysis.

Briefly, we obtained four sets of experimentally-derived amino-acid preferences, determined using deep-mutational scanning (DMS; Table 1). Twenty replicate simulations for each DMS parameter set were conducted along eight empirical phylogenies (Table 2), resulting in a total of 640 simulations (see *Methods* for additional details). Each simulation was designed to mimic the evolutionary landscape of a real protein: Each alignment site was parameterized according to that respective site's DMS-derived fitness values, and the length of each simulation matched the length of its originating protein (either the bacterial LAC, H1N1 NP, H1N1 HA, or HIV Env; Table 1).

Table 1: Deep-mutational scanning experimental data used for simulation.

Name	Description	Number of Sites	Reference
LAC	TEM-1 β-Lactamase	262	Firnberg et al. (2014); Bloom (2014b)
NP	Influenza H1N1 nucleoprotein	497	Bloom (2014a); Doud et al. (2015)
HA	Influenza H1N1 hemagglutinin	564	Thyagarajan and Bloom (2014)
HIV	HIV-1 Env Protein	661	Haddox et al. (2018)

After translating simulated codon alignments to amino acids, `ModelFinder` was then used to identify a set of five models (out of a default set of 168) with range of fit to the data, as measured by BIC, where `m1` indicates the best-fitting model for this alignment, and `m5` indicates the worst-fitting model for this alignment. Models `m2`, `m3`, `m4` represent, respectively, models whose BIC score fell at the first, second (i.e. median), and third quartile

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	Marlétaz 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.

of the full distribution of BIC values. These five models, as well as the equal-rates `JC` (Jukes and Cantor 1969) and `GTR` model (where all amino-acid exchangeabilities were optimized from the data) were used to reconstruct a total of seven phylogenies for each simulated alignment. A full overview of this procedure is shown in Figure 1, using an HA simulation replicate along the Placental Mammals tree as an example.

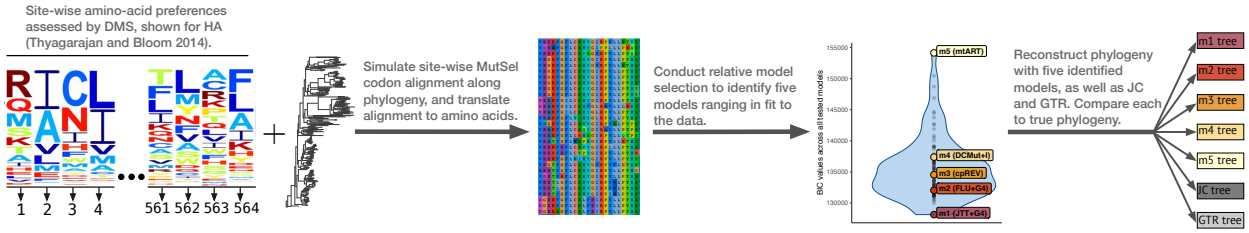

Figure 1: Flowchart describing simulation and phylogenetic reconstruction approach, using HA site-wise preferences (Thyagarajan and Bloom 2014) along the Placental Mammals phylogeny as an example (dos Reis et al. 2012). Sequence logos depicting HA preferences were adapted from Thyagarajan and Bloom (2014). Model names shown in parentheses in the BIC distribution next to `m1-m5` correspond to the actual models used for this example HA simulation replicate.

Model selection on simulated data

As ModelFinder does not evaluate the relative fit of the JC or the GTR models, we first examined their relative fit compared to m1–m5 for each simulation (Figure S1). Across all simulations, the JC showed consistently poor fit and always ranked between models m4 and m5. That JC was never actually the worst-fitting model suggests that entirely uninformative models may be more suitable for use compared to models with unequal exchangeabilities which inaccurately describe the evolutionary process. By contrast, the GTR model was consistently either the best-fitting model (for the larger NP, HA, and HIV simulations) or ranked in between the m1 and m2 models for most LAC simulations. Thus, GTR was a relatively high fit to the data for most simulations.

We generally expect that the best-fitting model determined by relative model selection (m1) should be the model that best reflects evolutionary properties of data. If this is indeed the case, we further expect that the m1 should be broadly consistent among simulations with the same set of 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 a JTT-based (Jones et al. 1992), HIVb-based (Nickle et al. 2007), or (for two LAC simulations) a WAG-based matrix emerging as the best-fitting model (Figure 2a). 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 [with one exception of a single LAC simulation whose m5 model was mtMAM, a model trained on mammalians mitochondrial sequences (Yang et al. 1998)] 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 (Table S1).

