Gene 865 (2023) 147336

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Influence of substitution model selection on protein phylogenetic tree reconstruction

Roberto Del Amparo [a,b], Miguel Arenas [a,b,c,*]

[a] CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[b] Department of Biochemistry, Genetics and Immunology, Universidade de Vigo, 36310 Vigo, Spain
[c] Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain

ARTICLE INFO

Edited by: Lakshminarayan M. Iyer

Keywords:
Substitution models of protein evolution
Substitution model selection
Molecular evolution
Phylogenetic tree reconstruction
Protein evolution
Phylogenetics

ABSTRACT

Probabilistic phylogenetic tree reconstruction is traditionally performed under a best-fitting substitution model of molecular evolution previously selected according to diverse statistical criteria. Interestingly, some recent studies proposed that this procedure is unnecessary for phylogenetic tree reconstruction leading to a debate in the field. In contrast to DNA sequences, phylogenetic tree reconstruction from protein sequences is traditionally based on empirical exchangeability matrices that can differ among taxonomic groups and protein families. Considering this aspect, here we investigated the influence of selecting a substitution model of protein evolution on phylogenetic tree reconstruction by the analyses of real and simulated data. We found that phylogenetic tree reconstructions based on a selected best-fitting substitution model of protein evolution are the most accurate, in terms of topology and branch lengths, compared with those derived from substitution models with amino acid replacement matrices far from the selected best-fitting model, especially when the data has large genetic diversity. Indeed, we found that substitution models with similar amino acid replacement matrices produce similar reconstructed phylogenetic trees, suggesting the use of substitution models as similar as possible to a selected best-fitting model when the latter cannot be used. Therefore, we recommend the use of the traditional protocol of selection among substitution models of evolution for protein phylogenetic tree reconstruction.

1. Introduction

Phylogenetic reconstructions are common analyses for understanding multiple biological processes such as the evolution of genes and proteins (Lijavetzky et al. 2003), the emergence of protein function change (Pellegrini et al. 1999; Pascual-García et al., 2010), the molecular clock and its violations (Pascual-García et al. 2019), the strength of selection (Dutheil et al. 2012; Del Amparo et al. 2021) and the stability and function of ancestral sequences (Liberles 2007), among others. Although some of the first phylogenetic reconstruction methods based on the maximum parsimony approach ignored patterns of substitution among character states (Fitch 1971), subsequent [and more accurate (Zhang and Nei 1997)] methods based on probabilistic approaches [i.e., maximum likelihood (ML) or Bayesian inference (Felsenstein 1988; Nascimento et al. 2017; Posada and Crandall 2021)] considered these patterns by implementing substitution models of molecular evolution (Arenas 2015) into the likelihood function (Yang 2006). For more than 20 years, a variety of works found that phylogenetic tree reconstructions based on probabilistic approaches should consider a substitution model of evolution that best fit with the study data to obtain accurate inferences in terms of likelihood (Yang et al. 1994; Zhang and Nei 1997; Zhang 1999; Minin et al. 2003; Lemmon and Moriarty 2004). Because of

Abbreviations: CATH, protein structure classification database; Dayhoff, empirical substitution model of protein evolution based on nuclear proteins; HIV, human immunodeficiency virus; HIVw, empirical substitution model of protein evolution based on HIV proteins; JTT, empirical substitution model of protein evolution based on nuclear proteinss; LG, empirical substitution model of protein evolution based on nuclear proteinst; NMR, nuclear magnetic resonance; ML, maximum likelihood; MtArt, empirical substitution model of protein evolution based on Arthropoda mitochondrial proteins; MtMam, empirical substitution model of protein evolution based on Mammalia mitochondrial proteins; PDB, protein data bank; PFAM, protein families database; +F, empirical amino acid frequencies; rtREV, empirical substitution model of protein evolution based on retrovirus proteins; WAG, empirical substitution model of protein evolution based on nuclear proteinsy; +I, proportion of invariable sites; +G, variation of the substitution rate among sites according to a gamma distribution.

* Corresponding author at: Department of Biochemistry, Genetics and Immunology, Universidade de Vigo, 36310 Vigo, Spain.
E-mail addresses: rdelamparo@uvigo.es (R. Del Amparo), marenas@uvigo.es (M. Arenas).

https://doi.org/10.1016/j.gene.2023.147336
Received 4 January 2023; Received in revised form 22 February 2023; Accepted 28 February 2023
Available online 3 March 2023

that, the selection of the best-fitting substitution model [among a set of substitution models implemented in well-established likelihood-based frameworks such as *ModelTest* (Posada and Crandall 1998) and *ProtTest* (Abascal et al. 2005), among others (Kalyaanamoorthy et al. 2017; Lefort et al. 2017)] is traditionally included in the protocol of probabilistic phylogenetic tree reconstruction (Anisimova et al. 2013). Nevertheless, a few recent studies suggested that the selection of a substitution model of evolution is unnecessary due to observing similar phylogenetic trees reconstructed under different substitution models. Most of those studies focused on DNA models (Abadi et al. 2019; Tao et al. 2020). At the DNA level, the probability of correctly assigning by chance a particular nucleotide is high (1/4) compared to that at the protein level (1/20). As a consequence, at the nucleotide level one could use an uninformative substitution model but still obtain the correct nucleotide state by chance (probability of 1/4) while that would be much less likely at the amino acid level (probability of 1/20). This suggests that the consequences of substitution model selection could be detected easier at the amino acid level, since at this level obtaining the correct state is less likely by chance and requires the information provided by the model. At the protein level, the study (Spielman 2020) is, to our knowledge, the only one suggesting that the selection of a substitution model of protein evolution is not required for proper phylogenetic tree reconstruction. However, we noted that the methodology applied in that study could be biased (see Discussion). On the other hand, some studies indicated that the development and application of new substitution models is crucial to obtain more realistic ancestral protein reconstructions (Arenas et al., 2015; Arenas and Bastolla, 2020; Duchêne et al., 2016; Sumner et al., 2012).

