Relative Evolutionary Rates in Proteins Are Largely Insensitive to the Substitution Model

Stephanie J. Spielman*[,1] and Sergei L. Kosakovsky Pond[1]

[1]Department of Biology, Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA

***Corresponding author:** E-mail: stephanie.spielman@temple.edu.

Associate editor: Meredith Yeager

Abstract

The relative evolutionary rates at individual sites in proteins are informative measures of conservation or adaptation. Often used as evolutionarily aware conservation scores, relative rates reveal key functional or strongly selected residues. Estimating rates in a phylogenetic context requires specifying a protein substitution model, which is typically a phenomenological model trained on a large empirical data set. A strong emphasis has traditionally been placed on selecting the "best-fit" model, with the implicit understanding that suboptimal or otherwise ill-fitting models might bias inferences. However, the pervasiveness and degree of such bias has not been systematically examined. We investigated how model choice impacts site-wise relative rates in a large set of empirical protein alignments. We compared models designed for use on any general protein, models designed for specific domains of life, and the simple equal-rates Jukes Cantor-style model (JC). As expected, information theoretic measures showed overwhelming evidence that some models fit the data decidedly better than others. By contrast, estimates of site-specific evolutionary rates were impressively insensitive to the substitution model used, revealing an unexpected degree of robustness to potential model misspecification. A deeper examination of the fewer than 5% of sites for which model inferences differed in a meaningful way showed that the JC model could uniquely identify rapidly evolving sites that models with empirically derived exchangeabilities failed to detect. We conclude that relative protein rates appear robust to the applied substitution model, and any sensible model of protein evolution, regardless of its fit to the data, should produce broadly consistent evolutionary rates.

Key words: **evolutionary rates, model selection, protein evolution, phylogenetics.**

Introduction

That the rates of substitution are not constant across a sequence, and are modulated by a multitude of processes and forces has been recognized since the dawn of modern comparative evolutionary analysis (Uzzell and Corbin 1971; Echave et al. 2016). Failing to account for site-to-site rate heterogeneity would be considered a neophyte error in contemporary applications, since it can lead to biased parameter estimation or incorrect phylogenies (Yang 1996). In addition to being an important confounder that needs to be corrected for, the distribution of rates across sites is of essential interest in its own right in many applications. As an example, the majority of analyses that seek imprints of natural selection on sequence data do so by inferring and interpreting the distributions of synonymous and nonsynonymous substitution rates and interpreting their properties (Delport et al. 2009).

In the context of protein sequence analysis, low site-specific substitution rates have served as a proxy for evolutionary conservation. Similarly, high rates have been regarded as a correlate of adaptation or positive selection (Sydykova and Wilke 2017). Positions which play key roles in protein functions, including those involved in protein–protein or protein–ligand interactions or those at or near active regions, tend to evolve very slowly and are highly conserved (Echave et al. 2016; Jack et al. 2016; Sydykova et al. 2018). By contrast, sites will tend to evolve rapidly if they interact with rapidly changing external stimuli, for example, if they are involved in chemosensory activity (Spielman and Wilke 2013; Almeida et al. 2015), or mediate immunity, that is, pathogen surface proteins or key regions of host immune genes (Tusche et al. 2012). Indeed, searching for conserved immune epitopes is a popular approach to finding vaccine or drug targets (Garcia-Boronat et al. 2008).

Evolutionary rates at individual sites in proteins are commonly measured as *relative* quantities, that indicate how quickly the site evolves relative to the "mean" protein rate. In such approaches, empirically derived models of protein evolution are fit to data in a phylogenetic framework, using either maximum-likelihood or Bayesian methods (Pupko et al. 2002; Mayrose et al. 2004; Fernandes and Atchley 2008; Nguyen et al. 2015; Ashkenazy et al. 2016; Spielman and Kosakovsky Pond 2018). Available substitution models can be loosely dichotomized into two classes. "Generalist" models are inferred from training data that comprise proteins from many domains of life, for example, JTT (Jones et al. 1992) or LG (Le and Gascuel 2008), and are meant to represent the shared evolutionary predilections of many different proteins.

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Temple University user on 07 September 2018

Article

Table 1. Data Sets Used for Rate Estimation.

Data Set	Class	N[a]	Median Sites (IQR[b])	Median Sequences (IQR)	Median Tree Length[c] (IQR)
Enzyme	Generalist	100	564 (371)	301 (104)	82.4 (34.0)
Mitochondria	Specialist	13	464 (357)	344 (2)	86.2 (49.4)
Chloroplast	Specialist	79	238 (328)	341 (15)	7.9 (7.52)
GPCR	Generalist	227	385 (156)	22 (4)	1.6 (1.75)

[a]Number of alignments in the data set.
[b]IQR stands for "interquartile range," defined as the difference between the 75th and the 25th percentiles of the distribution.
[c]Tree length is computed as the sum of branch lengths, measured in expected substitutions per site, from the phylogeny (all built under the LG+F model) used as LEISR input. See supplementary figures S1 and S2, Supplementary Material online, for details on how models affect tree lengths.

"Specialist" models, by contrast, are trained on data aimed to capture the properties of a particular taxonomic or biological group, for example, mtREV for mammalian mitochondrial sequences (Adachi and Hasegawa 1996) or AB for human antibody sequences (Mirsky et al. 2015).

Because site-specific rates are usually estimated by conditioning on a specific model of sequence evolution, one can reasonably assume that the estimated distribution of rates among sites will be influenced by the choice of the evolutionary model, which can be considered a nuisance parameter when rate estimates are of primary interest. It comes as no surprise that the question of model selection has assumed a prominent role in evolutionary rate inference, with popular implementations, such as *ProtTest* (Darriba et al. 2011) garnering thousands of citations.

Historically, protein substitution and similarity scoring models have primarily been developed and benchmarked in the specific context of homology identification (Henikoff and Henikoff 1992) and phylogenetic reconstruction (Le and Gascuel 2010). As a consequence, the relative performance of these models for inferring evolutionary rates from protein sequences has not been extensively studied. When this facet of model performance is mentioned, it is usually done in passing. For example, Landau et al. (2005) suggested that rates inferred with different models may be similar but feature "nonnegligible" differences.

We undertook a systematic comparison of site-specific rate estimates inferred using three generalist models, three specialist models, and the Jukes–Cantor equal-rates model (also known as JC, Jukes and Cantor 1969). As expected, standard likelihood-score based information criteria used in phylogenetic model selection revealed very strong model preferences for all alignments. Contrary to prevailing expectation, models yielded rates that were nearly perfectly correlated across alignments ranging in taxonomic scope and levels of sequence divergence. Even when we deliberately misapplied a specialist model, for example, by using a mitochondrial model on chloroplast data, we obtained rates that were almost perfectly correlated with the rates inferred under the cognizant specialist model. Only the extreme case of the equal-rates JC model yielded, albeit rarely, rates meaningfully different, and potentially indicative of positive selection, from models with unequal residue exchangeabilities.

