ModelTeller: Model Selection for Optimal Phylogenetic Reconstruction Using Machine Learning

Shiran Abadi,[1] Oren Avram,[2] Saharon Rosset,[3] Tal Pupko,[2] and Itay Mayrose*,[1]

[1]School of Plant Sciences and Food security, Tel-Aviv University, Tel-Aviv, Israel

[2]School of Molecular Cell Biology & Biotechnology, Tel-Aviv University, Tel-Aviv, Israel

[3]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

*Corresponding author: E-mail: itaymay@tauex.tau.ac.il.

Associate editor: Li Liu

Abstract

Statistical criteria have long been the standard for selecting the best model for phylogenetic reconstruction and downstream statistical inference. Although model selection is regarded as a fundamental step in phylogenetics, existing methods for this task consume computational resources for long processing time, they are not always feasible, and sometimes depend on preliminary assumptions which do not hold for sequence data. Moreover, although these methods are dedicated to revealing the processes that underlie the sequence data, they do not always produce the most accurate trees. Notably, phylogeny reconstruction consists of two related tasks, topology reconstruction and branch-length estimation. It was previously shown that in many cases the most complex model, GTR+I+G, leads to topologies that are as accurate as using existing model selection criteria, but overestimates branch lengths. Here, we present ModelTeller, a computational methodology for phylogenetic model selection, devised within the machine-learning framework, optimized to predict the most accurate nucleotide substitution model for branch-length estimation. We demonstrate that ModelTeller leads to more accurate branch-length inference than current model selection criteria on data sets simulated under realistic processes. ModelTeller relies on a readily implemented machine-learning model and thus the prediction according to features extracted from the sequence data results in a substantial decrease in running time compared with existing strategies. By harnessing the machine-learning framework, we distinguish between features that mostly contribute to branch-length optimization, concerning the extent of sequence divergence, and features that are related to estimates of the model parameters that are important for the selection made by current criteria.

Key words: **model selection, phylogenetic reconstruction, simulations, nucleotide substitution models, machine learning, Random Forest for regression.**

Introduction

The abundance of substitution models (Jukes and Cantor 1969; Kimura 1980; Felsenstein 1981; Cowan 1984; Hasegawa et al. 1985; Tamura 1992; Tamura and Nei 1993; Schöniger and Von Haeseler 1994; Zharkikh 1994; Huelsenbeck and Crandall 1997) and the need to choose one (or few) has established model selection as a prerequisite for phylogeny reconstruction (Goldman 1993a; Huelsenbeck and Rannala 1997; Sullivan and Swofford 1997, 2001; Posada and Crandall 2001; Pupko et al. 2002). This is evident by the wide use of model selection as an inherent component of phylogenetic analysis. For example, the most widely used method, MODELTEST (Posada and Crandall 1998), was included in the 100 all-time top-cited papers by Web of Science (Van Noorden et al. 2014). However, although criteria for phylogenetic model selection have been adapted from the general statistical literature, they rely on assumptions that do not hold for phylogenetic data analysis (Posada and Buckley 2004). For example, the likelihood ratio test (LRT) for comparing between a pair of nested models, as approximated using the chi-square distribution, has been expanded to the comparison of multiple models via the hierarchical likelihood ratio test (hLRT) criterion, which performs a sequence of LRTs between pairs of nested substitution models, until a model that cannot be rejected is reached. However, LRTs assume that at least one of the compared models is adequate and might be incorrect when the models are misspecified (Foutz and Srivastava 1977; Kent 1982; Golden 1995). Obviously, no substitution model can fully capture the genuine complexity of the evolutionary process, such that even the most adequate one merely provides an approximation of reality (Box 1976), therefore posing model misspecifications that may bias the results of LRTs in phylogenetics (Zhang 1999). Furthermore, it has been shown that the choices determined by the hLRT are influenced by the order of pairwise tests and the significance threshold used to reject the simpler model in each paired comparison (Yang et al. 1995; Ripplinger and Sullivan 2008). Other information criteria compute the maximum-likelihood scores for all the candidate models

doi:10.1093/molbev/msaa154 Advance Access publication June 25, 2020

simultaneously. As the maximum-likelihood score generally increases with the inclusion of more parameters, the Akaike and Bayesian Information Criteria, AIC (Akaike 1973, 1974) and BIC (Schwarz 1978), assign different penalties according to the number of parameters included in the model while also considering the data size. AIC assumes that the sample size is large enough to maintain the asymptotic property of the likelihood function and thus penalizes the likelihood score only for the number of model parameters. This assumption, however, rarely holds for phylogenetic data. In contrast, the corrected AIC (AICc; Sugiura 1978; Hurvich and Tsai 1989) and BIC criteria penalize the likelihood score by the number of parameters as well as the data size. Unfortunately, data size is not well defined for phylogenetic data because—owing to shared evolutionary history as well as functional and structural constraints—aligned sequences are surely dependent, as are the sites along the alignment. Although previous studies advocated to use the alignment length as the sample size (Posada and Crandall 2001; Minin et al. 2003; Posada 2008; Ripplinger and Sullivan 2008; Darriba et al. 2012), much debate exists regarding the effective data size in phylogenetics (Churchill et al. 1992; Goldman 1998; Morozov et al. 2000; Posada and Buckley 2004).