In order to shed light on this topic, here we evaluated the consequences of selecting among well-established empirical substitution models of protein evolution, which are traditionally used in protein phylogenetics, on phylogenetic tree reconstruction. We applied the traditional pipeline used for validation in phylogenetics and that included the analysis of molecular data simulated upon previously simulated evolutionary histories (Hoban et al. 2012; Arenas, 2012) and the analysis of a variety of real protein families. We also explored how the genetic diversity of the study data could affect the consequences of substitution model selection on protein phylogenetic tree reconstruction. Overall, we found that the accuracy of the reconstructed phylogenetic trees increases when using a best-fitting substitution model and models similar to that best-fitting model, particularly for data with large genetic diversity.

2. Materials and methods

In this section we present the methodologies to evaluate the influence of substitution model selection on protein phylogenetic tree reconstruction by the analyses of real and simulated protein data.

2.1. Analysis of the influence of substitution model selection on phylogenetic tree reconstruction using real protein families

Following a previous study (Arenas et al. 2015), we downloaded 12 protein families from the PFAM database (Finn et al. 2014) with curated sequences (seed) that present variable sequence length, sample size and sequence identity (Table 1). Protein families from PFAM present homology and thus, in contrast to other databases, are convenient to perform phylogenetic tree reconstructions. As a prudent procedure, we realigned the sequences with *MAFFT* (Katoh and Standley 2013). For every multiple sequence alignment, we identified the best and worst-fitting substitution models of protein evolution implemented in the framework *ProtTest3* (Darriba et al. 2011) (Table 1). Next, we inferred ML phylogenetic trees with *RAxML-NG* (Kozlov et al. 2019) under the selected best-fitting, worst-fitting, JTT (Jones et al. 1992) (which is one of the most frequently used substitution model), HIVw (Nickle et al. 2007) and MtMam (Yang et al. 1998) substitution models, which include a variety of close and distant substitution models in terms of relative substitution rates of their amino acid exchangeability matrices (see Figs. 1 and S1; Supplementary Material). Finally, we calculated the distance between phylogenetic trees inferred under the best and worst-fitting substitution models, the best-fitting and JTT substitution models and, HIVw and MtMam substitution models. For this we applied the well-established metrics *Robinson-Foulds* (Robinson and Foulds 1981), *Branch score* (Kuhner and Felsenstein 1994) and *K tree score* (Soria-Carrasco et al. 2007) that assess the distance between two phylogenetic trees. Notice that these metrics quantify discordances between phylogenetic trees in terms of topology (*Robinson-Foulds;* i.e., *comparing two trees, if every tree has 1 clade that is not present in the other tree then RF = 1 + 1 = 2*) or both topology and branch lengths (*Branch score and K tree score*) (Robinson and Foulds 1981; Kuhner and Felsenstein 1994; Soria-Carrasco et al. 2007). Indeed, they are based on different algorithms and thus their results can quantitatively differ. In addition to the 12 cited protein families, we applied the same procedure to analyze 200 real protein families from (Spielman 2020) by comparing phylogenetic trees inferred under the best-fitting, the worst-fitting and JTT substitution models.

In a subsequent analysis based on the 12 protein families (Table 1), we evaluated the particular influence of considering the empirical amino acid frequencies (+F) and models of sequence evolution by the proportion of invariable sites (+I) and/or the variation of the substitution rate among sites according to a gamma distribution (+G) (Yang 1994) on the phylogenetic tree reconstruction. In particular, we calculated the distance between phylogenetic trees inferred under only the exchangeability matrix of the best-fitting substitution model and under

Table 1
Real protein families analyzed in this study. For each studied protein family the table shows the PFAM code, number of sequences, sequence length (number of amino acids), sequence identity (average of sequence identities from all the pairs of sequences of the corresponding multiple sequence alignment, ranging from 0 to 1) and, best-fitting and worst-fitting substitution models selected with *ProtTest3*. We also analyzed 200 real protein families presented in (Spielman 2020).

Protein family	PFAM code	Number of sequences	Sequence length	Sequence identity	Best-fitting substitution model	Worst-fitting substitution model
Ferredoxin	PF05996	26	278	0.27	LG + I + G	MtMam
Cytochrome P450	PF00067	50	597	0.31	LG + I + G	MtMam
Triosephosphate isomerase	PF00121	41	275	0.34	LG + I + G	MtMam
Retroviral aspartil protease	PF00077	8	123	0.35	rtREV + G	MtMam
Glucokinase	PF02685	36	393	0.45	LG + I + G	MtArt
Pancreatic ribonuclease	PF00074	149	154	0.46	JTT + I + G	MtArt
Rubredoxin	PF00301	24	52	0.53	WAG + I + G	MtMam
Heat shock protein	PF00012	27	691	0.55	LG + G	MtMam
Oxysterol-binding protein	PF01237	363	1144	0.64	LG + I + G	MtMam
Homogentisate 1,2-dioxygenase	PF04209	11	443	0.64	LG + G	MtMam
Kinesin	PF00225	71	688	0.66	LG + I + G	MtMam
DNA ligase	PF13298	159	194	0.67	LG + I + G	MtArt