Our results imply that some features of the evolutionary inference are quite robust to model misspecification and that any sensible model of protein evolution is likely to produce largely consistent evolutionary rate patterns in many settings. On the one hand, this finding is not as surprising as it may appear, because many evolutionary-rate analyses have been reported robust to various severe modeling violations at least based on simulated data and practical use cases (Anisimova et al. 2001). On the other hand, our results suggest that, depending on what estimates are of primary interest, standard model selection approaches may be suboptimal, since they do not identify the source of improvement in fit. In particular, our finding suggests that for some important applications, it is not necessary to waste CPU cycles on exhaustive model selection, and that alternative evaluative measures of goodness-of-fit, such as the use of posterior predictive approaches, could be more informative about the impact of evolutionary model choice on interpretable parameter inference (Brown 2014).

Results

We compiled 419 alignments (table 1) selected to represent both "general" and "specialist" proteins: a data set of enzyme alignments randomly selected from Jack et al. (2016), a data set of mammalian G protein-coupled receptor (GPCR) alignments from Spielman and Wilke (2013), a data set of green land plant chloroplast alignments, and a data set of Metazoan mitochondrial alignments (see Materials and Methods for details).

We inferred site-specific relative evolutionary rates from each alignment under three generalist models (LG, WAG, JTT), three specialist models (mtMet, gcpREV, HIVb), and the JC model with equal rates. Respectively, these three specialist models were originally trained on Metazoan mitochondrial sequences (Le et al. 2017), green plant chloroplast sequences (Cox and Foster 2013), between-host HIV-1 (subtype M) sequences (Nickle et al. 2007).

We inferred rates with each model twice, with and without including a gamma distributed component to model rate heterogeneity during the branch length optimization step, producing 14 sets of relative rate estimates per alignment. Throughout, we use +G to refer to a model which incorporated gamma-distributed rate heterogeneity. For a given alignment, rates inferred with any model employed the same equilibrium frequencies, empirically derived from the alignment, but a different exchangeability matrix. Any differences we identified between model inferences, therefore, were directly attributable to the differences in amino-acid

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

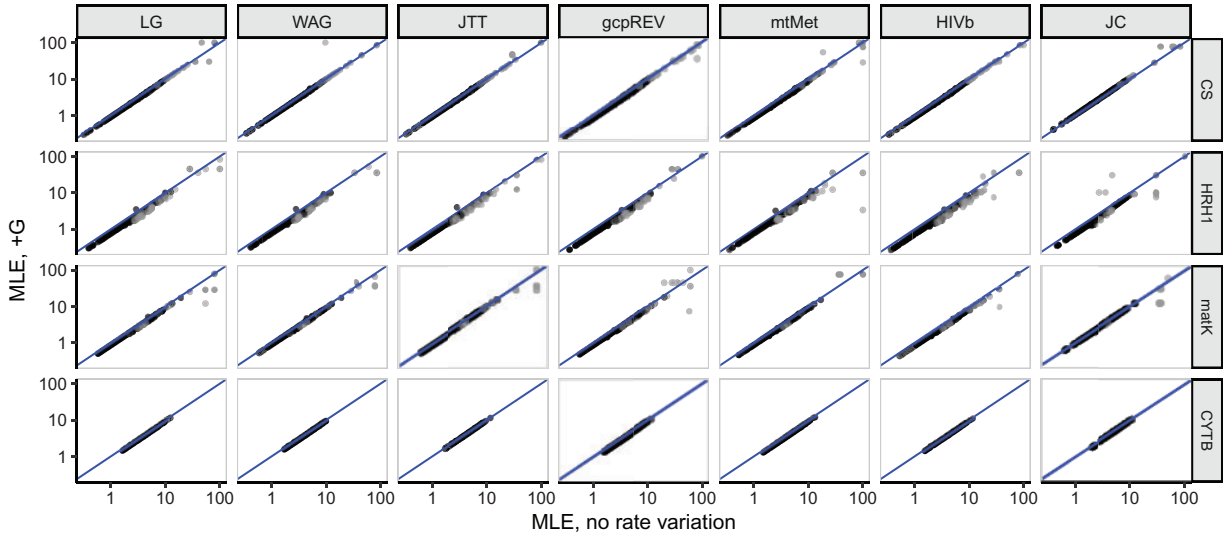

Fig. 1. Relationship of site-specific MLEs inferred with a given model, with and without specifying $+G$ in branch length optimization. Points in each log–log plot represent a single alignment site, and the line in each plot represents $x = y$. Representative alignments shown for enzyme, mitochondria, chloroplast, and GPCR data sets, respectively, are *CS* (*citrase synthase*), *HRH1* (*human histamine receptor 1*), *maturase K* (*matK*), and *cytochrome B* (*CYTB*). Black points represent MLEs with reliable estimates, and gray points represent those with unbounded (range $\geq$1,000) confidence intervals, for either axis. For visual clarity, all sites where MLE$<10^{-8}$, on either axis, have been removed from the figure.

exchangeabilities between model matrices. As such, we use the terms "model" and "matrix" interchangeably. We use the term "MLE" to refer to maximum-likelihood point estimates of relative rates.

A priori, we expected that gcpREV should best fit alignments from the chloroplast data set, and that mtMet should best fit alignments from the mitochondrial data set. Further, no model would be expected to fully recapitulate the evolutionary dynamics of the GPCR data set, as these transmembrane proteins are subject to unique evolutionary constraints imposed by the membrane environment (Stevens and Arkin 2001; Spielman and Wilke 2013). We included the HIVb model as an example of a specialist model expected to fit poorly to the data sets used here, all of which are dissimilar from the data on which HIVb had been trained.

Although other analyses of protein evolutionary rates have opted to normalize rates by the gene-wide mean or median (Jack et al. 2016; Spielman and Kosakovsky Pond 2018; Sydykova et al. 2018) or convert rates to standard *Z*-scores (Pupko et al. 2002), we directly analyzed rates yielded by LEISR without any normalization. Because of this choice, we consider Spearman (rank) correlations (ρ) when comparing MLEs. We adopt a rank-based test because MLEs are *relative* to the whole-protein rate, so the relative rank of MLEs is a more natural measure than the MLEs themselves. To complement this measure, we also consider Pearson correlation coefficients (r), on log-transformed data. This transformation is necessary to satisfy assumptions of the Pearson correlation, in part by mitigating the issue that the variance of rate estimates increases, possibly nonlinearly, with increasing rate magnitude (Scheffler et al. 2014). In addition, MLE distributions for a given gene are typically very skewed, with most rates falling below $\sim$10 with a few outlying MLEs several orders of magnitude larger.