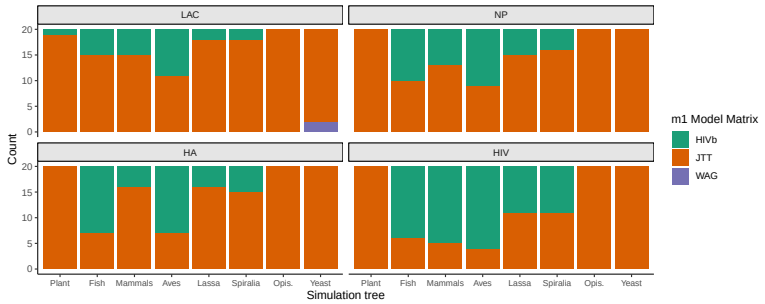

Figure 2: Best-fitting model (`m1`) matrix across simulations, where each column shows the selected model matrix for 20 simulation replicates. For visual clarity, the following abbreviations have been applied: Plant for Green plant, Fish for Ray-finned fish, Mammals for Placental mammals, Lassa for Lassa virus, and Opis. for Opisthokonta.

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. For example, 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 preferences themselves. 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 evolutionary information that generally represent protein evolution, in spite of the strong biological differences between the training datasets for each of these models. To probe relationship among these models further, we calculated the Pearson correlation among all model instantaneous rate matrices that `ModelFinder` considers (Figure S2). In fact, compared to all other models, `HIVb` model exchangeabilities are most strongly correlated with `JTT` exchangeabilities at $R = 0.906$. Furthermore, both `JTT` and `HIVb` models show the lowest correlations with `mtArt` ($R = 0.792$

and $R = 0.645$, respectively). The `HIVb` and `JTT` models therefore much more similar than their origins suggest. That said, `JTT` itself is most strongly correlated with its variant `JTTDCMUT` (Kosiol and Goldman 2005) ($R = 0.999$), followed by models `VT` (Müller and Vingron 2000) ($R = 0.954$) (Müller and Vingron 2000), `WAG` (Whelan and Goldman 2001) ($R = 0.935$), `LG` (Le and Gascuel 2008) and `mtInv` (Le et al. 2017) ($R = 0.916$ and $R = 0.914$, respectively), and finally `HIVb` at $R = 0.906$. Even so, the strong correlation between `HIVb` and `JTT` may explain why these two specific models predominated as `m1` models.

Inferred tree topologies show consistent distances from the true tree, regardless of relative 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.

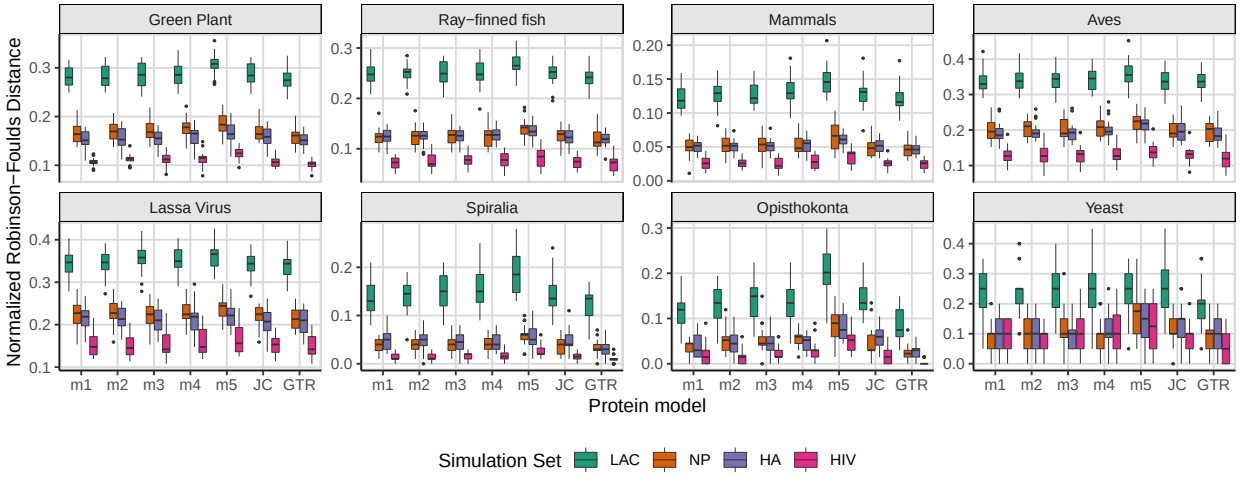

Figure 3: Normalized Robinson-Foulds distance (nRF) between tree inferences and the respective true tree used in simulation. Each boxplot represents the distribution of nRF values for 20 simulation replicates.