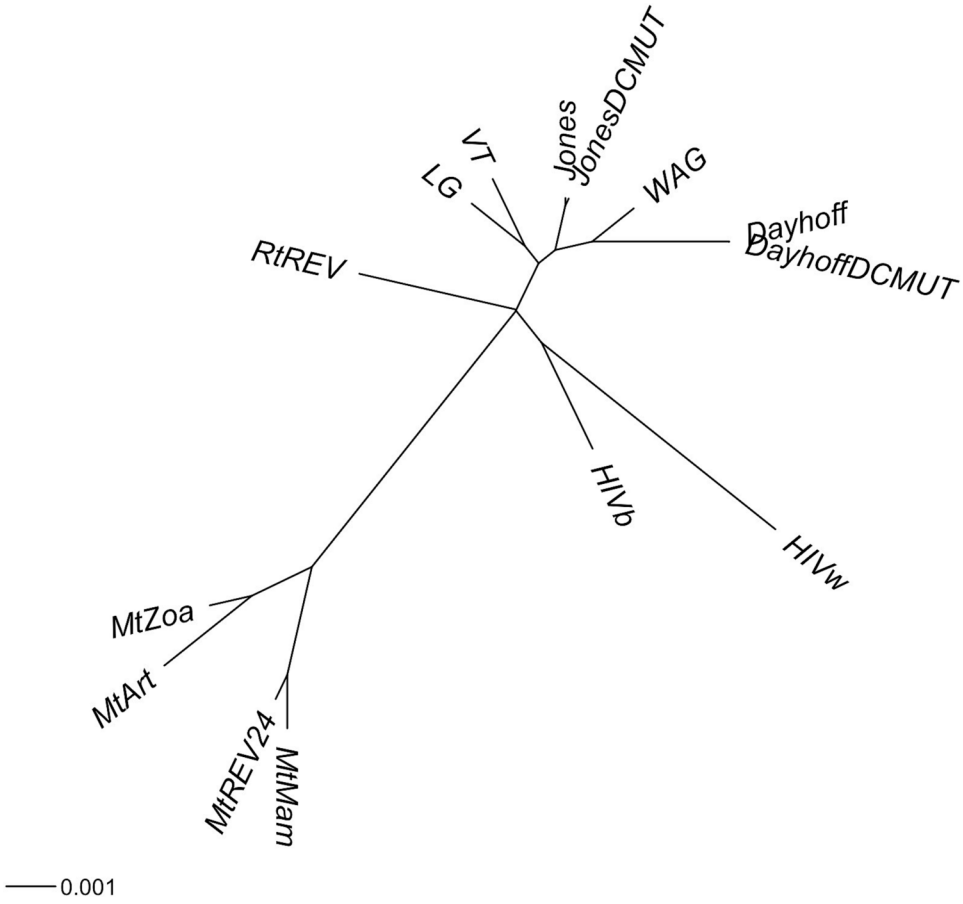

Fig. 1. Agglomerative clustering of common empirical substitution models of protein evolution. We normalized the exchangeability matrix and amino acid frequencies at the equilibrium of commonly used empirical substitution models of protein evolution. Next, we calculated distances between the exchangeability matrices and amino acid frequencies among every pair of substitution models. Finally, we applied the bottom-up agglomerative clustering method neighbor joining. Thus, the figure shows the clustering based on the distance between the normalized exchangeability matrices and the amino acid frequencies (with same weight) of the substitution models. Clusters only based on the exchangeability matrix or the amino acid frequencies of the substitution models are presented in Figures S1 and S2 (Supplementary Material), respectively.

the exchangeability matrix with + F, +I, +G, +I + G and + F + I + G.

2.2. Analysis of the influence of substitution model selection on phylogenetic tree reconstruction using simulated protein data

We used simulated data to evaluate the distance between phylogenetic trees inferred under the true (simulated) substitution model and phylogenetic trees inferred under other (close or far from the true comparing their exchangeability matrices; Figs. 1 and S1) substitution models. First we simulated phylogenetic trees with random topologies using the function *rtree* implemented in the library *ape* of *R* (Paradis et al. 2004). Next, for each simulated tree, we simulated protein sequence evolution (we assumed a sequence length of 250 amino acids, which is a length commonly observed in nature, as shown in Table 1, and presented sufficient molecular signatures of evolutionary patterns to distinguish between substitution models) under a particular substitution model with the function *simSeq* implemented in the *phangorn* library of *R* (Schliep 2011). We arbitrarily applied the HIVw substitution model (Nickle et al. 2007) (true model) in the simulations (but see later simulations under other models). We evaluated the influence of substitution model selection on phylogenetic tree reconstruction in 6 evolutionary scenarios that included simulated data with different number of protein sequences (50 and 100) and variable sequence identity (pairwise sequence comparisons, 0.2, 0.5 and 0.8). For each evolutionary scenario, we simulated a total of 100 multiple sequence alignments. As a control check of our simulations of protein evolution, we applied *ProtTest3* to verify that the true substitution model was selected as the best-fitting substitution model (among all the substitution models implemented in the framework) in all the simulated data. Next, for each simulated dataset, we reconstructed ML phylogenetic trees with *RAxML-NG* under the HIVw (true model), HIVb, JTT, Dayhoff, MtMam and MtArt substitution models. These substitution models were selected due to their common use in phylogenetics (i.e., JTT) and, close (i.e., HIVw), intermediate (i.e., Dayhoff) or far (i.e., MtMam) distance of their exchangeability matrices (Figs. 1 and S1) with that of the true substitution model. Finally, we obtained the distance between the simulated (true) phylogenetic tree and the phylogenetic trees reconstructed under each substitution model with the metrics *Robinson-Foulds*, *Branch score* and *K tree score*. We applied the *Wilcoxon signed-rank* test to evaluate statistical significance among estimates from different substitution models.

Using the same methodology, we additionally explored the particular influence of considering and ignoring the variation of the substitution rate among sites according to a gamma distribution (+G) on the phylogenetic tree reconstruction. In particular, we simulated protein sequences under the JTT substitution model with and without + G. Next, we inferred phylogenetic trees under the true substitution model, the true substitution model with or without + G and, under other substitution models (also with and without + G) such as Dayhoff (which is a model similar to JTT; Fig. 1) and, HIVw and MtMam (both models are distant from JTT; Figs. 1 and S1).

3. Results

3.1. Evaluation of selecting a substitution model of evolution for phylogenetic tree reconstruction based on real protein families

The analysis of real protein families (Table 1) showed that phylogenetic trees reconstructed under substitution models of protein evolution with distant exchangeability matrices always differ to a greater or lesser extent (Fig. 2). For example, the distance between phylogenetic trees inferred under the selected best and worst-fitting substitution