It was previously reported that model selection adds little to the performance of specific inferential tasks. For example, we and others have demonstrated that employing the most parameter-rich model, GTR+I+G, leads to reconstruction of topologies that are as accurate as those produced by the models selected by established model selection strategies (Arbiza et al. 2011; Abadi et al. 2019; Spielman 2020), meaning that performing model selection only increases computation time but does not add to the topological accuracy. On the other hand, for the task of branch-length estimation, a parameter-rich model is not guaranteed to produce more accurate estimates than a simpler model. For example, using extensive simulations, Abadi et al. (2019) demonstrated that the consistent use of GTR+I+G generally results in inferior branch-length estimates compared with those obtained by model selection criteria, with the BIC and DT criteria obtaining the most accurate estimates. Buckley et al. (2001) demonstrated on empirical data that the use of site-specific rate (SSR) model, which accounts for even more parameters than the GTR+I+G model since multiple rate parameters are assigned according to pre-determined site categories, was better fitted to the data but resulted in reduced accuracy compared with the GTR+I+G model, probably due to the false assumption of equal rates within site categories which led to underestimation of branch lengths. These studies and others (Posada 2001; Minin et al. 2003; Abdo et al. 2005) show that simpler models may yield better branch-length estimates and therefore model selection should still be advocated for this task. Notwithstanding, the current practices for model selection focus at revealing the model that best fits the data in terms of likelihood scores, but may fail to identify parameters that have a greater impact on phylogenetic accuracy (Buckley and Cunningham 2002). Selecting a model that is not suitable, with respect to branch-length estimation, may result in large deviations from reasonable estimates. These, in turn, may lead to substantial biases in downstream analyses, such as divergence time estimation (Buckley et al. 2001; Sanderson and Doyle 2001). For instance, Tao et al. (2020) demonstrated on several nucleotides and amino acid data sets that even though divergence time estimates derived from simple and complex models were overall similar, they led to considerable discrepancies for several nodes, in some cases deviations of 130 My. Thus, a method that is optimized to select the best model for phylogeny reconstruction is in need (Sanderson and Kim 2000; Kelchner and Thomas 2007). Here, we employ the emerging machine-learning computational framework to tackle this task.

A supervised machine-learning technique consists of training an algorithm over descriptive data sets to learn the effect of the data characteristics on the outcome. In practice, a data sample is represented by a set of explanatory features. According to these features, an algorithm is trained to predict the most appropriate label in the case of classification, or a continuous target value in the case of regression. For example, decision trees are constructed such that every node splits the training data according to a threshold value of a certain feature (Rokach and Maimon 2008). Essentially, the features along paths of the tree repeatedly separate the data to smaller subsets such that the samples within a subset at a tip own similar feature values beneficial to characterizing it with a certain label. A successful reconstruction of a decision tree is represented by a categorization of instances of the same label with high accuracy, and indicates the usage of powerful features. Numerous algorithms for classification or regression exist, for example, K-means, Naïve Bayes, support vector machine, and K-nearest neighbors (Kotsiantis 2007). High performance of a method for a certain data set and learning task does not guarantee that it would perform well for another, thus the suitability of alternative algorithms should be examined (Caruana and Niculescu-Mizil 2006). In any case, supplying the algorithm with a training set that extends over the full range of realistic instances is crucial for ultimate modeling of possible occurrences.

New Approach

To date, few studies have introduced emerging machine-learning techniques to phylogenetics, for example, for the task of topology reconstruction (Suvorov et al. 2020) and for detecting autocorrelation of evolutionary rates among ancestral and descendant lineages (Tao et al. 2019). Here, we present ModelTeller, which is based on the Random Forest learning algorithm, for the prediction of the optimal model for branch-length estimation. The Random Forest algorithm consists of construction of many decision trees according to subsamples of the training data. During the construction of each decision tree, the features that separate the samples most accurately at each split of the tree are selected. Thus, representative data samples and extraction of meaningful features are inherent for good performance. ModelTeller was trained over data sets that were simulated using parameter estimates that were drawn from an extensive cohort of empirical phylogenetic data sets, thus representing realistic data characteristics. Given a multiple sequence