If accuracy tracks relative model fit in protein phylogenetic inference, 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, shown in Figure 3, 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 most apparent trend we observed was that nRF as the number of sites in the dataset increased, with HIV simulations showing much smaller nRF values compared to HA, NP, and LAC simulations. To further investigate this trend, we calculated the average entropy H of all simulations as $H = -\sum_i \Pi_i \ln \Pi_i$, where Π_i is the global frequency of amino-acid i, and the sum runs over all twenty amino-acids. In this case, entropy is bounded by $H \in [0, 2.996]$, where the maximum value indicates an equal frequency among all twenty amino acids. On average, HIV simulations had the highest average entropy ($H = 2.85$), followed by HA ($H = 2.83$), NP ($H = 2.81$), and LAC ($H = 2.78$) (Figure S5). These differences were highly significant (linear model, all $P < 4 \times 10^{-27}$), and the specific tree used for simulation had no significant effect on entropy (all $P > 0.03$). Therefore, the relative increase in amount of evolutionary signal in HIV simulations, combined with its larger number of sites, may have driven its overall lower nRF.

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. Trees inferred with models `m1`, `m2`, `m3`, `m4`, and `JC` (with the single exception of the `m1-m4` comparison) did not have significantly different nRF values, suggesting highly comparable performance among most models regardless of fit. Trees inferred with `m4` models did show a significantly larger nRF compared to `m1` trees, but with a vanishingly-small effect size of merely 0.7%. Trees inferred with `GTR` showed significantly smaller nRF compared to all other models, and trees inferred with `m5` models show significantly larger normalized RF compared to all other models. All significant differences detected had exceedingly small effect sizes, with the largest difference in nRF of 3.3% from the comparison between `GTR` and `m5`. As such, the average nRF improvement from applying the best-fitting model compared to the worst-fitting model is, at most, roughly 3%, ultimately demonstrating that relative model fit does not have a strong systematic influence on phylogenetic inference accuracy.

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 exactly matched the true tree (RF distance of 0). Out of the total 4480 tree inferences conducted (seven trees inferred for each of 640 simulated alignments), only 130 inferences across 42 distinct simulations achieved RF distance of 0 (Table S2). 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), with either HIV (87/130), NP (23/130), HA (20/130) simulation parameters. 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).

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 ultrafast bootstrap approximation (`UFBoot2`) implemented in `IQ-TREE` (Minh et al. 2013; Hoang et al. 2017). Because `UFBoot2` is presumed 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 supported splits is expected to correspond to a false positive rate (FPR) 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 FPR as well as the accuracy at this `UFBoot2` threshold. FPR is calculated as $\frac{FP}{FP+TN}$, where FP is the number of false positive nodes and TN is the number of true negative nodes. The accuracy is calculated as $\frac{TP+TN}{TP+FP+TN+FN}$, where TP is the number of true positive nodes, TN is the proportion of

true negative nodes, FP is the number of false positive nodes, and FN is the number of false negative nodes. The resulting classification metrics, specifically for HA simulations, are shown in Figures 4, with corresponding results for LAC, NP, and HIV simulations in Figures S3 and S4.