Gamma Distributed Rate Variation Has Little Effect on Site-Specific Relative Rates

We first compared, for each alignment, MLEs between each model's inference, with or without $+G$. As expected, adding gamma variation resulted in increased tree lengths (supplementary figs. S1 and S2, Supplementary Material online) compared with constant-rate models. However, the effect of $+G$ on relative site rates was negligible. In figure 1, we show how these inferences relate for single representative alignment from each of our four data sets. Rates track the $x = y$ line of equality nearly perfectly across all data sets, with minor deviations generally appearing only at very high rates. Disagreements tend to emerge only for rapidly evolving sites whose rates are difficult to estimate and have numerically unbounded confidence intervals (CI range $\geq$ 1,000).

This near-perfect agreement was consistent across all models considered, and all alignments examined here (fig. 2*a*). The lowest measured correlation between model parameterizations was $\rho = 0.924$, and $\rho \geq 0.99$ for 98% of comparisons. Pearson correlations (fig. 2*b*) were even higher, with all $r > 0.98$. We therefore concluded that modeling rate variation during relative branch-length estimation has virtually no impact on the inferred site-specific rates MLEs. Consequently, we considered only rates inferred without $+G$ for the remainder of our analyses.

Models Yield Virtually Identical Relative Rate Inferences

We next assessed the extent to which the evolutionary model affected relative rate MLEs at individual sites. In figure 3, we show the relationship between LG MLEs and those inferred by all other models for four representative alignments across our data sets. A remarkably strong agreement was apparent for all model types (generalist, specialist, and equal-rates JC)

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

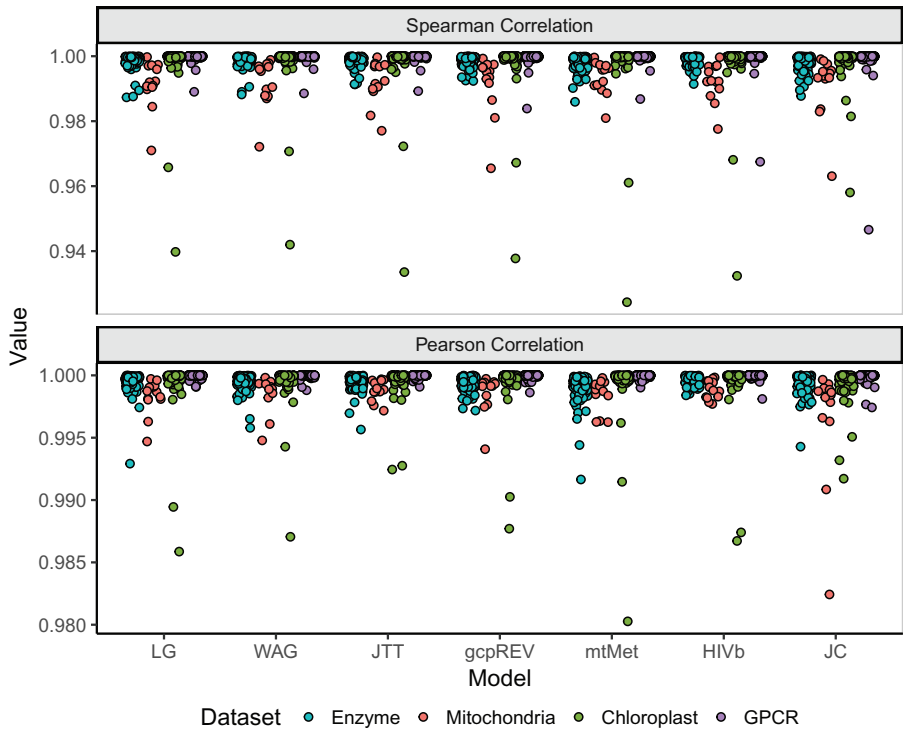

Fig. 2. Spearman and Pearson correlation coefficients of MLEs inferred with a given model, with and without specifying $+G$ in branch length optimization. Each point represents the respective correlation between MLEs for a single alignment. Note the limited range for the *y* axes, where panel (*a*) ranges from 0.92 to 1.0 and panel (*b*) ranges from 0.98 to 1.0.

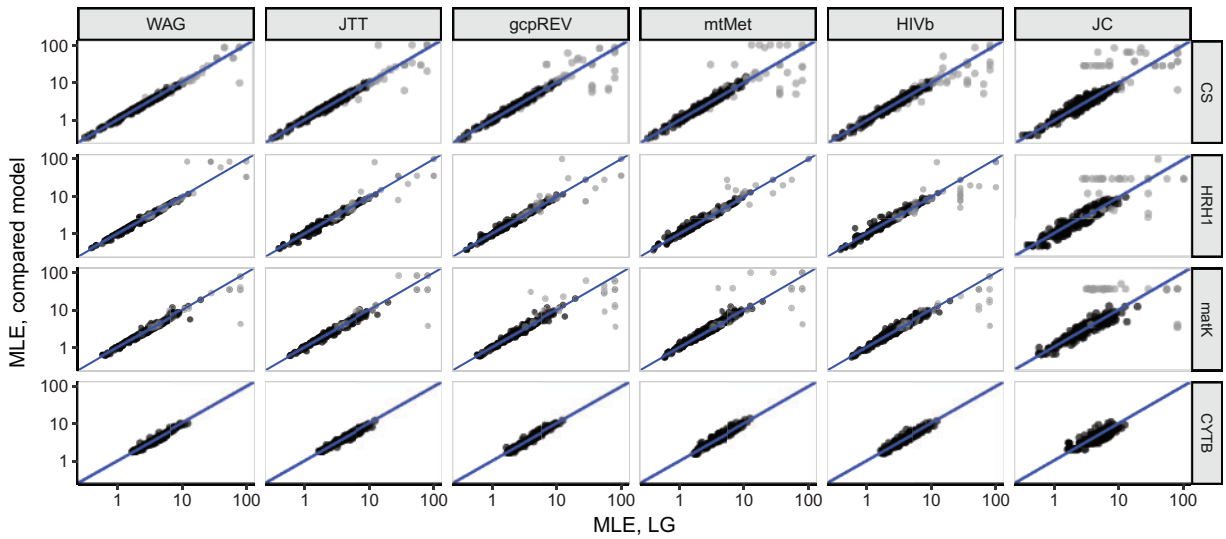

Fig. 3. Relationship of inferred rates ("MLE") between the LG and other models. Points in each log–log plot represent a single alignment site, and $x = y$ line is also shown. Representative alignments shown for enzyme, mitochondria, chloroplast, and GPCR data sets, respectively, are *CS* (*citrase synthase*), *HRH1* (*human histamine receptor 1*), *maturase K* (*matK*), and *cytochrome B* (*CYTB*). Black points represent MLEs with reliable estimates, and gray points represent those with unbounded (range$\geq$1,000) confidence intervals, for either axis. For visual clarity, all sites where MLE$<10^{-8}$, on either axis, have been removed from the figure.

and similarly for all data sets, regardless of taxonomy. Rate comparisons between LG and JC, however, did show somewhat more noise, although a clear rank correspondence for most sites was still present. Similar to the patterns observed in figure 1, points which do not fall close to the $x = y$ line in figure 3 nearly always corresponded to sites with imprecise MLEs, that is, sites with unbounded confidence intervals (CI range $\geq$1,000).