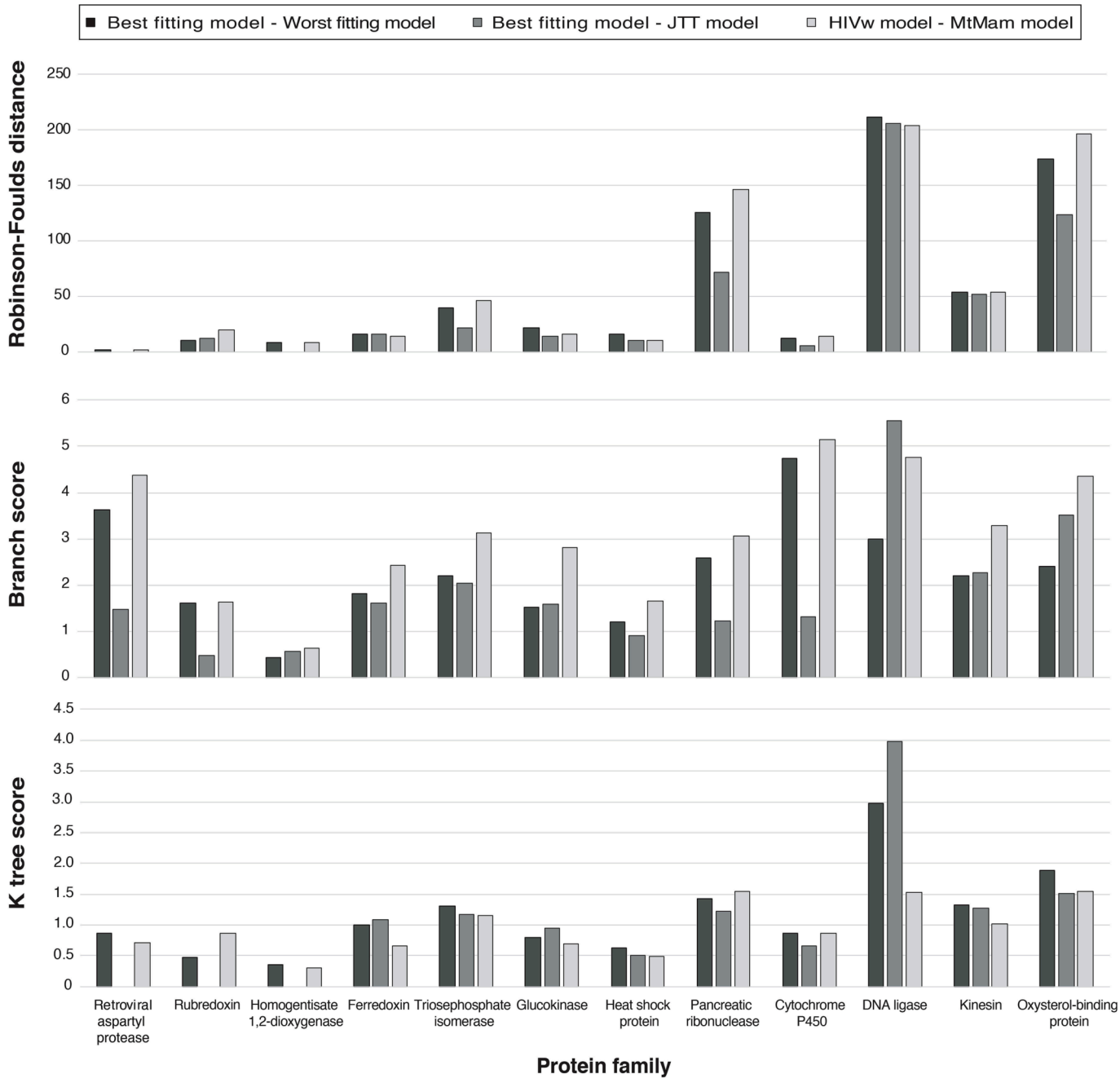

Fig. 2. Influence of substitution model selection on phylogenetic tree reconstruction using real data. Distances between phylogenetic trees reconstructed under the selected best-fitting and worst-fitting substitution models (Table 1) (black bars), best-fitting and JTT substitution models (dark grey bars) and, HIVw and MtMam substitution models (clear grey bars).

models, or between phylogenetic trees inferred under HIVw and MtMam substitution models, was above 0 for every metric. The differences between phylogenetic trees not only involved branch lengths, but also topology as shown with the *Robinson-Foulds* metric and in the illustrative examples presented in Figs. S3-S10 that visually highlight differences between phylogenetic trees reconstructed under the best and the worst fitting substitution models for the studied protein families presenting 50 or less sequences (selected for simplicity; Supplementary Material). We only found similar phylogenetic trees inferred under different substitution models in 3 out of 12 protein families, in particular when the phylogenetic trees were reconstructed under substitution models with similar exchangeability matrices (i.e., JTT and a best-fitting model that is close to JTT; Fig. 2). Concerning the analyses of the 200 protein families from (Spielman 2020), we found that all of them showed that phylogenetic trees inferred under different substitution models differ using any phylogenetic tree discordance metric (Fig. S11; Supplementary Material), thus indicating differences in terms of both topology and branch lengths. Finally, we explored the particular influence of empirical amino acid frequencies (+F) and models of sequence evolution by including the parameters proportion of invariable sites (+I) and variation of the substitution rate among sites according to a gamma distribution (+G) in the phylogenetic tree reconstruction. In general, we found that + G affects reconstructed phylogenetic trees (especially branch lengths) to a greater extent than the other studied parameters (+F and + I) (Fig. S12; Supplementary Material), but this observation should be carefully interpreted because this influence varies among data

(see Discussion).

3.2. Evaluation of selecting a substitution model of evolution for phylogenetic tree reconstruction based on simulated protein data

We evaluated the error of inferring a phylogenetic tree with a substitution model that is not the true model using computer simulations. In general, we found that phylogenetic trees inferred under the true (simulated) substitution model are more similar to the true phylogenetic trees than phylogenetic trees inferred under other substitution models (Figs. 3 and S13; Supplementary Material). However and importantly, the effects of substitution model selection on the reconstructed phylogenetic trees were affected by the molecular diversity of the study data. Specifically, for data simulated with large genetic diversity (i.e., average of pairwise sequence identity around 0.2), phylogenetic trees inferred with a substitution model that is close (in terms of relative substitution rates of the exchangeability matrix) to the true substitution model present topologies significantly similar to those from the true phylogenetic tree, while branch lengths are significantly different (first column in Figs. 3 and S13). Indeed, phylogenetic trees inferred with a substitution model that is far from the true substitution model present both topology and branch lengths statistically different to those obtained using the true substitution model in the inference (first column in Figs. 3 and S13). For example, for proteins simulated with that low sequence identity, phylogenetic trees reconstructed under the HIVw substitution model (true model) are significantly different (Wilcoxon signed-rank test $P < 0.05$) from phylogenetic trees reconstructed under the Dayhoff, MtMam and MtArt substitution models in terms of topology and under HIVb, JTT, Dayhoff, MtMam and MtArt substitution models in terms of branch length (first column in Figs. 3 and S13). When the average of sequence identity of the simulated data is around 0.5 (which is common in protein families, i.e. Table 1), only substitution models far from the true substitution model produced phylogenetic trees with topology and branch lengths significantly different to those of the phylogenetic trees reconstructed under the true substitution model (second column in Figs. 3 and S13). Under high levels of sequence identity (i.e., 0.8), only substitution models far from the true substitution model produced phylogenetic trees with branch lengths significantly different to those of the phylogenetic trees reconstructed under the true substitution model. Indeed, at this high level of sequence identity, phylogenetic trees reconstructed under different substitution models did not present significantly different topologies (last column in Figs. 3 and S13). Therefore, data with low sequence identity were statistically more sensitive to the selection of the substitution model than data with high sequence identity (Figs. 3 and S13). If the sequence identity is high, ignoring substitution model selection does not affect significantly the accuracy of the reconstructed topologies, although branch lengths could still significantly differ for distant substitution models (Figs. 3 and S13). Next, despite one could hypothesize that increasing the number of sequences (while maintaining genetic diversity) of the input data can lead to a more pronounced effect of substitution model selection on the accuracy of the reconstructed phylogenetic trees (because of dealing with more nodes and branches), we found that data with different number of sequences qualitatively displayed similar consequences of substitution model selection on the reconstructed phylogenetic trees (compare Figs. 3 and S13). Finally, we evaluated the influence of considering and ignoring the substitution rate variation among sites according to a gamma distribution (+G) on phylogenetic tree reconstruction. We found that accounting for + G mainly affects branch lengths of the reconstructed trees (the topology was only slightly affected), especially in data presenting low sequence identity (Figs. S14-S17; Supplementary Material).