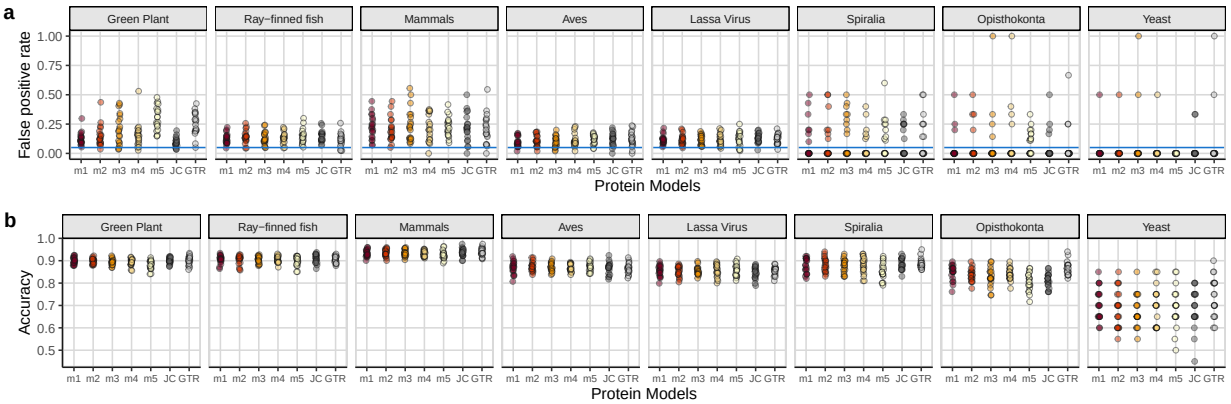

Figure 4: a) False positive rate (FPR) in inferred splits, for HA simulations, using 95% `UFBoot2` as a threshold. Each point represents the FPR of a single simulated alignment. The horizontal line in each panel is the $y = 0.05$ line, representing the expected FPR. b) Proportion of accurately-classified splits in tree inferences, for HA simulations, using 95% `UFBoot2` as a threshold. Each point represents the accuracy of a single simulated alignment. Figures S3 and S4 show complementary results for FPR and accuracy, respectively, for all other simulations and are broadly consistent to the HA results 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, with both FPR and accuracy being remarkably similar across all inference models. However, the FPR was not well-bounded at the expected value of 5%, suggesting that `UFBoot2` may not be as robust to model violations as has been previously presumed (Hoang et al. 2017). Further investigation of the sensitivity of this approximate bootstrap measure to model misspecification is therefore merited.

We again analyzed this data with two linear models, considering either FPR 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. Similar to results from nRF analyses, there was no significant difference in FPR among trees inferred with all models `m1`, `m2`, `m3`, `m4`, and `JC`. By contrast, the `GTR` model had significantly larger FPR compared to `m1`, `m2`, and `JC`, and the `m5` model had significantly larger FPR compared to `m1`, `m2`, `m3`, and `m4`. There was no significant difference in FPR

between m5 and GTR. Even so, the largest effect size for any comparison was still extremely small at a maximum of 3.7% (for the comparison between m1 and m5). Accuracy was not significantly different among models m1, m2, m3, m4, and JC, but GTR did show significantly higher accuracy compared to all other models, and m5 showed significantly lower accuracy compared to all other models. Again, however, effect sizes for all significant comparisons were very modest, with at most a 2.3% difference in accuracy for the comparison between GTR and m5 models.

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 neither FPR nor accuracy were significantly different among m1, m2, m3, m4, and JC models provides further evidence that relative model fit does not have substantial bearing on phylogenetic inference from protein data.

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 m1 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.

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 S3). 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. This result was somewhat surprising, given the substantial differences observed between inferred and true topologies (Figure 3). It is therefore a distinct possibility that trees with substantial topological differences may have more similar than anticipated likelihoods, an issue which

merits future investigations.

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

We next examined how relative model fit affects phylogenetic inference for a set of natural protein sequence alignments. Because such analyses cannot truly assess inference accuracy (as the true phylogeny is unknown), we asked whether protein models ranging in goodness-of-fit to the data yielded 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 were substantially more varied compared to selected models for simulated data (Table S4), although over 75% of datasets selected an `LG`-based `m1` 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). Similar to results from simulated data, the vast majority of `m5` models were either mtArt (Abascal et al. 2006) or mtMam (Yang et al. 1998). The general poor fit of certain mitochondrial models for both simulated and natural sequence data here may reflect the highly unique nature of the data on which these models were originally trained.

Starkly contrasting with simulated alignments, the `GTR` model only emerged as the best-fitting model for a single PANDIT alignment (Figure 5b). Even so, `GTR` was generally a better fit to each dataset than were `JC` and `m5` models. As the PANDIT alignments analyzed here were substantially more sparse compared to 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 examined 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). Thus, `GTR` tended to be a better relative fit for more informative alignments.