In figure 4, we show correlations between all pairs of models, averaged across alignments for each data set. As with correlations between rates with and without $+G$ (fig. 2), Pearson correlations were consistently larger than

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

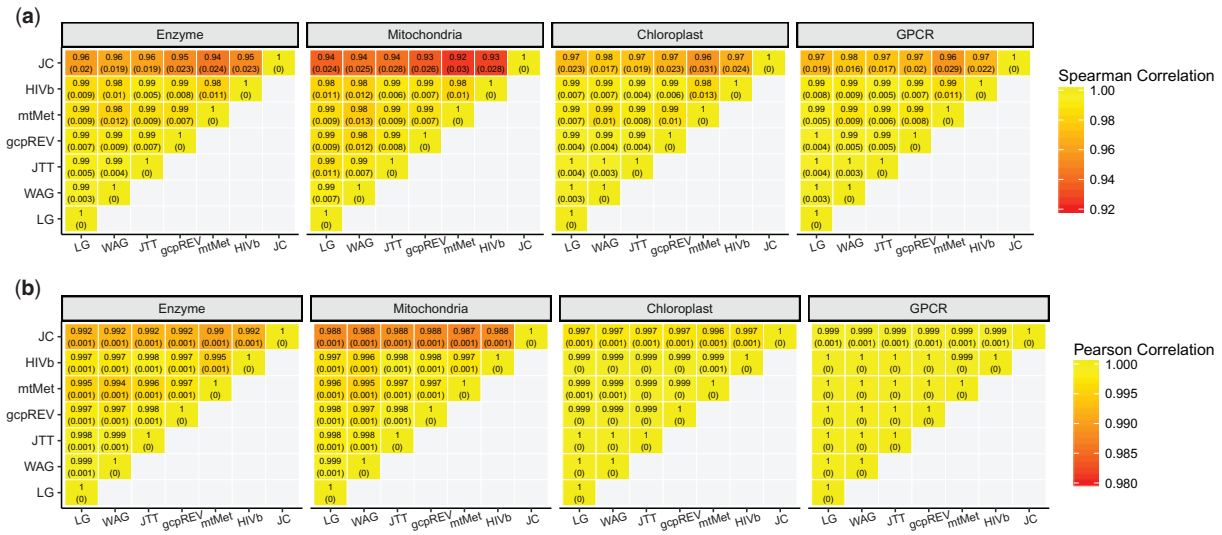

Fig. 4. Heatmap of correlations between model inferences, considering only rates inferred without branch length heterogeneity. Spearman correlation coefficients are shown in panel (*a*), and Pearson correlation coefficients are shown in panel (*b*). Values in each cell show the mean and SD of the respective correlation coefficient. Note that the heat scale ranges have limited ranges, from 0.92 to 1 in panel (*a*) and from 0.98 to 1 in panel (*b*).

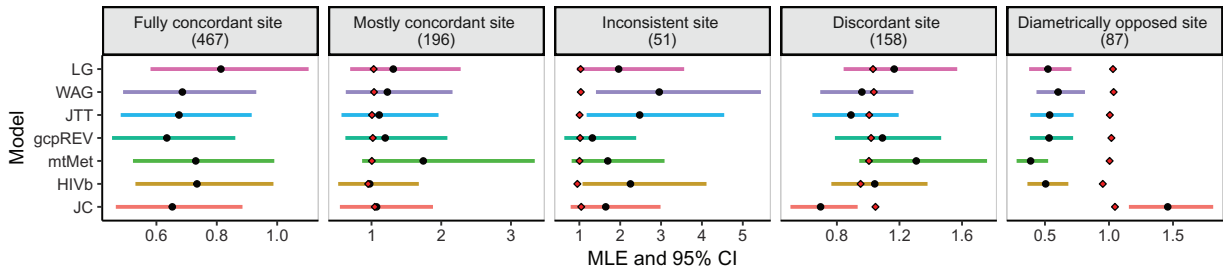

Fig. 5. Relative rate inferences, bounded by 95% confidence intervals, for five sites from an enzyme alignment of *citrase synthase* sequences. In each panel, black points represent site MLEs, red diamonds (when present) represent the median gene-wide rate under the given model, and horizontal bars represent 95% CIs. The title of each panel includes the site of interest's index from the *citrase synthase* alignment. See text for more details on site classifications.

Spearman correlations, but all values were still exceptionally high. The strength of these correlations was consistent across all models and data set types, generalist and specialist alike. For example, rates inferred on chloroplast with a chloroplast (gcpREV) model show nearly perfect correlations with rates inferred with a mitochondrial (mtMet) or HIV-specific (HIVb) model. JC was the only model with lower correlations with all other models, although the minimum averaged correlation was still extremely strong at $\rho = 0.92 \pm 0.03$ and $r = 0.987 \pm 0.001$.

Significant Rate Differences Are Infrequent but Generally Associated with JC

In spite of the near-perfect correlations among rates inferred with different models, relative rate estimates at individual sites are occasionally influenced by the choice of the model to a noticeable extent. For each individual site in all alignments (426, 678 sites), we assessed the extent to which the approximate 95% confidence intervals (CI) from different models overlapped one another. For example, in figure 5, we show five sites, each representing a different category of agreement or disagreement, from the enzyme *citrase synthase* alignment.

(1) A site was "fully concordant," when the MLE from any model fell within the 95% CI from every other model, indicating that the relative rate at this site was insensitive to model choice.

(2) A site was "mostly concordant," when the rate MLE from at least one model fell outside CIs from at least one other model (e.g., the MLE for mtMet fell outside the CI for HIVb), but all CIs included the median gene rate inferred from each model. In this case all MLEs did not significantly differ from the median rate, so a such a site would not be considered "interesting" for most downstream analyses. We considered the median gene-wide rate for this analysis, rather than the mean gene-wide rate, because difficult-to-estimate sites can yield inflated and highly outlying MLEs which overly bias the mean as a measure of location.