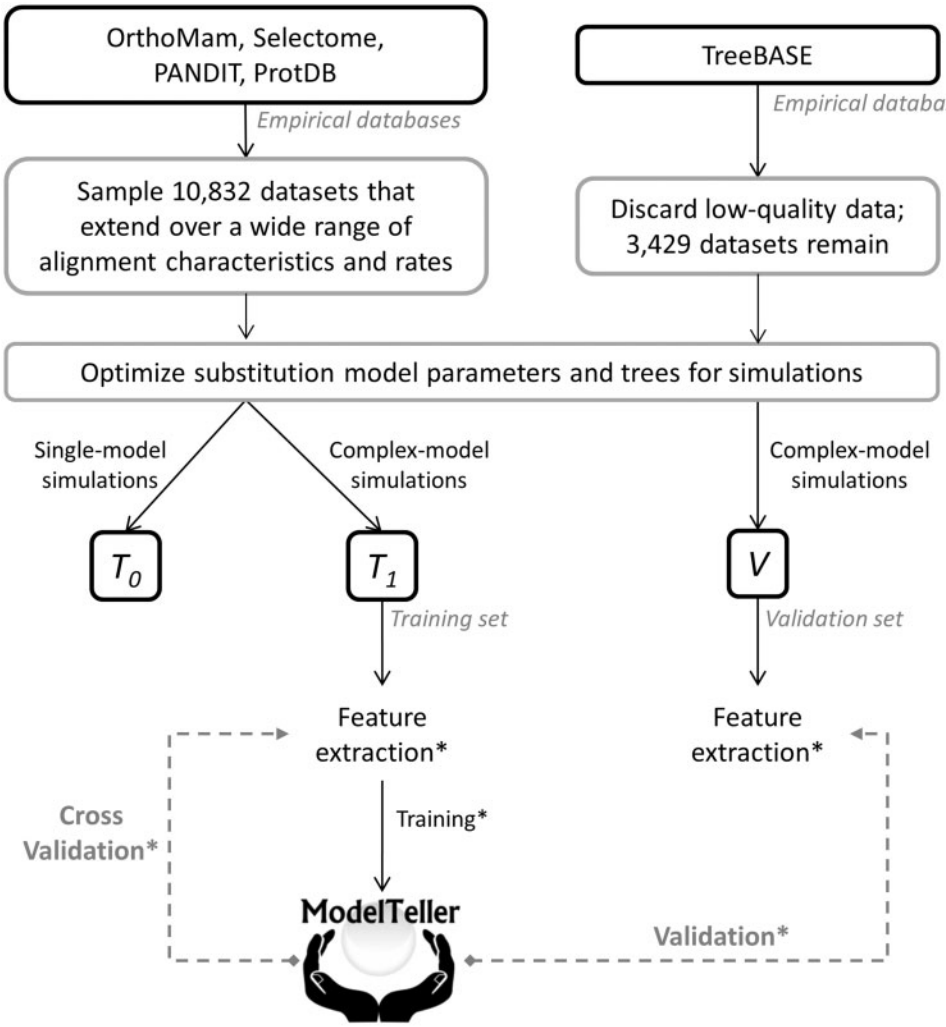

OrthoMam, Selectome, PANDIT, ProtDB
Empirical databases
Sample 10,832 datasets that extend over a wide range of alignment characteristics and rates
TreeBASE
Empirical database
Discard low-quality data; 3,429 datasets remain
Optimize substitution model parameters and trees for simulations
Single-model simulations
Complex-model simulations
Complex-model simulations
T_0
T_1
V
Training set
Validation set
Feature extraction*
Feature extraction*
Cross Validation*
Training*
ModelTeller
Validation*

FIG. 1. A schematic flowchart of the data and procedures applied in this study. The black solid arrows represent the computational pipeline and the dashed gray arrows represent a prediction scheme, either cross-validation or validation procedures. Given an empirical sequence database, we first computed the optimized parameters for each of the 24 nucleotide substitution models. Given these parameters, simulations were made to generate the training/validation data and informative features were extracted from the simulated alignments. The features of the training set were used to train ModelTeller and assess its performance in cross-validation. The features extracted from the validation set were used for assessing the predictions of ModelTeller. The asterisk represents procedures that were done once for the analysis of ModelTeller and once for ModelTeller$_G$: for ModelTeller, the features were computed over a quick reconstruction of a BioNJ tree; for ModelTeller$_G$, the GTR+I+G tree was first optimized, and the features were computed over this tree. The training and prediction were made according to the relevant sets of features.

alignment (MSA) of the examined nucleotide sequences, ModelTeller extracts features that characterize the sequence data and the assumed phylogeny. Some of these features are accounted for by current practices for model selection and some were previously suggested as test statistics for examining whether a model is adequate to the data (for a full description of the features, see supplementary table S1, Supplementary Material online). These features could be assigned to three main groups. The first group consists of features that were computed directly from the MSA. These represent the nucleotide composition and measures concerning the similarity between the sequences. For the second group of features, we computed several tree features based on a rapid reconstruction of a BioNJ tree (Gascuel 1997) together with an approximation of the +I and +G model parameters. The third set includes features of the MSA that were computed for a subset of sequences, excluding those of a distant group. This subset was defined by the larger cluster in the partition induced by the longest branch in the BioNJ tree. Altogether, 54 features were extracted to train ModelTeller.

For conducting this study, three data sets were assembled (fig. 1). The T_0 and T_1 sets were assembled by simulating sequence data according to 10,832 empirical alignments that extend over a wide range of alignment characteristics and rates. These were obtained from four databases that differ in the evolutionary relationships between the sequences: OrthoMam (Ranwez et al. 2007; Douzery et al. 2014), Selectome (Moretti et al. 2014), PANDIT (Whelan et al. 2003), and ProtDB (Carroll et al. 2007). The T_0 set, which was used for analysis of current methodologies for model selection, was simulated according to a single-model simulation procedure. That is, a single model was randomly selected for each empirical alignment (out of 24 commonly used nucleotide substitution models, see Materials and Methods), the parameters of that model were estimated for the data set, and in turn were used to simulate an alignment. The T_1 set, which was used to train ModelTeller, mimics empirical data more realistically by using a complex simulation scheme. To generate this set, multiple models were estimated across each empirical alignment along with rate-variation across sites that resemble empirical data more closely than using the gamma distribution. The performance of ModelTeller was examined on the T_1 training set in a cross-validation procedure. Namely, the data were divided to ten subsets, such that in each iteration the training data consisted of nine subsets and the remaining tenth subset was used for predictions. In this procedure, the training data may not differ much from the test set because they both originate from the same repository of data characteristics. Therefore, even though ModelTeller performed well in cross-validation, its performance was verified on data samples that were not used to train it. To this end, we assembled a validation set, V, that represents relevant empirical cases. The data sets that were employed to estimate the simulations parameters in V were derived from 3,429 phylogenetic data sets deposited in the heterogeneous TreeBASE public repository (Piel et al. 2009; Vos et al. 2012).

The focus of this study is optimal phylogeny reconstruction. First, we use the T_0 set to show that current methodologies are aimed at uncovering the underlying model, but do not produce the most accurate phylogenies. Then, we analyze the performance of ModelTeller both in a cross-validation procedure over the T_1 set and by validation over the V set. We also demonstrate that the models that are best for branch-length estimation generally yield inferior topologies compared with the most complex model, GTR+I+G. Therefore, we suggest a two-step procedure for optimal phylogeny reconstruction, ModelTeller$_G$, in which topologies are first inferred by optimizing a maximum-likelihood phylogeny using the GTR+I+G model and then the best model for branch-length estimation is predicted given the GTR+I+G topology. Finally, we pinpoint the features that contribute most to the prediction of models for branch-length estimation and compare them with those that affect the selection of the best-fitted models, in order to conceive where the gap between the two lies.