4. Discussion

Selecting and applying the best-fitting substitution model of evolution to perform phylogenetic tree reconstruction is a well-established protocol in the field. It is based on the intuitive reasoning of the

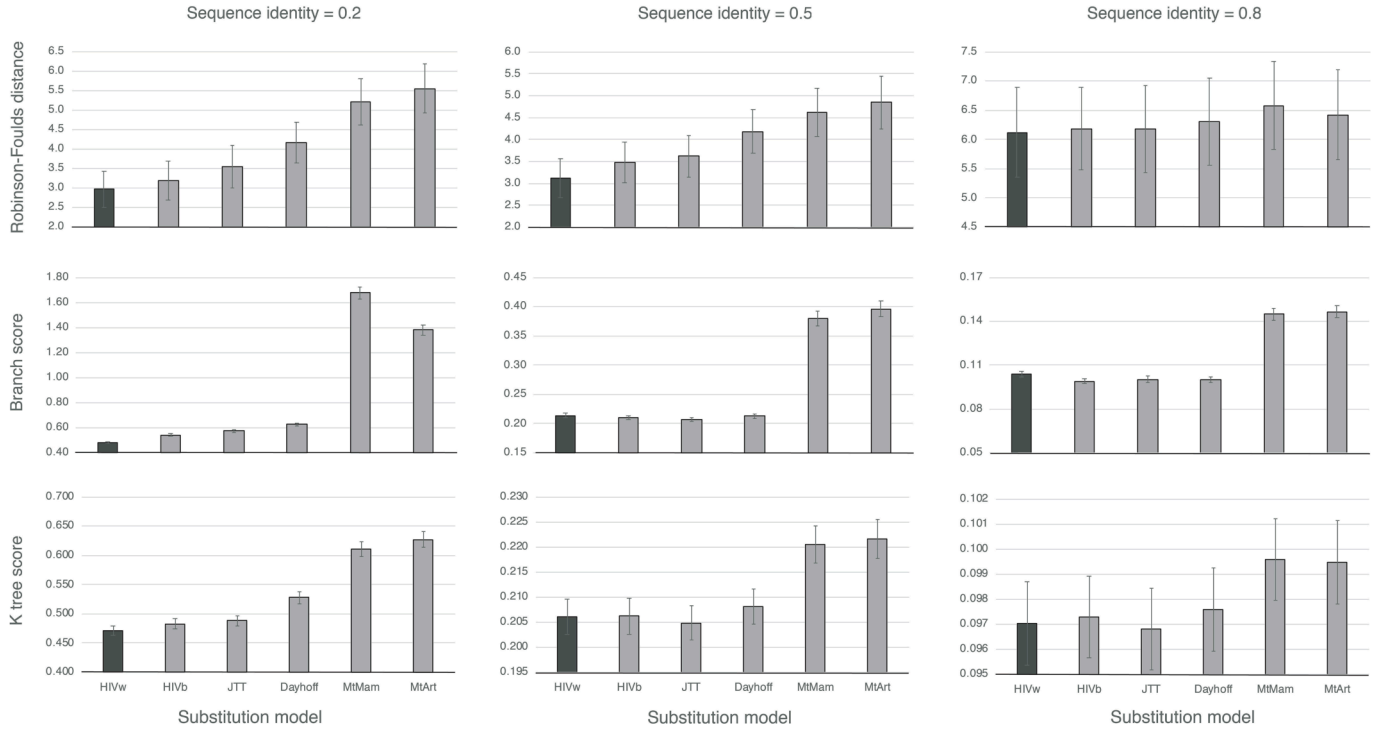

Fig. 3. Influence of substitution model selection on phylogenetic tree reconstruction using simulated data. Distances between true phylogenetic trees and phylogenetic trees reconstructed under HIVw (true model, shown in dark grey), HIVb, JTT, Dayhoff, MtMam and MtArt substitution models. The study is based on simulated datasets presenting 50 protein sequences. Error bars indicate 95% confidence interval of the mean over 100 simulations. Results for datasets simulated with 100 sequences are presented in Figure S13.