(3) An "inconsistent site" was a site where model inferences do not fully agree. At the example site in figure 5,

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

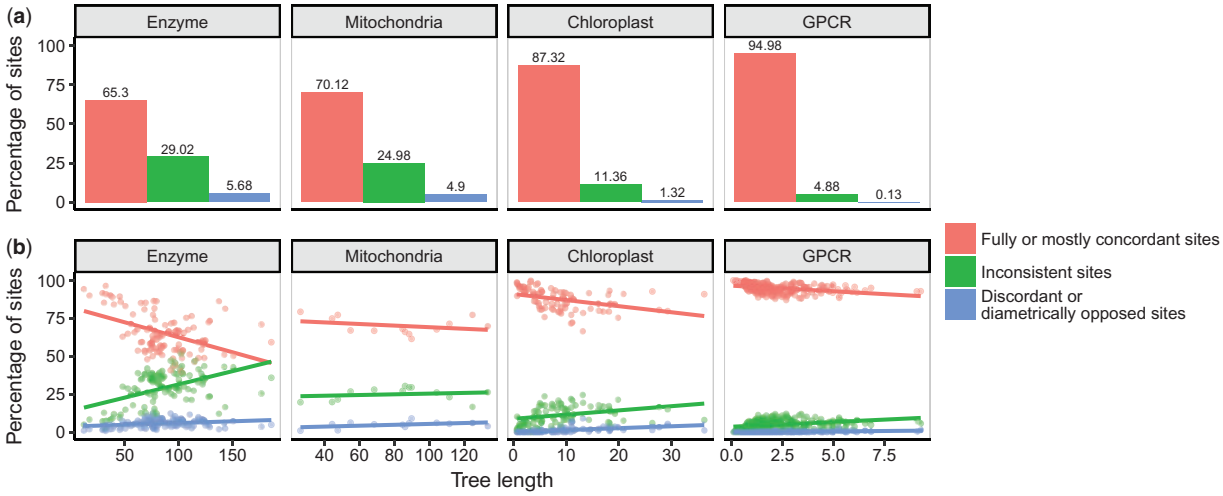

FIG. 6. Classification of model agreement on individual sites. (*a*) Average (over all alignments) percentage of sites, in each of our four data sets, which show different degrees of concordance among model inferences. (*b*) Alignment-level model concordance as a function of tree length (substitutions/site/unit time), where each point represents a single alignment, and linear regressions are drawn as solid lines.

the MLE from the WAG model fell outside the CI for the gcpREV model; JTT and WAG models indicate that the site was evolving faster than the median site, while the other five models suggest that the site's MLE did not significantly differ from the median. However, all MLEs here displayed the same general trend of being larger than the median rate, and therefore sites like this would usually also not be considered "interesting."

(4) A site was "discordant" if at least one model (in fig. 5, JC), yielded a fully inconsistent MLE compared with most models. Specifically, the JC MLE for the example site was reliably below the median rate, whereas all other models yielded MLEs that did not significantly differ from the median rate.

(5) A site was "diametrically opposed" when, depending on the model, its rate was either reliably below or above the median rate.

The last two site classifications (discordant and diametrically opposed) reveal the most interesting sites for our purposes, in that different models inferred different and at times opposite levels of evolutionary constraint.

We queried all 426, 678 sites across all alignments to determine how frequently the following scenarios occurred: 1) fully or mostly concordant sites, 2) inconsistent sites, and 3) discordant or diametrically opposed sites. We found that only relatively few sites were impacted by the choice of model (fig. 6*a*), and the extent of discordance was influenced by tree length (fig. 6*b*). As tree length increased, the proportion of sites where models agree tended to decrease, while the proportion of sites where models disagree tended to increase. This observation was highly significant, as assessed with a linear model with proportion of sites as a response, and the interaction of tree length and model agreement represented by the three scenarios shown in figure 6 as the predictor ($P < 2 \times 10^{-16}$). In other words, the specific model chosen should have a larger effect on rate estimates for a more diverged alignment. By contrast, the specific model chosen may have virtually no impact on an alignment with relatively low divergence. This was the case for the GPCR data set, which contained the fewest overall per-site substitutions and very rarely showed discordant model inferences.

Only 118 sites (0.03% of all alignment columns in this study) belonged to the diametrically opposed category (fig. 5*e*), that is, where at least one model reported a site's rate as significantly lower than the median rate, while at least one other model reported it as significantly higher than the median rate. Strikingly, these 118 sites, found among 53 enzyme and 3 mitochondrial alignments, all followed the same pattern: JC was the outlying model, all JC MLEs were above the median rate, and MLEs from all other models were below the median rate.

We hypothesized that these sites represent fast-evolving residues where the fixed amino acids have relatively high exchangeabilities in empirical matrices. In such matrices, this fast-evolving site would appear to have a relatively low rate simply because the exchangeabilities would contain information that "should" be incorporated into the rate. By contrast, in JC, the exchangeabilities make no a priori assumption of fast evolution, and thus the rate parameter would be able to capture the truly high rate.

To probe this question, we directly counted the number of substitutions among each pair of amino acids at all sites, across all alignments. We employed HyPhy (Kosakovsky Pond et al. 2005) to count substitutions using joint maximum-likelihood (Pupko et al. 2000) under a specified amino-acid model (here, LG) to reconstruct ancestral sequences. We performed this procedure a second time with the JC model, results of which were indistinguishable from counts under the LG model; supplementary fig. S3, Supplementary Material online. We tabulated substitutions under the principle of minimum evolution directly from the inferred ancestral sequences. Visual inspection of "diametrically opposed sites" revealed a high proportion of substitutions among the amino acids isoleucine (I), leucine (L), and valine (V), all