Results

Existing Methodologies for Phylogeny Reconstruction

We first analyzed the accuracy of phylogeny reconstruction by existing methods for model selection. To this end, AIC and BIC were executed to select the best model for each simulated data set in T_0. Although we expected that successfully selecting the true model out of 24 nested models would be a difficult task, BIC and AIC recovered the models that generated the data in 75% and 69% of the data sets, respectively. However, in many phylogenetic analyses, the generating model can be considered as a nuisance parameter, whereas the accuracy of the resulting phylogeny is of utmost importance. To examine the accuracy of phylogenetic reconstruction, we reconstructed the maximum-likelihood tree according to each of the 24 models, and measured which one is more similar to the true tree using the Branch-Score (BS) distance, that is, the sum of the squared differences between corresponding branches of the trees (Kuhner and Felsenstein 1994). Note that this measure also incorporates topological dissimilarities because branches that are not found in one tree are treated as having a length of 0. To compare across the different sizes of trees throughout the data, we computed the branch-length error (termed hereafter BLE), that is, the square-rooted BS divided by the sizes of the corresponding true trees. Our results demonstrated that in 84% and 87% of the data sets, alternative models led to estimates that are better than the ones inferred by the models selected by AIC and BIC, respectively. On average, AIC and BIC selected models that were ranked 8.44 and 8.24, respectively (in a ranking of 1–24 from lowest to highest BLE), meaning that in many cases the selected models were not even in the top five. Taken together, these results indicate that although AIC and BIC select the generating models with high frequencies, these selections often yield suboptimal branch-length estimates. Indeed, there were nonsignificant differences in the BLE obtained by the generating model and the models selected by AIC and BIC (average BLE of 0.104 for all three predictions; $P > 0.15$ in all three comparisons; paired t-tests). Moreover, in 92% of the data sets, the generating models resulted in less accurate estimates than those of alternative models, with an average rank of 8.27, again indicating that the generating models were not optimal for branch-length estimation. However, when employing the model that yields the minimal BLE (termed hereafter *the minBLE model*), the average BLE was 0.099 and was significantly better than that of AIC, BIC, and the generating model, meaning that there is room for improvement ($P < 10^{-52}$ in all three comparisons; paired t-tests).

Next, we examined whether the most complex model, GTR+I+G, can be used as a fixed model instead of applying a model selection procedure. To this end, we compared the branch-length estimation and topologies accuracy measurements between three reconstruction strategies: AIC, BIC, and a consistent reconstruction with the GTR+I+G model. To account for topological accuracy, we computed the Robinson–Foulds (RF) distance between the inferred and true trees (i.e., the number of splits that are different between the two trees). To enable comparison across the different tree sizes in the data, we computed the topological error (TE) as the RF distance divided by the number of inner nodes. When the topological accuracy was examined for the T_0 set, the three strategies performed similarly. GTR+I+G was only marginally inferior to AIC and BIC, but these results were statistically nonsignificant. Since these data sets were generated using a single model, thereby assuming the same evolutionary process along all parts of the phylogeny and across all positions, they do not realistically represent the complex patterns observed within empirical data. To this end, we repeated the analysis over the T_1 set that was generated using more

Table 1. Average Rank of the Models Selected by the Various Strategies According to Topological and Branch-Length Accuracy over the T_1 and V Sets.

	T_1 Set		V Set	
	Average BLE Rank	Average TE Rank	Average BLE Rank	Average TE Rank
ModelTeller	8.75	4.47	10.28	6.08
ModelTeller$_G$	8.63[a]	3.79	10.44[a]	4.70
AIC	10.76	3.97	12.58	5.16
BIC	10.55	4.12	12.73	5.25
GTR+I+G	11.10	3.79	12.88	4.70
ModelTest-NG: AIC	12.10	4.58	12.67	5.26
ModelTest-NG: BIC	12.21	4.60	12.82	5.32
AIC$_G$	10.97[a]	3.79	12.60[a]	4.70
BIC$_G$	10.69[a]	3.79	12.82[a]	4.70
minBLE model[b]	1	3.75	1	5.39
maxBLE model[b]	24	5.90	24	6.59
Random model[c]	12.60	4.65	12.56	5.87

NOTE.—The table presents the accuracy measures of the trees reconstructed according to the models selected by the various strategies. The reported averages are across all data sets in T_1 (two left columns) and in V (two right columns): 1) Average BLE rank: the average rank of the selected model within the ranking of the 24 models according to the branch-length errors and 2) Average TE rank: the average rank of the selected model within the ranking of the 24 models according to the topological errors. Different models always lead to different BLEs in our data and therefore always extend from ranks 1 to 24. In contrast, different models might lead to identical topologies. Trees with identical TEs are assigned with the same rank with no skips between increasing ranks.

[a] The ranking of the 24 candidate models was computed where all models were based on a fixed GTR+I+G topology.

[b] The minBLE and maxBLE models are the models that obtained the minimal and maximal BLEs for each data set.

[c] For each data set, a model was randomly selected among the 24 models.

Table 2. Branch-Length Estimation Accuracy across the Different Strategies over the T_1 and V Sets.

	T_1 Set		V Set	
	Average BLE	%Better than Random	Average BLE	%Better than Random
minBLE model	0.141	—	0.330	—
ModelTeller	0.144	68	0.339	63
BIC	0.145	60	0.340	51
AIC	0.146	60	0.340	52
GTR+I+G	0.146	59	0.340	51
Random model	0.147	—	0.341	—
maxBLE model	0.155	—	0.355	—

NOTE.—The table presents the accuracy measures of the trees reconstructed according to the models selected by the various strategies. The reported averages are across all data sets in T_1 (two left columns) and in V (two right columns): 1) Average BLE: the average BLE of the trees inferred by the selected models and 2) % better than random: the percentage of data sets for which the selected model yields lower or equal BLE compared with a model selected by random.

complex simulation patterns and thus mimics empirical data more realistically. When the topological accuracy was examined over the T_1 set, GTR+I+G performed significantly better than both AIC and BIC ($P = 10^{-7}$ and 10^{-15}, respectively, Wilcoxon signed-rank tests). Similar to previous conclusions (Arbiza et al. 2011; Abadi et al. 2019; Spielman 2020), these results indicate that GTR+I+G should be preferred over existing model selection criteria. However, when branch-length estimates were examined, the performance of GTR+I+G was significantly worse than AIC and BIC for both T_0 and T_1 sets ($P < 10^{-8}$ for all comparisons, Wilcoxon signed-rank tests; tables 1 and 2).