evolutionary consequences of the type of substitution events [i.e., more than 50 years ago (Zuckerkandl and Pauling, 1965) indicated that '*it is the type rather than number of amino acid substitutions that is decisive*'] and it was supported by a number of likelihood-based studies published since more than 20 years ago (Yang et al. 1994; Zhang and Nei 1997; Zhang 1999; Minin et al. 2003; Lemmon and Moriarty 2004). However, the use of substitution model selection in phylogenetic tree reconstruction was recently explored in a few works that evaluated DNA (Abadi et al. 2019; Tao et al. 2020) and protein (Spielman 2020) data to indicate that accounting for this protocol has little effect on the accuracy of the topology of reconstructed phylogenetic trees. We noted that those findings are particularly surprising for proteins due to the variety of selection processes caused by functional constraints (Fay and Wu 2003; Chi and Liberles 2016) and observed in the large diversity of currently available empirical substitution models (Thorne 2000; Arenas 2015). Thus, here we investigated the influence of substitution model misspecification on protein phylogenetic tree reconstruction. We focused on protein evolution for several reasons. First, some data can only be available as protein sequences because most of protein sequencing methodologies (i.e., based on NMR and X-ray diffraction) and protein databases (i.e., PFAM, PDB and CATH, among others) do not include DNA sequences. Second, a variety of empirical substitution models of protein evolution have been inferred from different taxonomic groups and protein families. These models include a 20×20 exchangeability matrix of relative rates of change among amino acids, and amino acids frequencies, that can differ among models at a greater or lesser extent (Fig. 1). Next, the number of states is higher at the protein level than at the DNA level and thus assignments of a correct state by chance could be more frequent when evaluating DNA sequences than protein sequences. Indeed, phylogenetic analyses based on protein sequence comparisons allow to extend well beyond the nucleotide saturation distances. Finally, the influence of substitution model selection on phylogenetic tree reconstruction was explored for DNA sequences in some studies (Abadi et al. 2019; Tao et al. 2020), while for protein sequences it was much less investigated (Spielman 2020).

First, we performed the phylogenetic tree reconstruction of 12 protein families under different substitution models and, for all of them, we found clear differences (in terms of topology and branch lengths) between phylogenetic trees inferred especially under distant empirical substitution models of evolution. Indeed, the analysis of the 200 protein families from (Spielman 2020) also showed differences between phylogenetic trees reconstructed under different models. These findings already show that substitution model misspecification can affect both topology and branch lengths of the reconstructed phylogenetic trees. This could not be the case for real datasets that are not analyzed in the present study, but still we show that the bias can occur and, therefore, selecting and applying a best-fitting substitution model is recommended. When comparing phylogenies reconstructed under close empirical substitution models, we found that a total of 209 out of 212 (98.6%) studied protein families produced different phylogenetic trees. Despite these findings could be expected, we believe that they should be demonstrated and especially considering the conclusions made in a previous study (discussed later). Our results suggest (also verified with computer simulations, which is discussed later) that using substitution models as similar as possible to a selected best-fitting model is recommended when that best-fitting substitution model cannot be used for any reason (i.e., it is not implemented in the evolutionary framework of the phylogenetic tree reconstruction). Finally, we found that the substitution rate variation among sites according to the traditional gamma distribution (+G) especially affects branch lengths of the phylogenetic tree. The variation of the substitution rates among protein sites is common in nature (Baele et al. 2011; Pentinsaari et al. 2016; Jimenez-Santos et al. 2018) and accounting for it provides more realistic phylogenetic inferences as we, and others (Yang 1996; Jia et al. 2014), found analyzing real and simulated data. In summary, for the studied real protein families we found that selecting a best-fitting substitution model can affect the reconstructed protein phylogenetic trees.

Next, we simulated protein sequences to quantify the distance between phylogenetic trees inferred under different substitution models, including the true substitution model. In general, we found that phylogenetic tree reconstruction under the true substitution model produces the most accurate phylogenies compared to phylogenetic trees reconstructed under substitution models with distant amino acid replacement matrices. Indeed, in general, we found again that applying a substitution model close to the true substitution model results in more accurate phylogenetic trees than applying a substitution model that is far from the true substitution model. Actually, we observed a high correlation between the distance among the exchangeability matrices of the studied substitution models and the distance between the phylogenetic trees reconstructed under the corresponding substitution models (Table S1; Supplementary Material). In practice, this means that if the selected best-fitting substitution model is not available to perform a phylogenetic tree reconstruction, applying a substitution model as close as possible to the best-fitting substitution model is recommended. Again, despite these findings could be expected in advance, we believe that they should be formally demonstrated. Importantly, we found that the influence of the substitution model on phylogenetic tree reconstruction is affected by the genetic diversity of the study data. In particular, data presenting large genetic diversity produced phylogenetic trees more dependent on the applied substitution model than data with low genetic diversity; the latter can even produce phylogenetic trees insensitive to the selection of a substitution model. We believe that this result can be explained by the following intuitive reasoning. Data presenting low genetic diversity have accumulated less substitution events during their evolutionary histories. If substitutions are rare (short branches) the weight of the substitution model on the calculated likelihood is lower due to the reduced number of evolutionary pathways (Yang 2006), finally affecting the resulting phylogenetic tree. Moreover, these short branches can favor topology fluctuations caused by the probabilistic method of phylogenetic reconstruction and that are more frequent when the genetic diversity is small, reducing the signatures of the substitution model on the reconstructed phylogenetic tree. In long branches, where the number of substitutions increases, different models can lead to more different estimates because of the presence of more evolutionary information.

Our conclusions partially differ from those presented in (Spielman 2020), where protein substitution model selection was suggested as generally unnecessary to obtain accurate phylogenetic tree topologies. Using real and simulated data we found that the topology of a reconstructed phylogenetic tree can be affected by the substitution model used for the reconstruction. Of course and in agreement with (Spielman 2020), we found that phylogenetic trees inferred under similar empirical substitution models, in datasets with low genetic diversity, may not differ [although one can also properly consider that minor differences in the topology of a phylogenetic tree could be dramatic for diverse evolutionary studies (e.g., Soltis and Soltis 2003; Davis et al. 2010; Pace et al. 2012; Moreira et al. 2021)]. However, in other scenarios (i.e., data with large genetic diversity) we found relevant influences of substitution model selection on protein phylogenetic tree reconstruction. We consider that some of our conclusions differ from those presented in (Spielman 2020) due to technical aspects and the interpretation of some results. First, that study simulated sequences upon phylogenetic trees reconstructed in previous studies from real data. That procedure can be problematic because those phylogenies were already reconstructed under particular substitution models [even after performing substitution model selection (Salichos and Rokas 2013; Ruhfel et al. 2014)] and thus could lead to phylogenies biased toward the originally applied substitution models as we show in the simulation section of the present study. Another technical aspect that could affect results from that study is that the input phylogeny branch lengths were scaled up by a factor of three to attempt to better fit branch lengths based on DNA sequences with the number of observed amino acid substitutions, but ignoring unfixed,

synonymous and nonsynonymous changes. These artificially longer branch lengths placed upon a fixed topology can reduce the impact of the substitution model on the tree topology (Sullivan et al. 1996; Kück et al. 2012) because in that situation the number of real substitutions can be too small to allow topology changes. Here, we avoided those potential biases by simulating phylogenetic trees without assuming a substitution model, which is a traditional procedure in population genetics (Hoban et al. 2012; Arenas, 2012) and that can be performed by simulating random, coalescent and birth–death phylogenetic trees [among other approaches to simulate phylogenies (Hoban et al. 2012; Arenas, 2012)]. In any case, in our opinion from a biological perspective, just a single change in a tree topology could be biologically relevant (e.g., Soltis and Soltis 2003; Wiley 2010; Som 2015).