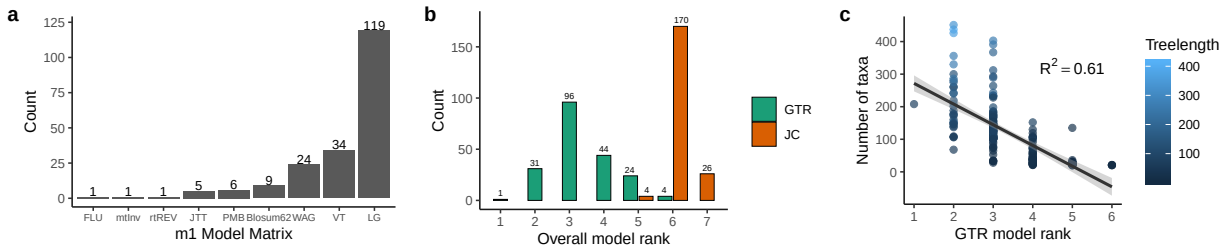

Figure 5: Model selection results on 200 PANDIT alignments. a) Best-fitting model (`m1`) matrix across PANDIT alignments. b) Relative rank, among all seven inference models, for the `JC` and `GTR` models, which were not considered by `ModelFinder`. c) Scatterplot showing key features in PANDIT datasets which predicted the relative `GTR` model rank within the seven inference models employed. 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, and ii) AU tests for each inferred set of trees to assess whether they fell in the respective `m1`-inferred tree's confidence set. In Figure 6a, 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 ($P < 0.01$), but we emphasize that the `m1-GTR` comparison showed the second highest median nRF difference. 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 trees inferred across these models, no single protein model of these five stood out as 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 qualitative difference between `JC` and `m5` trees. Indeed, while both of these models fit the natural sequence data very poorly, the poor fit of `JC` likely derived from their equal and therefore uninformative exchangeabilities, but the poor fit of `m5` models more likely derived from their misleading unequal exchangeabilities.

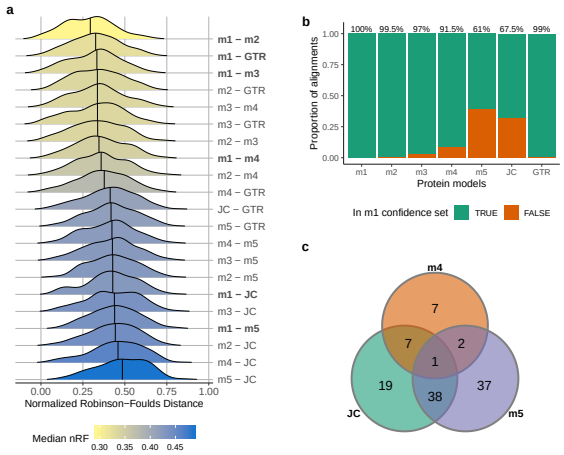

Figure 6: Results from PANDIT analysis. a) Distributions of nRF differences between trees inferred with each given pair of models. Distributions are colored by median nRF and arranged in increasing order of median nRF. The vertical line through each distribution represents its median nRF, and rows showing `m1` comparisons are bolded for clarity. b) 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. c) Venn diagram depicting the number of PANDIT alignments whose trees inferred with `m4`, `m5`, and/or `JC` fell outside the respective `m1` confidence set of trees.

Further analysis with AU tests revealed that, for each alignment, most trees indeed fell in

the m1 confidence set of trees (Figure 6b). For the 200 alignments examined, 199 (99.5%) of m2 trees, 194 (97%) of m3 trees, and 198 (99%) of GTR trees fell in the m1 confidence set of trees ($P > 0.01$). Therefore, although m2, m3, GTR models were poorer fits to 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 different PANDIT alignments, as depicted with the Venn Diagram in Figure 6c. 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 which m5 and JC trees fell outside the m1 confidence set. 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.

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 6a, 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 relative model fit is not exceptionally poor, the specific model is unlikely to strongly mislead or bias phylogenetic inference on protein data.

Discussion

We have investigated whether relative model fit has a systematic effect on inference accuracy in single-gene protein phylogenetic inference. From both simulated and natural sequence data, we find that inferred topologies are highly robust to the fit of the employed protein

model. We emphasize that the GTR model, which is often unused in protein phylogenetics due to the large amount of free parameters to optimize, may be a very reasonable candidate model for phylogenetic inference even if/when it is not the relatively best fit to the data.

An unexpected but key finding in the simulation study presented here is that most models, regardless of relative fit, will recover similar proportions of highly-supported but incorrect nodes (Figure 4, Figures S3-4). This insight has important consequences for fundamental questions in phylogenetics and systematics. 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 resolved phylogenies may have been elusive because 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 splits, but no model is substantially overrepresented for such splits.