Optimal Model for Branch-Length Estimation

We devised ModelTeller, a machine learning-based algorithm optimized to predict the model that yields the most accurate branch-length estimates. To compose the prediction model, we extracted 54 features from each data set of the T_1 set (supplementary table S1, Supplementary Material online) and used these to train a Random Forest for regression algorithm to predict the ranking of the 24 candidate models according to the BLE accuracy measure. ModelTeller was trained and examined in a 10-fold cross-validation. For performance evaluation, the predicted rankings of models across all data sets were compared with the true rankings (from 1 to 24, being the models that yield the minimal and maximal BLE, respectively). The average Spearman correlation coefficient between the true and predicted rankings across all data sets was 0.46, and 24% of the data sets resulted in a coefficient of $r > 0.9$ (fig. 2). On average, the models that were selected (i.e., ranked first) by ModelTeller were ranked 8.75 within the true rankings of models, with a BLE of 0.144, where choosing the best and worst models for every data set yielded average BLEs of 0.141 and 0.155 (tables 1 and 2). ModelTeller performed better than random, as a selection of a random model for each data set resulted in an average rank of 12.6 and an average BLE of 0.147. In comparison, the Spearman correlation coefficients of BIC and AIC were significantly worse than that of ModelTeller, such that the average Spearman correlation coefficients between the true rankings and those computed by BIC and AIC were 0.22 and 0.21, with only 8% of the data

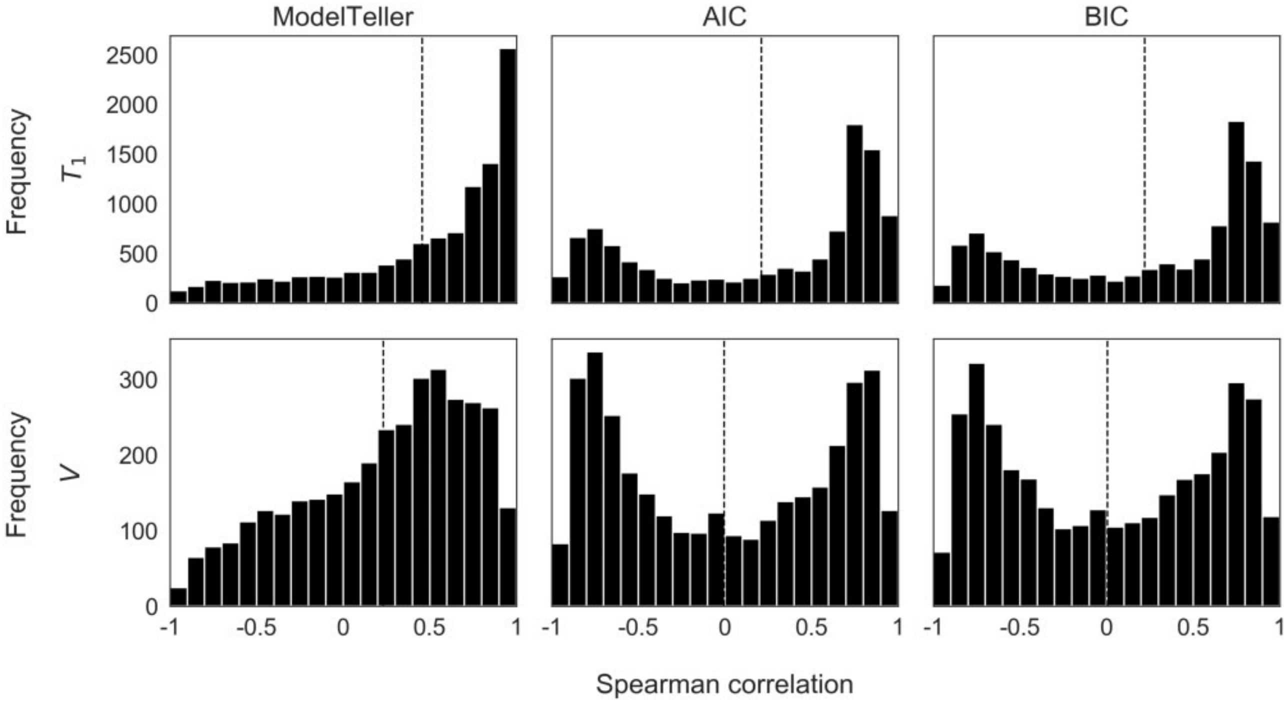


FIG. 2. Performance evaluation of ModelTeller and existing model selection criteria. Distribution of the Spearman correlation coefficients between the observed and predicted rankings of each of the data sets for the T_1 and V sets (top and bottom three panels, respectively, each column panels represent the correlation coefficient distributions of ModelTeller, AIC, and BIC). The x axis represents the Spearman correlation coefficients binned across the interval $[-1,1]$. The y axis represents the number of data sets for which the correlation between the observed ranking of models and the