Currently, around 100 empirical amino acid exchangeability matrices (several hundreds of substitution models in case of additionally considering + F, +I and + G) are available [and more continue being developed (e.g., Le et al. 2017; Chang et al. 2020; Le and Vinh 2020; Del Amparo and Arenas 2022)] due to the need of realistically mimic the evolution of the diverse protein families observed in nature. Indeed, real data often shows genetic signatures of complex substitution processes. For example, substitution patterns can vary among genomic regions (Arbiza et al. 2011) or the fitness of viruses can be intensely affected by particular amino acid substitution events that alter the protein stability and function (Lorenzo-Redondo et al. 2014; Arenas et al. 2016; Duchêne et al., 2016; Echave et al. 2016; Kirchner et al. 2017; Jimenez-Santos et al. 2018; Geoghegan and Holmes 2018).

In practice, the researcher may not know in advance the influence of applying a particular substitution model of evolution on the phylogenetic tree reconstruction for a certain dataset. Considering our findings, we recommend applying the selected best-fitting substitution model by default (if it is not possible, we recommend applying a substitution model as similar as possible to the selected best-fitting substitution model), thus following the traditional phylogenetics protocol. Indeed, the researcher can infer phylogenetic trees under different substitution models and compare them (i.e., following the procedure presented here for the study of real data) and, in case of observing phylogenetic tree discordances, we recommend applying a selected best-fitting substitution model.

The reconstruction of phylogenetic trees should be as realistic as possible if one aims to obtain reasonable biological conclusions and here we show that applying a proper substitution model of evolution can be crucial. We believe that future research on phylogenetic tree reconstruction from protein data will involve structurally constrained substitution (SCS) models that are more complex, but also more realistic, than the empirical substitution models implemented in most of currently available phylogenetic reconstruction frameworks (Bordner and Mittelmann 2013; Arenas et al. 2015), but efforts are still needed in this direction. Considering the results of this study we also believe that the development of more realistic substitution models of protein evolution is required to improve the accuracy of protein phylogenetic tree reconstructions.

CRediT authorship contribution statement

Roberto Del Amparo: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Miguel Arenas:** Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The real and simulated data are available at Zenodo repository from the URL https://doi.org/10.5281/zenodo.6377152.

Acknowledgments

We thank CESGA "*Centro de Supercomputación de Galicia*" for the computer resources. Funding for open access charge: Universidade de Vigo/CISUG.

Funding

This work was supported by the Spanish Ministry of Science and Innovation, Agencia Estatal de Investigación, through the Grant [PID2019-107931GA-I00/AEI/10.13039/501100011033].

***Data** Availability*

The real and simulated data are available at Zenodo repository from the URL https://doi.org/10.5281/zenodo.6377152.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2023.147336.

References

Abadi, S., Azouri, D., Pupko, T., Mayrose, I., 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nat. Commun. 10, 934–934.
Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21 (9), 2104–2105.
Anisimova, M., Liberles, D.A., Philippe, H., Provan, J., Pupko, T., von Haeseler, A., 2013. State-of the art methodologies dictate new standards for phylogenetic analysis. BMC Evol. Biol. 13 (1), 161.
Arbiza, L., Patricio, M., Dopazo, H., Posada, D., 2011. Genome-wide heterogeneity of nucleotide substitution model fit. Genome Biol. Evol. 3, 896–908.
Arenas, M., 2015. Trends in substitution models of molecular evolution. Front. Genet. 6, 319.
Arenas, M., Sánchez-Cobos, A., Bastolla, U., 2015. Maximum likelihood phylogenetic inference with selection on protein folding stability. Mol. Biol. Evol. 32 (8), 2195–2207.
Arenas, M., Bastolla, U., 2020. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. Methods Ecol. Evol. 11 (2), 248–257.
Arenas, M., Lorenzo-Redondo, R., Lopez-Galindez, C., 2016. Influence of mutation and recombination on HIV-1 in vitro fitness recovery. Mol. Phylogenet. Evol. 94, 264–270.
Arenas, M., 2012. Simulation of Molecular Data under Diverse Evolutionary Scenarios. PLoS Comput. Biol. 8, e1002495.
Baele, G., Van de Peer, Y., Vansteelandt, S., 2011. Context-dependent codon partition models provide significant increases in model fit in atpB and rbcL protein-coding genes. BMC Evol. Biol. 11, 145.
Bordner, A.J., Mittelmann, H.D., 2013. A new formulation of protein evolutionary models that account for structural constraints. Mol. Biol. Evol. 31, 736–749.
Chang, H., Nie, Y., Zhang, N., Zhang, X., Sun, H., Mao, Y., Qiu, Z., Huang, Y., 2020. MtOrt: an empirical mitochondrial amino acid substitution model for evolutionary studies of Orthoptera insects. BMC Evol. Biol. 20 (1).
Chi, P.B., Liberles, D.A., 2016. Selection on protein structure, interaction, and sequence. Protein Sci. 25 (7), 1168–1178.
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.
Davis, C.C., Willis, C.G., Primack, R.B., Miller-Rushing, A.J., 2010. The importance of phylogeny to the study of phenological response to global climate change. Philos. Trans. Roy. Soc. B Biol. Sci. 365 (1555), 3201–3213.
Del Amparo, R., Arenas, M., 2022. HIV Protease and Integrase Empirical Substitution Models of Evolution: Protein-Specific Models Outperform Generalist Models. Genes 13 (1), 61.
Del Amparo, R., Branco, C., Arenas, J., Vicens, A., Arenas, M., 2021. Analysis of selection in protein-coding sequences accounting for common biases. Brief Bioinform 22 (5), bbaa431.
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. Mol. Biol. Evol. 33 (1), 255–267.
Dutheil, J.Y., Galtier, N., Romiguier, J., Douzery, E.J.P., Ranwez, V., Boussau, B., 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol. 29, 1861–1874.