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

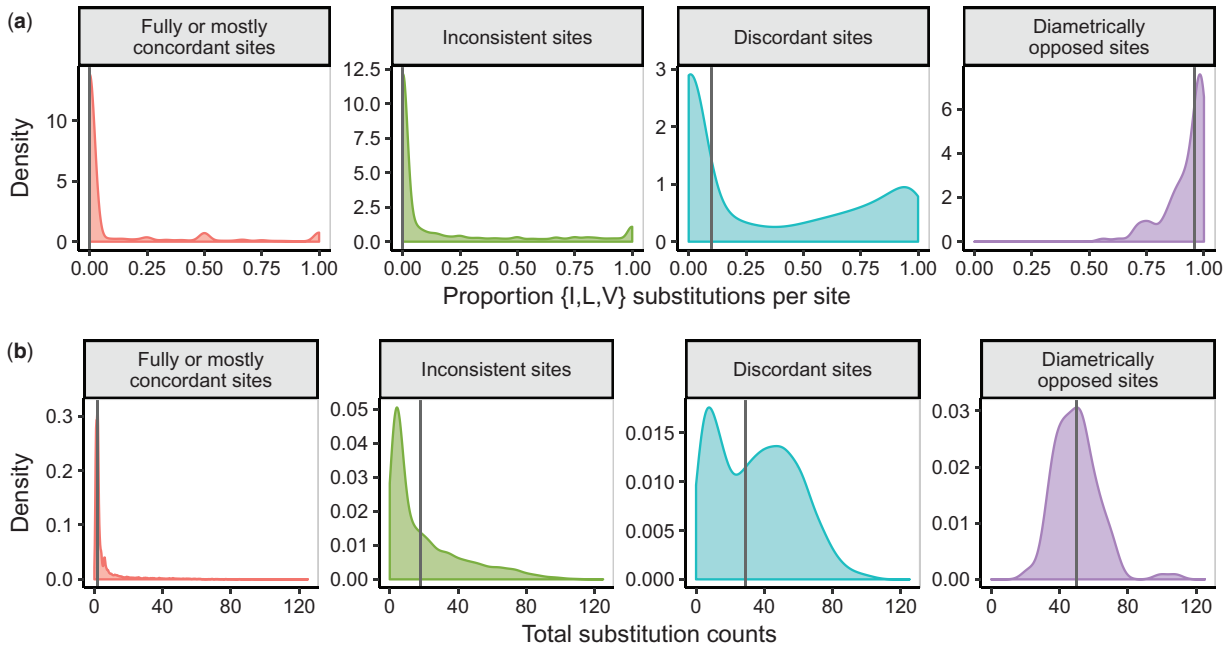

Fig. 7. Distribution of substitution counts across sites. (*a*) Proportion of {*I*, *L*, *V*} substitutions for all sites which have experienced at least one substitution, across site classifications as seen in figures 5 and 6. The vertical line in each panel represents the median proportion of {*I*, *L*, *V*} substitutions. (*b*) Total number of substitutions across site classifications, again only considering which had experienced at least one substitution. The vertical line in each panel represents the median total number of substitutions. Results in this figure were obtained using substitution counts under the LG model, but results are unaffected by the model chosen to count substitutions; see supplementary figure S3, Supplementary Material online, for analogous results obtained by counting substitutions under JC.

of which have extremely similar biochemical properties and consistently high exchangeabilities across empirical protein models. We therefore determined the proportion of substitutions at each site, considering only sites which experienced at least one substitution, which were between either IL, IV, or LV ({*I*, *L*, *V*} substitutions) (Similar results were obtained for a more generic group of "highly exchangeable" residues; not shown). We found that diametrically opposed sites, and to some extent discordant sites, were strongly enriched for {*I*, *L*, *V*} substitutions (ANOVA $P < 2 \times 10^{-16}$, fig. 7*a*), with at least 56% of substitutions at the diametrically opposed sites occurring among I, L, and V. We additionally found that diametrically opposed sites tended to experience more substitutions (ANOVA $P < 2 \times 10^{-16}$, fig. 7*b*) relative to other site classifications. One interpretation for these results is that despite its poor fit to the data (based on information criteria or log likelihood), JC may uniquely capture sites of potential interest, where evolution rapidly occurs among substitutions with high model exchangeabilities.

Strong Model Fit differences are Present In Spite of High Rate Agreement

We determined which model would be preferred for each alignment using standard procedures in phylogenetic model selection (Posada and Buckley 2004). Specifically, we calculated the Akaike Information Criterion $[AIC = 2(\log L - K)$, where log *L* represents the log-likelihood and *K* represents the number of estimated model parameters)] for each model fitted to each alignment, and we ranked all models accordingly. We performed this model ranking separately for models with and without +*G*. Because any model under a given rate variation setting has the same number of parameters, other commonly used information criterion measures such as small-sample *AIC* (AIC_c) or Bayesian Information Criterion would yield the same results as *AIC* does here.

Considering only models without +*G*, the best-fitting model generally matched expectations given the scope of an alignment. All enzyme alignments were best fit by a generalist model (LG, WAG, or JTT), the majority of chloroplast alignments were best fit by the gcpREV model, and the majority of mitochondrial alignments were best fit by the mtMet model (fig. 8*a*, upper panel). GPCR alignments, which have no corresponding specialist model, were best fit by either a generalist model, mtMet, or HIVb (fig. 8*a*, lower panel). As expected, JC never emerged as a best-fitting model. These trends were broadly consistent with model selection results for +*G* models, with a few minor differences. Curiously, relatively more GPCR models showed a preference for HIVb than for a generalist model when +*G* was applied.

We determined the relative level of support for these preferred models (combining both +*G* models and models without rate variation), by calculated each model's (relative) Akaike Weight. For each alignment, we calculated the weight *w* for each fitted model *i* as:

$$w_i = \frac{\exp[-0.5\Delta AIC]}{\sum_i \exp[-0.5\Delta AIC]}, \quad (1)$$

where $\Delta AIC = AIC_i - AIC_{\min}$, and $AIC_{\min}$ refers to the model with the smallest *AIC* value for the given alignment. For all but one chloroplast alignment, the best-fitting model ($AIC_{\min}$)

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

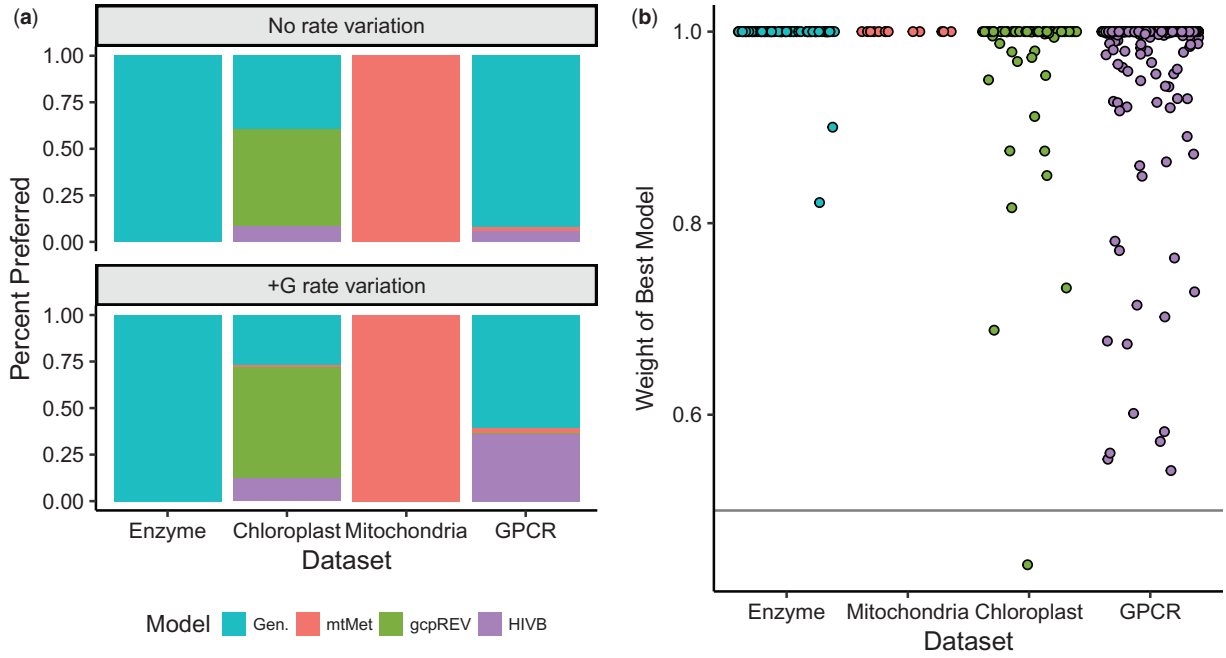

FIG. 8. Model selection results. (*a*) Distribution of preferred models, considering models without $+G$ (top panel) and with $+G$. The legend abbreviation "Gen." refers to one of the three generalist models (JTT, WAG, and LG) examined here. JC never emerged as a best-fitting model under either rate variation setting. (*b*) Relative Akaike weight of top model for all alignments, grouped by data set. The horizontal line is $y = 0.5$.

was robustly supported with $w \geq 0.5$ with the vast majority of weights at $w \approx 1$ (fig. 8*b*).