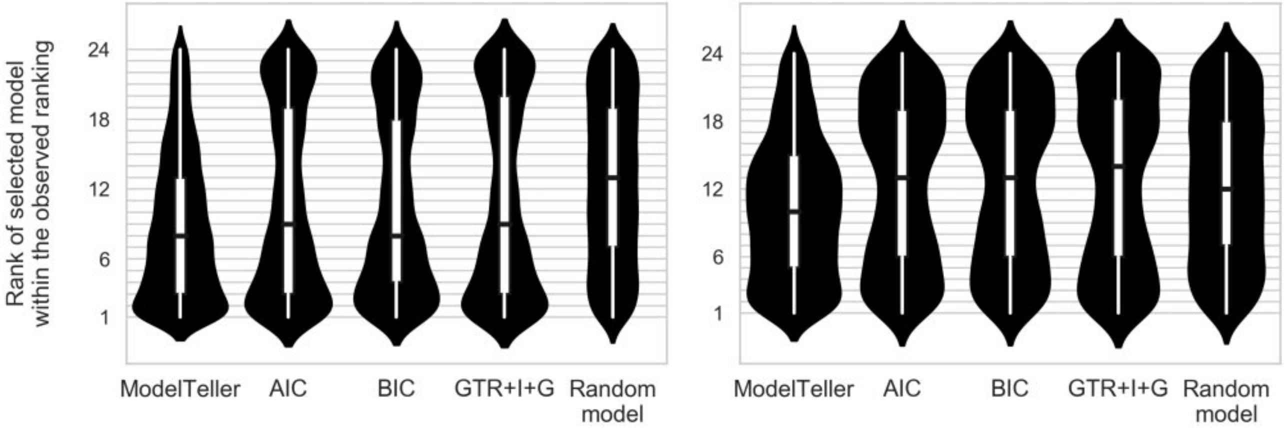


FIG. 3. Distribution of the true ranks of the selected models. The x axis represents five strategies for model selection: ModelTeller, AIC, BIC, consistently using the GTR+I+G model, and selecting a model at random for each data set. The y axis represents the true ranks of the models selected by the methodologies (from 1 to 24 according to the observed BLEs). Left panel: distribution over the T_1 set with medians: 8 for ModelTeller and BIC, 9 for AIC and GTR+I+G, and 13 for a random selection of model. Right panel: distribution over V with medians: 10 for ModelTeller, 13 for AIC, BIC, and for a random selection of model, and 15 for GTR+I+G. The black horizontal lines represent the medians, the thick white bars represent the interquartile (IQR) range, and the thin white lines extend beyond 1.5× the IQR range. The violin plots represent a kernel density estimation of the underlying distribution of the true rank of the selected models; wider sections of the violin plot represent a higher probability that the methodology selects models of that rank, whereas skinnier sections represent a lower probability.

sets with a coefficient $r > 0.9$ (fig. 2; $P = 10^{-291}$ and 10^{-290} for BIC and AIC, respectively, in comparison to ModelTeller, paired t-test). The models selected by BIC and AIC also resulted in significantly higher ranks compared with ModelTeller (BIC = 10.55 and AIC = 10.76 on average, fig. 3, table 1; $P = 10^{-42}$ between ModelTeller and AIC, 10^{-37} between ModelTeller and BIC, and 10^{-7} between BIC and AIC, paired t-tests).

To examine the performance of ModelTeller on data that were not used to train it, we executed it for the validation set, V, that represents relevant empirical cases. The performances of all three strategies, ModelTeller, AIC, and BIC were poorer for the 3,429 alignments in V. The average rank of models selected by ModelTeller was10.28 with an average BLE of 0.339, where the best and worst models yielded average BLEs of 0.33 and 0.355. Here again, ModelTeller performed better than random, as choosing a random model for each data set resulted in an average rank of 12.56 and BLE of 0.341. In contrast, the average ranks of models selected by BIC and AIC were higher than that of ModelTeller and even higher than random (BIC = 12.73 and AIC = 12.58, fig. 3, tables 1 and 2). BIC and AIC selected models that are better than a random selection in 51% and 52% of the data sets, compared with 63% by ModelTeller. The average BLEs of BIC and AIC were both 0.34, which was significantly higher than ModelTeller ($P = 10^{-4}$ for either BIC or AIC, compared with ModelTeller, paired t-tests). Notably, the frequencies at which AIC or BIC selected models that produced the top five trees were symmetric to the frequencies at which they selected models that produced the worst five trees (fig. 3). Exploration of the data showed that AIC and BIC tended to avoid very simple models and particularly tended to select certain models. That is, AIC frequently selected HKY+G, GTR+G, and GTR+I+G and BIC frequently selected K80+G and HKY+G (supplementary fig. S1, Supplementary Material online) and thus generated a distribution pattern that more closely resembles a consistent usage of a single model, like that of the most parameter-rich model, GTR+I+G. Similarly, for both BIC and AIC, the distribution of the Spearman correlation coefficients was bimodal with increasing frequencies toward the boundaries (+1 and −1) and an average ~0 (0.0 for BIC and −0.04 for AIC; fig. 2), implying that correct ranking of models (from best to worst) was obtained at a similar frequency to ranking the models in a completely reverse manner (from worst to best; fig. 3). In comparison, the distribution of the Spearman correlation coefficient of ModelTeller presented a pattern by which correct rankings were more frequent than reverse rankings, with a significantly higher average of $r = 0.23$ ($P = 10^{-59}$ when ModelTeller was compared either with AIC or with BIC, paired t-test).

Running Time

The rapid feature extraction and their assignment in a readily implemented prediction model grants ModelTeller an advantage over existing model selection criteria that are based on the maximum-likelihood computation, which consists of an iterative optimization of the parameter estimates, topologies, and branch lengths for each candidate model. Over the validation set V, the execution of AIC and BIC took 56 min on average as opposed to ModelTeller which took an average of 39 s, a decrease by a factor of 72.

We also compared the performance of the abovementioned strategies to ModelTest-NG (Darriba et al. 2020), a recent reimplementation of jModelTest that was shown to be faster but not less accurate than the original. The models selected by ModelTest-NG were inferior to AIC and BIC as implemented in jModelTest (and therefore also to ModelTeller), both for the T_1 and V sets (table 1). In spite of the reduced accuracy, the computation by ModelTest-NG took, on average, 3.8 min for the validation set V, much faster than jModelTest, but still seven times slower than ModelTeller.