Critically, the results presented here do not suggest that all models infer identical trees. The at-times wide distribution of nRF distances demonstrates that phylogenies inferred across varying models have, in many areas, distinct branching patterns (Figure 3 and Figure 6a). As such, our results do not imply that all protein models should infer the same tree for a given dataset. Instead, our results demonstrate that there is no clear, systematic shift towards more accurate inferences when relative model fit, as measured with information theoretic criteria, increases. While applying model selection procedures will not necessarily worsen a given analysis, there is similarly no robust evidence that applying relative model selection will improve analysis, as many users often presume. This is not to say relative model selection itself is either unnecessary or inaccurate, but rather that relative model selection is not a reliable "litmus test" for identifying which model will produce the most accurate and reliable inferences. In addition, because this study focused on relative model fit rather than absolute model fit, it remains a distinct possibility that all protein models had similar (potentially poor) absolute fits to the data. Future research endeavors should therefore assess the precise relationship between relative and absolute model fit measures to

achieving a unified understanding of the merits of phylogenetic model selection approaches.

This study focused exclusively on the practical ramifications of relative model selection when a single protein exchangeability models are 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+I+G` 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 predictive of accuracy in phylogenetic inference, in spite of decades of tradition. Similar to findings from Abadi et al. (2019), we also observed strong differences between the free-rate `GTR` and `JC` models as applied to natural sequence data, meaning that `JC` may not be as robust as simulations suggested.

Moreover, this study focused specifically on measures of topological accuracy when investigating how relative model fit might affect phylogenetic inference, and we did not investigate the effects of model fit on branch length inference. While such an analysis is clearly desirable, branch lengths under the simulation model (which operated at the codon level) and under inference models (which operated at the amino-acid level) are not directly comparable, and their relationship cannot be generalized. Under the MutSel simulation model, branch lengths are defined as the number of neutral codon changes per unit time, whereas under protein models used for inference, branch lengths are defined as the number of amino-acid changes per unit time, rendering these two quantities fundamentally distinct. Future efforts

therefore may seek a unifying framework to compare branch lengths and divergence levels across model formulations.

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 relative 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 relative 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.

Acknowledgements

We thank Joseph Bielawski, Christopher T. Jones, Noor Youssef, Edward Susko, and Andrew J. Roger for helpful discussions about phylogenetic comparisons and uncertainty. We additionally thank one anonymous reviewer and Shiran Abadi for valuable input to improve and clarify findings in this study.

References

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nature Communications 10(1):934.

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

Adachi J, Waddell P J, Martin W, M H. 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, Shapiro B J, Matranga C B, Sealfon R, Lin A E, Moses L M, Folarin O A, Goba A, Odia I, Ehiane P E, Momoh M, England E M, Winnicki S, Branco L M, Gire S K, Phelan E, Tariyal R, Tewhey R, Omoniwa O, Fullah M, Fonnie R, Fonnie M, Kanneh L, Jalloh S, Gbakie M, Saffa S, Karbo K, Gladden A D, Qu J, Stremlau M, Nekoui M, Finucane H K, Tabrizi S, Vitti J J, Birren B, Fitzgerald M, McCowan C, Ireland A, Berlin A M, Bochicchio J, Tazon-Vega B, Lennon N J, Ryan E M, Bjornson Z, Milner D A, Lukens A K, Broodie N, Rowland M, Heinrich M, Akdag M, Schieffelin J S, Levy D, Akpan H, Bausch D G, Rubins K, McCormick J B, Lander E S, Günther S, Hensley L, Okogbenin S, Schaffner S F, Okokhere P O, Khan S H, Grant D S, Akpede G O, Asogun D A, Gnirke A, Levin J Z, Happi C T, Garry R F, Sabeti P C. 2015. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell 162(4):738–750. ISSN 0092-8674, 1097-4172. doi:10.1016/j.cell.2015.07.020.

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

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48. doi:10.18637/jss.v067.i01.

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

Bloom J D. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. Mol. Biol. Evol. 31:2753–2769.

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(3):289–92. doi:10.1093/sysbio/syu009.

Brown J M, Thomson R C. 2018. Evaluating Model Performance in Evolutionary Biology. Annual Review of Ecology, Evolution, and Systematics 49(1):95–114.

Dang C C, Le Q S, Gascuel O, Le V S. 2010. FLU, an amino acid substitution model for influenza proteins. BMC Evolutionary Biology 10(1):99.