We conclude that although virtually all alignments show strong model preferences measured by goodness-of-fit, empirically the models revealed broadly consistent and effectively interchangeable estimates of site-specific relative rates, with the possible exception of JC in certain circumstances.

Discussion

We have investigated how the choice of empirical amino-acid evolutionary model affects inferred relative evolutionary rates in protein alignments. We conclude that for most sites in most alignments, the choice of substitution model has very little effect on the estimate. Even a priori poor choices, such as fitting a mitochondrial model to chloroplast data, failed to appreciably move the needle on site-specific relative rate estimates. More surprisingly, the model devoid of any biological realism (JC, Jukes and Cantor 1969), returned essentially the same estimates as the other models, with the exception of about one site in a thousand. Finally, while other applications of evolutionary rate have considered normalized and/or standardized rates (Pupko et al. 2002; Jack et al. 2016; Sydykova et al. 2018), which might emphasize agreement between models, our analysis found concordance between raw relative rates.

While rate inference appears to be quite robust to model choice, information theoretic criteria of model fit display very strong preferences toward a specific model of evolution (fig. 8), consistent with previous results and prior expectation. The goodness-of-fit improvements observed must therefore derive principally from features of the evolutionary process other than relative substitution rates at individual sites.

In general, the models examined here were developed for and have primarily been applied to questions of phylogenetic inference. Our results reveal both similar and distinct trends to those previously observed in the literature. For example, with respect to protein model performance, Keane et al. (2006) found that, in the context of phylogenetic tree inference, specialist models may not yield improved inferences relative to generalist models, even on specialist data. We have similarly shown that specialist and generalist models perform highly similarly when inferring relative evolutionary rates, for any given data set. By contrast, Huelsenbeck et al. (2008) suggested that, again in the context of phylogenetic inference, no single protein model may be suitable for a given alignment.

Our results instead suggest that, in the context of evolutionary rates, *any* protein model may be equally suitable for a given alignment, with the distinct possibility that they are all equally bad. However, we do emphasize that JC identified, albeit only very few, sites with salient signals of rapid evolution, where other more "realistic" models failed to identify these sites due to the high exchangabilities among substituting amino acids. As such, while JC consistently fits the data very poorly, this model that can identify rapid evolutionary toggling among highly similar amino acids. Such evolutionary patterns can be highly biologically meaningful; for example, previous work has demonstrated that certain sites in HIV-1 experience strong selection pressure to undergo such amino-acid toggling (Delport et al. 2008). Therefore, it is possible that the simplistic equal-rates JC could in fact be most useful for identifying certain types of selection pressures in proteins.

A careful analysis of properties of amino acid models including WAG, JTT, and LG by Goldstein and Pollock (2016)

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

suggested that, due to the fact that these models strive to capture the propensities of an average protein, they effectively describe evolution neutrally, that is, they do not fit any particular protein especially well. Our findings on the application of such models to evolutionary rate inference echo the conclusion from Goldstein and Pollock (2016) that these models may not contain substantially different information about site-specific rates of protein evolution. However, our observation that rate inferences with JC, on occasion, uniquely deviated from models with unequal exchangeabilities implies that there is some difference between "averaged" matrices and neutral evolution, as JC would represent the exact neutral scenario of protein evolution, that is, that any amino acid can be substituted for another with no selection preference or biochemical bias.

A wide variety of platforms have been popularized for selecting the best protein model in the context of both phylogenetic and evolutionary rate inference (Keane et al. 2006; Darriba et al. 2011; Ashkenazy et al. 2016; Kalyaanamoorthy et al. 2017; Lanfear et al. 2017). Model selection is considered a part of due diligence and good practice in these applications. However, an improvement in general goodness of fit may not translate into a quantifiable impact on the quantities of interest. For example, Spielman and Wilke (2015) found that, in the context of models of codon evolution (i.e., dN/dS-based models), *AIC* and *BIC* can positively mislead one to prefer a model with empirically worse *rate* estimates, while relatively poorly fit models in fact may produce the most accurate measures of selection strength. An alternative avenue for model selection is the use of posterior predictive distributions, in a Bayesian context (Gelman et al. 2013). We suggest that such avenues, which are starting to gain some traction in phylogenetic modeling (Bollback 2002; Rodrigue et al. 2009; Brown 2014; Lewis et al. 2014; Duchene et al. 2016), may prove more reliable than the use of theoretic information criteria for assessing the fit of evolutionary models to sequence data.

Phylogenetic modeling assumptions commonly made for the sake of inferential tractability (e.g., site independence, using a fixed topology inferred from the same data for rate analysis, or stationarity and time-homogeneity of the substitution process) are not biologically justifiable, but they are tolerated because they produce biologically meaningful inferences. As George E.P. Box wrote, "*Since all models are wrong the scientist must be alert to what is importantly wrong*" (Box 1976). Our evidence is that choice of substitution model is "mostly harmless" (Adams 1979) for the purposes of site-level rate inference, and it is not necessary to chase the elusive best fit model for each protein alignment.

Materials and Methods

Data Collection and Processing

We collected alignments from four distinct classes of proteins: enzymes, Metazoan mitochondrial data, green land plant chloroplast data, and mammalian G protein-coupled receptor data (GPCRs).

For the enzyme data set, we randomly selected 100 alignments with at least 25 unique sequences and corresponding phylogenies from Jack et al. (2016) for analysis. We prepared the mitochondrial and chloroplast data sets as follows. First, we compiled a list of complete Metazoan mitochondrial and green land plant genomes from the NCBI genomes database, of which there were 7,515 and 1,026, respectively. For each data set, we randomly chose 350 taxa to include in analysis. For these taxa, we obtained all genomic protein sequences from NCBI. After discarding sequences with amino acid ambiguities, and retaining only genes for which at least 100 taxa contained full-length sequences, we made gene-specific alignments using MAFFT v7.305b (Katoh and Standley 2013) and inferred phylogenies using FastTree2 (Price et al. 2010) with the LG+F substitution model (Le and Gascuel 2008).

Finally, we retrieved 227 alignments of mammalian GPCRs analyzed by Spielman and Wilke (2013), filtered to include at least 20 sequences. We reconstructed phylogenies for these alignments using FastTree2 (Price et al. 2010) with the LG+F model (Le and Gascuel 2008), as the original phylogenies had been constructed using masked alignments with reduced information.