Optimal Model for Branch-Length Estimation on a Fixed Topology

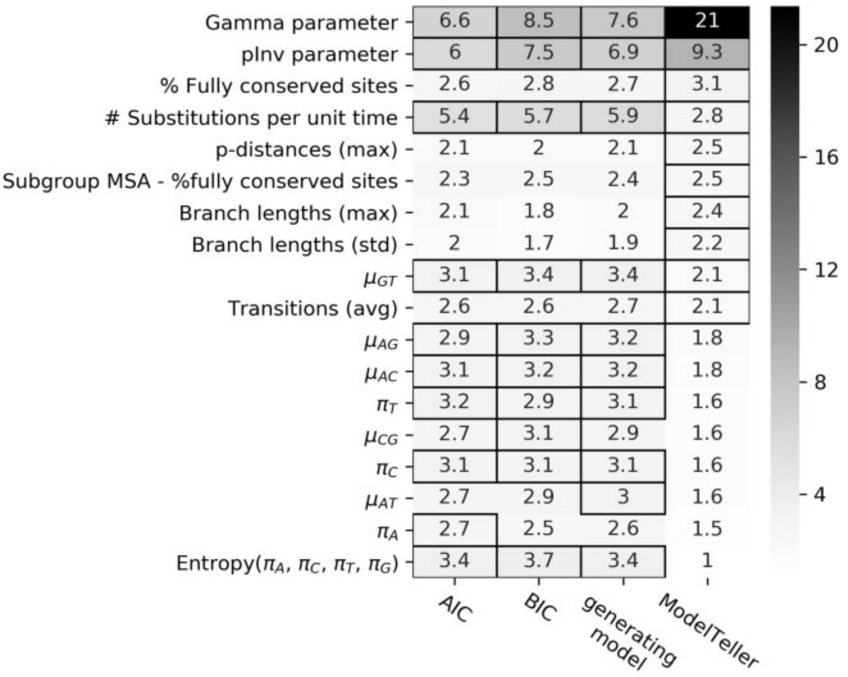
Even though ModelTeller was more accurate in estimating branch lengths compared with AIC and BIC, both in cross-validation over the T_1 set and over the validation set V, it showed lower accuracy when considering only TEs, whereas the highest topological accuracy was obtained by GTR+I+G (table 1). Since ModelTeller is optimized at branch-length accuracy, we suspected that using the best model for branch-length estimation comes at the cost of poorer topology reconstruction. To examine this hypothesis, we measured the rank of the minBLE model according to the TE for each data set. This analysis revealed that the average TE rank of the minBLE model was 5.39 for the V set, significantly higher (i.e., worse) than those obtained for GTR+I+G, AIC, and BIC ($P = 10^{-60}$).

In light of this, we devised a two-step phylogeny reconstruction procedure (referred to as ModelTeller$_G$), such that a topology is first reconstructed using the GTR+I+G model and then the Random Forest model is trained to predict the ranking of the 24 candidate models in terms of branch-length accuracy given the GTR+I+G topology. On the training set T_1, in a 10-fold cross-validation, ModelTeller$_G$ resulted in an average rank of 8.63 according to BLE, similar to that obtained using ModelTeller ($P = 0.15$, paired t-test). On the validation set V, models selected for branch-length estimation by ModelTeller$_G$ were ranked 10.44 according to the BLE. These were nonsignificant when compared with ModelTeller ($P = 0.35$) but significantly better than AIC or BIC ($P = 10^{-5}$ in both comparisons between ModelTeller$_G$ and AIC or BIC, paired t-test). Despite the similarity to ModelTeller in branch-length estimation, these came with a significant improvement in topology reconstruction, owing to the use of the GTR+I+G topologies (table 1).

To examine whether the improvement of ModelTeller over AIC and BIC stems from the better topology reconstruction of the GTR+I+G model, we executed AIC and BIC in a similar two-step manner, that is, we used a maximum likelihood reconstruction of the GTR+I+G model as a fixed topology, optimized the parameters and the tree branch lengths according to each of the candidate models, and used AIC and BIC for model selection among those (termed AIC$_G$ and BIC$_G$). When examined over both the T_1 and V sets, although the GTR+I+G topologies were better than those of AIC and BIC, the average ranks of AIC$_G$ and BIC$_G$ according to the BLE measure were similar to those obtained by a regular execution of AIC and BIC, and worse than those of ModelTeller$_G$ (table 1).

Feature Importance

The underlying algorithm of ModelTeller, Random Forest, consists of selecting the feature that separates the samples at each split of the tree most accurately during the construction of the decision trees. Eventually, the accumulated effect



	AIC	BIC	generating model	ModelTeller
Gamma parameter	6.6	8.5	7.6	21
pInv parameter	6	7.5	6.9	9.3
% Fully conserved sites	2.6	2.8	2.7	3.1
# Substitutions per unit time	5.4	5.7	5.9	2.8
p-distances (max)	2.1	2	2.1	2.5
Subgroup MSA - %fully conserved sites	2.3	2.5	2.4	2.5
Branch lengths (max)	2.1	1.8	2	2.4
Branch lengths (std)	2	1.7	1.9	2.2
μ_{GT}	3.1	3.4	3.4	2.1
Transitions (avg)	2.6	2.6	2.7	2.1
μ_{AG}	2.9	3.3	3.2	1.8
μ_{AC}	3.1	3.2	3.2	1.8
π_T	3.2	2.9	3.1	1.6
μ_{CG}	2.7	3.1	2.9	1.6
π_C	3.1	3.1	3.1	1.6
μ_{AT}	2.7	2.9	3	1.6
π_A	2.7	2.5	2.6	1.5
Entropy(π_A, π_C, π_T, π_G)	3.4	3.7	3.4	1

FIG. 4. Top features of the AIC, BIC, and the generating model classifiers versus ModelTeller. A Random Forest classifier was trained to predict the models selected by AIC and BIC and the generating model. The heat map presents the union of the top ten features for each classifier and ModelTeller. Every column represents the feature importance scores for the respective trained model (given as percentages among all features following removal of the features that represent the substitution model for ModelTeller; see Materials and Methods and supplementary table S1, Supplementary Material online). The cells encircled with gray rectangles represent the features that were among the top ten for each strategy.