Echave, J., Spielman, S.J., Wilke, C.O., 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17 (2), 109–121.
Fay, J.C., Wu, C.-I., 2003. Sequence Divergence, Functional Constraint, and Selection in Protein Evolution. Annu. Rev. Genomics Hum. Genet. 4 (1), 213–235.
Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22 (1), 521–565.
Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42 (D1), D222–D230.
Fitch, W., 1971. Toward defining the course of evolution: minimal change for a specific tree topology. Syst. Zool. 20, 406–416.
Geoghegan, J.L., Holmes, E.C., 2018. The phylogenomics of evolving virus virulence. Nat. Rev. Genet. 19 (12), 756–769.
Hoban, S., Bertorelle, G., Gaggiotti, O.E., 2012. Computer simulations: tools for population and evolutionary genetics. Nat. Rev. Genet. 13 (2), 110–122.
Jia, F., Lo, N., Ho, S.Y.W., 2014. The Impact of Modelling Rate Heterogeneity among Sites on Phylogenetic Estimates of Intraspecific Evolutionary Rates and Timescales. Plos One 9, e95722.
Jimenez-Santos, M.J., Arenas, M., Bastolla, U., 2018. Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution. PeerJ 6, e5549.
Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8 (3), 275–282.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14 (6), 587–589.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780.
Kirchner, S., Cai, Z., Rauscher, R., et al., 2017. Alteration of protein function by a silent polymorphism linked to tRNA abundance. PLoS Biol. 15, e2000779.
Kozlov, A.M., Darriba, D., Flouri, T., et al., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35, 4453–4455.
Kück, P., Mayer, C., Wägele, J.-W., Misof, B., 2012. Long Branch Effects Distort Maximum Likelihood Phylogenies in Simulations Despite Selection of the Correct Model. Plos One 7, e36593.
Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.
Le, V.S., Dang, C.C., Le, Q.S., 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evol. Biol. 17, 136.
Le, T.K., Vinh, L.S., 2020. FLAVI: An Amino Acid Substitution Model for Flaviviruses. J. Mol. Evol. 88 (5), 445–452.
Lefort, V., Longueville, J.-E., Gascuel, O., 2017. SMS: Smart Model Selection in PhyML. Mol. Biol. Evol. 34, 2422–2424.
Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in bayesian phylogenetics. Syst. Biol. 53, 265–77.
Liberles, D.A., 2007. Ancestral Sequence Reconstruction. Oxford University Press.
Lijavetzky, D., Carbonero, P., Vicente-Carbajosa, J., 2003. Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families. BMC Evol. Biol. 3, 17.
Lorenzo-Redondo, R., Delgado, S., Moran, F., Lopez-Galindez, C., 2014. Realistic three dimensional fitness landscapes generated by self organizing maps for the analysis of experimental HIV-1 evolution. PLoS One 9, e88579.
Minin, V., Abdo, Z., Joyce, P., Sullivan, J., 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52, 674–683.
Moreira, F., Arenas, M., Videira, A., Pereira, F., 2021. Molecular Evolution of DNA Topoisomerase III Beta (TOP3B) in Metazoa. J. Mol. Evol. 89 (6), 384–395.
Nascimento, F.F., Reis, M.D., Yang, Z., 2017. A biologist's guide to Bayesian phylogenetic analysis. Nat. Ecol. Evol. 1, 1446–1454.
Nickle, D.C., Heath, L., Jensen, M.A., et al., 2007. HIV-specific probabilistic models of protein evolution. PLoS One 2, e503.
Pace, N.R., Sapp, J., Goldenfeld, N., 2012. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc. Natl. Acad. Sci. 109, 1011.
Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290.
Pascual-García, A., Abia, D., Méndez, R., Nido, G.S., Bastolla, U., 2010. Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins 78 (1), 181–196.
Pascual-García, A., Arenas, M., Bastolla, U., 2019. The Molecular Clock in the Evolution of Protein Structures. Syst. Biol. 68, 987–1002.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96 (8), 4285–4288.
Pentinsaari, M., Salmela, H., Mutanen, M., Roslin, T., 2016. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Sci. Rep. 6, 35275.
Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.
Posada, D., Crandall, K.A., 2021. Felsenstein Phylogenetic Likelihood. J. Mol. Evol. 89 (3), 134–145.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53 (1-2), 131–147.
Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., Burleigh, J., 2014. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol. Biol. 14 (1), 23.
Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497 (7449), 327–331.
Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593.
Soltis, D.E., Soltis, P.S., 2003. The role of phylogenetics in comparative genetics. Plant Physiol. 132, 1790–1800.
Som, A., 2015. Causes, consequences and solutions of phylogenetic incongruence. Brief. Bioinform. 16 (3), 536–548.
Soria-Carrasco, V., Talavera, G., Igea, J., Castresana, J., 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23 (21), 2954–2956.
Spielman, S.J., 2020. Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics. Mol. Biol. Evol. 37, 2110–2123.
Sullivan, J., Holsinger, K.E., Simon, C., 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. 42 (2), 308–312.
Sumner, J.G., Jarvis, P.D., Fernandez-Sanchez, J., et al., 2012. Is the general time-reversible model bad for molecular phylogenetics? Syst. Biol. 61, 1069–74.
Tao, Q., Barba-Montoya, J., Huuki, L.A., et al., 2020. Relative Efficiencies of Simple and Complex Substitution Models in Estimating Divergence Times in Phylogenomics. Mol. Biol. Evol. 37, 1819–1831.
Thorne, J.L., 2000. Models of protein sequence evolution and their applications. Curr. Opin. Genet. Dev. 10 (6), 602–605.
Wiley, E.O., 2010. Why Trees Are Important. Evol. Educ. Outreach 3 (4), 499–505.
Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39 (3), 306–314.
Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. 11, 367–372.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.
Yang, Z., Goldman, N., Friday, A., 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11, 316–324.
Yang, Z., Nielsen, R., Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15, 1600–1611.
Zhang, J., 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. 16 (6), 868–875.
Zhang, J., Nei, M., 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. 44 (Suppl 1), S139–S146.
Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166.