Rate Inference

We then inferred relative protein evolutionary rates with the LEISR (Pupko et al. 2002; Spielman and Kosakovsky Pond 2018) method in HyPhy version 2.3.8 (Kosakovsky Pond et al. 2005), using seven different amino-acid evolutionary models. We used three generalist models: LG (Le and Gascuel 2008), WAG (Whelan and Goldman 2001), JTT (Jones et al. 1992), three specialist models: mtMet (Metazoan mitochondrial, Le et al. 2017), gcpREV (green-plant chloroplast, Cox and Foster 2013), and HIVb (between-host HIV-1, Nickle et al. 2007), and the equal-rates JC model (Jukes and Cantor 1969). We inferred rates under each model with (+*G*) and without a four-category discrete gamma distribution to model site heterogeneity during branch length optimization. All inferences used empirical (+*F*) equilibrium residue frequency estimates. We processed LEISR output for subsequent analysis using the Python helper package phyphy (Spielman 2018). All analyses interpreted the raw relative rates returned by LEISR, that is, rates were neither normalized nor standardized in any way. In addition, we right-censored any inferred rate with a maximum-likelihood estimate (MLE) of $MLE \geq 1{,}000$ to all have the value $MLE = 1{,}000$, because this value is effectively the numerical infinity in this context. We quantified estimation error of individual relative rates using profile likelihood, tabulating approximate 95% confidence intervals, using the critical values of the χ^2_1 distribution.

Statistical Analysis Availability

Data analysis was primarily conducted in the R programming language (R Core Team 2017), with substantial use of the tidyverse (Wickham 2017) suite of data analysis and visualization tools. All code and data associated with this work are freely available from the GitHub repository https://github.com/sjspielman/protein_rates_models, last accessed June 30, 2018.

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported in part by grants R01 GM093939 (NIH/NIGMS) and U01 GM110749 (NIH/NIGMS).

References

Adachi J, Hasegawa M. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput Sci Monogr*. 28:1–150.

Adams D. 1979. The hitchhiker's guide to the galaxy. 1st Am. ed. New York: Harmony Books.

Almeida F, Sanchez-Gracia A, Walden K, Robertson H, Rozas J. 2015. Positive selection in extra cellular domains in the diversification of *Strigamia maritima* chemoreceptors. *Front Ecol Evol*. 3:79.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*. 18(8):1585–1592.

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. 2016. Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*. 44(W1):W344–W350.

Bollback J. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol*. 19(7):1171–1180.

Box GEP. 1976. Science and statistics. *J Am Stat Assoc*. 71(356):791–799.

Brown JM. 2014. Predictive approaches to assessing the fit of evolutionary models. *Syst Biol*. 63(3):289–292.

Cox CJ, Foster PG. 2013. A 20-state empirical amino-acid substitution model for green plant chloroplasts. *Mol Phylogenet Evol*. 68(2):218–220.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164–1165.

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog*. 4(12):e1000242.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform*. 10(1):97–109.

Duchene S, Di Giallonardo F, Holmes EC. 2016. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Mol Biol Evol*. 33(1):255–267.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*. 17(2):109–121.

Fernandes AD, Atchley WR. 2008. Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. *Bioinformatics* 24(19):2177–2183.

Garcia-Boronat M, Diez-Rivero CM, Reinherz EL, Reche PA. 2008. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic Acids Res*. 36(Web Server):W35–W41.

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. 2013. Bayesian data analysis. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.

Goldstein RA, Pollock DD. 2016. The tangled bank of amino acids. *Prot Sci*. 25(7):1354–1362.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 89(22):10915–10919.

Huelsenbeck J, Joyce P, Lakner C, Ronquist F. 2008. Bayesian analysis of amino acid substitution models. *Philos Trans R Soc B* 363(1512):3941–3953.

Jack BR, Meyer AG, Echave J, Wilke CO. 2016. Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biol*. 14(5):e1002452.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 8(3):275–282.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. 3rd ed. New York: Academic Press. p. 21–132.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14(6):587–589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. 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 Evol Biol*. 6:29.

Landau M, Mayrose T, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. 2005. Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. *Nuclear Acids Res*. 33(Web Server):W299–W302.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol*. 34(3):772–773.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*. 25(7):1307–1320.

Le SQ, Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol*. 59(3):277–287.

Le VS, Dang CC, Le QS. 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. *BMC Evol Biol*. 17(1):136.

Lewis PO, Xie W, Chen MH, Fan Y, Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. *Syst Biol*. 63(3):309–321.

Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*. 21(9): 1781–1791.

Mirsky A, Kazandjian L, Anisimova M. 2015. Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. *Mol Biol Evol*. 32(3):806–819.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. 2007. HIV-specific probabilistic models of protein evolution. *PLoS One* 2(6):e503.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenetics. *Bioinformatics* 21(5):676–679.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over Likelihood Ratio Tests. *Syst Biol*. 53(5):793–808.

Price MN, Dehal PS, Arkin AP. 2010. FastTree2: approximately maximum–likelihood trees for large alignments. *PLoS One* 5(3): e9490.

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77.

Pupko T, Pe I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*. 17(6):890–896.

R Core Team. 2017. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of codong sequence

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018

evolution with dependence between codons. *Mol Biol Evol*. 26(7):1663–1676.
Scheffler K, Murrell B, Kosakovsky Pond SL. 2014. On the validity of evolutionary models with site-specific parameters. *PLoS One* 9(4):e94534.
Spielman SJ. 2018. phyphy: Python package for facilitating the execution and parsing of HyPhy standard analyses. *J Open Source Softw*. 3(21):514.
Spielman SJ, Kosakovsky Pond SL. 2018. Relative evolutionary rate inference in HyPhy with LEISR. *PeerJ* 6:e4339.
Spielman SJ, Wilke CO. 2013. Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. *J Mol Evol*. 76(3): 172–182.
Spielman SJ, Wilke CO. 2015. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 32(4):1097–1108.
Stevens TJ, Arkin IT. 2001. Substitution rates in alpha-helical transmembrane proteins. *Prot Sci*. 10(12):2507–2517.
Sydykova D, Jack B, Spielman S, Wilke C. 2018. Measuring evolutionary rates of proteins in a structural context [version 2; referees: 4 approved]. *F1000Research* 6(1845):1845.
Sydykova DK, Wilke CO. 2017. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. *PeerJ* 5:e3391.
Tusche C, Steinbrück L, McHardy AC. 2012. Detecting patches of protein sites of influenza A viruses under positive selection. *Mol Biol Evol*. 29(8):2063–2071.
Uzzell T, Corbin KW. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172(3988):1089–1096.
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(5):691–699.
Wickham H. 2017. tidyverse: easily install and load the 'Tidyverse'. R package version 1.2.1. https://www.tidyverse.org/
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*. 11(9):367–372.

Downloaded from https://academic.oup.com/mbe/article-abstract/35/9/2307/5040133 by Rowan University user on 07 September 2018