Darriba D, Posada D, Kozlov A M, Stamatakis A, Morel B, Flouri T. 2019. ModelTest-NG: a new and scalable tool for the selection of dna and protein evolutionary models. bioRxiv doi:10.1101/612903.

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

Darriba D, Taboada G L, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8):772. doi:10.1038/nmeth.2109.

Dayhoff M O, Schwartz R M, Orcutt B C. 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5:345–352.

dos Reis M, Inoue Jun, Hasegawa Masami, Asher, Robert J, Donoghue, Philip C J, 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(1742):3491–3500. doi:10.1098/rspb.2012.0683.

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

Duchêne D A, Duchêne S, Holmes E C, Ho S Y W. 2015. Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations. Molecular Biology and Evolution 32(11):2986–2995.

Duchêne S, Di Giallonardo F, Holmes E C. 2016. Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales. Molecular Biology and Evolution 33(1):255–267.

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

Firnberg E, Labonte J W, J G J, M O. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. Mol. Biol. Evol. 31:1581–1592.

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

Goldman N. 1993a. Simple diagnostic statistical tests of models for DNA substitution. Journal of Molecular Evolution 37(6):650–661.

Goldman N. 1993b. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36(2):182–198.

Haddox H K, Dingens A S, Bloom J D. 2016. 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). doi:10.1371/journal.ppat.1006114.

Haddox H K, Dingens A S, Hilton S K, Overbaugh J, Bloom J D. 2018. Mapping mutational effects along the evolutionary landscape of HIV envelope. eLife 7. ISSN 2050-084X. doi: 10.7554/eLife.34420.

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

Hoang D T, Chernomor O, von Haeseler A, Minh B Q, Vinh L S. 2017. UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution 35(2):518–522.

Hohna S, Coghill L, Mount G, Thomson R, Brown J. 2017. p^3: Phylogenetic posterior prediction in revbayes. Molecular biology and evolution 35:1028–1034.

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

Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. Biometrical Journal 50(3):346–363.

Huelsenbeck J P, Joyce P, Lakner C, Ronquist F. 2008. Bayesian Analysis of Amino Acid Substitution Models. Philosophical Transactions: Biological Sciences 363(1512,):3941–3953.

Hughes L C, Ortí G, Huang Y, Sun Y, Baldwin C C, Thompson A W, Arcila D, Betancur-R R, Li C, Becker L, Bellora N, Zhao X, Li X, Wang M, Fang C, Xie B, Zhou Z, Huang H, Chen S, Venkatesh B, Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences 115(24):6249–6254. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1719358115.

Jones C T, Youssef N, Susko E, Bielawski J P. 2016. Shifting balance on a static mutation–selection landscape: A novel scenario of positive selection. Molecular Biology and Evolution 34(2):391–407.

Jones C T, Youssef N, Susko E, Bielawski J P. 2018. Phenomenological load on model parameters can lead to false biological conclusions. Molecular Biology and Evolution 35(6):1473–1488.

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

Jukes T H, Cantor C R. 1969. Evolution of protein molecules. In H N Munro, editor, Mammalian protein metabolism, New York: Academic Press, pages 21–132. Iii edition.

Kainer D, Lanfear R. 2015. The Effects of Partitioning on Phylogenetic Inference. Molecular Biology and Evolution 32(6):1611–1627.

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

Keane T M, Creevey C J, Pentony M M, Naughton T J, McInerney J O. 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, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution 31(2):151–160.

Kosiol C, Goldman N. 2005. Different Versions of the Dayhoff Rate Matrix. Molecular Biology and Evolution 22(2):193–199.

Lanfear R, Frandsen P B, Wright A M, Senfeld T, Calcott B. 2017. Partitionfinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molecular Biology and Evolution 34(3):772–773.

Lartillot N, Philippe H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution 21(6):1095–1109.

Le S Q, Dang C C, Gascuel O. 2012. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. Molecular Biology and Evolution 29(10):2921–2936.

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

Le S Q, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philosophical Transactions of the Royal Society B: Biological Sciences 363(1512):3965–3976.

Le V S, Dang C C, Le Q S. 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evolutionary Biology 17(1):136.

Liberles D A, Teufel A I, Liu L, Stadler T. 2013. On the need for mechanistic models in computational genomics and metagenomics. Genome Biology and Evolution 5(10):2008–2018.

Luo A, Qiao H, Zhang Y, Shi W, Ho S Y, Xu W, Zhang A, Zhu C. 2010. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. BMC Evolutionary Biology 10(1):242.

Marlétaz F, Peijnenburg K T C A, Goto T, Satoh N, Rokhsar D S. 2019. A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. Current Biology 29(2):312–318.e3. ISSN 0960-9822. doi:10.1016/j.cub.2018.11.042.

Minh B Q, Nguyen M A T, von Haeseler A. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution 30(5):1188–1195.

Moran R J, Morgan C C, O'Connell M J. 2015. A Guide to Phylogenetic Reconstruction Using Heterogeneous Models—A Case Study from the Root of the Placental Mammal Tree. Computation 3(2):177–196.

Müller T, Vingron M. 2000. Modeling Amino Acid Replacement. Journal of Computational Biology 7(6):761–776.

Nguyen L T, Schmidt H A, von Haeseler A, Minh B Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32(1):268–274.

Nickle D C, Heath L, Jensen M A, Gilbert P B, Mullins J I, Kosakovsky Pond S L. 2007. HIV-specific probabilistic models of protein evolution. PLOS ONE 2(6):e503.

Philippe H, Brinkmann H, Lavrov D V, Littlewood D T J, Manuel M, Wörheide G, Baurain D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9(3):e1000602.

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proceedings of the National Academy of Sciences 112(50):15402–15407.

Posada D, Buckley T R. 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Systematic Biology 53(5).

Posada D, Crandall K A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9):817–818.

Prum R O, Berv J S, Dornburg A, Field D J, Townsend J P, Lemmon E M, Lemmon A R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526(7574):569–573. ISSN 1476-4687. doi:10.1038/nature15697.

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

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

Ripplinger J, Sullivan J. 2008. Does Choice in Model Selection Affect Maximum Likelihood Analysis? Systematic Biology 57(1):76–85.

Ripplinger J, Sullivan J. 2010. Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. Molecular Biology and Evolution 27(12):2790–2803.

Ruhfel B R, Gitzendanner M A, Soltis P S, Soltis D E, Burleigh J G. 2014. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evolutionary Biology 14(1):23. ISSN 1471-2148. doi:10.1186/1471-2148-14-23.

Ryan J F, Pang K, Schnitzler C E, Nguyen A D, Moreland R T, Simmons D K, Koch B J, Francis W R, Havlak P, NISC Comparative Sequencing Program, Smith S A, Putnam N H, Haddock S H D, Dunn C W, Wolfsberg T G, Mullikin J C, Martindale M Q, Baxevanis A D. 2013. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science 342(6164):1242592–1242592.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497(7449):327–331. ISSN 1476-4687. doi:10.1038/nature12130.

Self S G, Liang K Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82(398):605–610.

Sella G, Hirsh A E. 2005. The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. U.S.A. 102:9541–9546.

Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology 51(3):492–508.

Si Quang L, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24(20):2317–2323.

Spielman S J, Kosakovsky Pond S L. 2018. Relative Evolutionary Rates in Proteins Are Largely Insensitive to the Substitution Model. Molecular Biology and Evolution 35(9):2307–2317. ISSN 0737-4038, 1537-1719. doi:10.1093/molbev/msy127.

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

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

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

Spielman S J, Wilke C O. 2016. Extensively parameterized mutation–selection models reliably capture site-specific selective constraint. Molecular Biology and Evolution 33(11):2990–3002.

Sukumaran J, Holder M T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26(12):1569–1571. ISSN 1367-4803. doi:10.1093/bioinformatics/btq228.

Sullivan J, Joyce P. 2005. Model Selection in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 36(1):445–466.

Tamuri A U, Reis M d, Goldstein R A. 2012. Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. Genetics 190(3):1101–1115.

Tarver J E, dos Reis M, Mirarab S, Moran R J, Parker S, O'Reilly J E, King B L, O'Connell M J, Asher R J, Warnow T, Peterson K J, Donoghue P C J, Pisani D. 2016. The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. Genome Biology and Evolution 8(2):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, Bloom J D. 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(90001):D327–D331. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkj087.

Whelan S, Allen J E, Blackburne B P, Talavera D. 2015. ModelOMatic: Fast and Automated Model Selection between RY, Nucleotide, Amino Acid, and Codon Substitution Models. Systematic Biology 64(1):42–55.

Whelan S, Goldman N. 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. ISBN 978-3-319-24277-4.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. doi:10.21105/joss.01686.

Yang N, Nielsen R, Hasegawa M. 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.