of a feature across the forest of decision trees reflects its importance for the prediction accuracy. As shown above, AIC and BIC recover the true model with high frequencies, however, this model does not necessarily produce the optimal branch-length estimates. In order to better understand the data characteristics, represented by the set of examined features, that differentiate between the task of inferring the underlying model and the task of predicting the model that would yield the optimal branch-length estimation, we utilized the training platform of ModelTeller. To this end, we trained three Random Forest classifiers, each with a different target label: one was trained to predict the selection made by AIC, one by BIC, and one to predict the model that was used to generate the sequence alignment (the generating model). These classifiers used the same features that were computed to train ModelTeller. These trainings were carried out on the T_0 set, for which simple generating models exist and AIC and BIC performed well. For the comparison in this analysis, ModelTeller, which relies on a regression model and predicts a ranking of the models, was also trained on T_0. The three classifiers performed well in cross-validation, as the true-positive frequencies (the percentage of data sets for which the target model was successfully predicted) were 58% for AIC, 71% for BIC, and 65% for the generating model. As expected by the high performance of AIC and BIC in identifying the generating model (see the "Existing Methodologies for Phylogeny Reconstruction" subsection), the feature importance scores of the three classifiers were highly similar (Pearson correlation coefficient $r^2 = 0.99$, Spearman correlation coefficient $r > 0.98$, between every pair). However, these differed from those generated by ModelTeller, which is optimized at branch-length estimation (Pearson correlation coefficient $0.51 < r^2 < 0.62$, Spearman correlation coefficient $0.65 < r < 0.69$, between ModelTeller and any other classifier). For example, the alpha parameter of the gamma distribution and the proportion of invariant sites features were ranked first for all analyses, however, the importance values of the former and the latter were 6.6–8.5% and 6–7.5% for the three classifiers compared with importance scores of 21% and 9.3%, respectively, for ModelTeller (fig. 4). The additional top features selected by ModelTeller were related to the tree and sequence divergence, whereas the top features for the classifiers were the model parameters and the substitution matrix entries (fig. 4). The distinct data attributes corroborate the difference between the target tasks of current model selection practices (of identifying the generating model) and that of ModelTeller (optimized at branch-length inference).

Discussion

For decades, phylogenetic model selection has relied on likelihood computations of all alternative models. Many statistical criteria have been proposed over the years, focusing on the inference task of understanding the processes that underlie the data at hand. Nevertheless, these practices have been questioned when the accuracy of the resulting phylogeny is the objective (Sanderson and Kim 2000; Buckley and Cunningham 2002). In recent years, various learning tools have emerged, aimed at learning the patterns within the data for identifying the best course of action for a certain prediction task. In this study, we utilized the machine-learning framework to directly predict the optimal model for the most common prediction task—the reconstruction of the phylogenetic tree for a given set of sequences. Because the tree topology is quite robust to the model choice (Buckley et al. 2001; Posada and Crandall 2001; Sullivan and Swofford 2001; Abdo et al. 2005; Sullivan et al. 2005; Kelchner and Thomas 2007; Ripplinger and Sullivan 2008; Hoff et al. 2016), we have focused here on the accuracy of the estimated branch lengths.

Notably, two major goals in the study of biological systems are inference and prediction. In phylogenetics, statistical inference is an important component to identify the processes that generated the data, but carries multiple drawbacks. First, not all statistical criteria have been adjusted to phylogenomics. For example, AIC and BIC assume independence of the data particles, an assumption that does not hold for sequence data. Second, statistical criteria rely on an extensive search throughout the parameter space and optimization of the likelihood score or its components. These searches entail long computation time. The running time also increases considerably with the size of the data. The abundance of parameters and the accumulating amounts of data often make the computation infeasible, mainly for Bayesian methods. Third, the goal that is achieved by applying these methods is identifying the underlying processes. However, when the ultimate inference depends on the phylogeny, for example, topology reconstruction, divergence time estimation, and ancestral sequence reconstruction, the accuracy of the reconstructed phylogeny is crucial. As shown in our study, existing maximum likelihood criteria perform very well at revealing the underlying process, at least when it is relatively simple, but these models do not necessarily lead to the most accurate phylogenetic tree. In contrast, the learning paradigm makes minimal assumptions on the data generating process and is aimed at prediction. Once a trained model is set up, the prediction process is rapid and the running time relies merely on the extraction of informative features that describe the data.

With the accumulation of massive biological databases and advancements in automating data extraction, many biological frameworks have converted from inference of the underlying processes to generalizing the links between the patterns and the outcome in the observed data for prediction of future samples. For example, the amplification of brain measurements has shifted the research of cognitive neurosciences from isolating the effect of one variable to learning complex brain patterns, thus enabling prediction of future brain scans (Norman et al. 2006; Bzdok 2017). Studying disease profiles based on microarray data has advanced from attempting to detect a single gene that induces a condition in correlation tests to learning the complex interactions between genes and to predicting the condition (Guyon et al. 2002). Notwithstanding, machine-learning poses several limitations that should be addressed when setting up the training data. First, models that were not included in the training of ModelTeller cannot be assessed in the prediction process. Here, we used the nucleotide substitution models that are most commonly used as candidate models in phylogenetic model selection applications (Stamatakis et al. 2005; Posada 2008; Darriba et al. 2012). However, this is only a subset of all possible substitution models, and when new models are proposed the learning machinery should be retrained. Second, machine learning cannot make accurate predictions for patterns that were not found in the training set. To minimize this limitation, we assembled a large cohort of empirical data that extends over a large range of realistic patterns. We also showed that when given the validation data that were generated from patterns drawn from a public repository of contemporary investigated cases, ModelTeller outperformed AIC and BIC. Notably, many machine-learning algorithms exist. In testing, we applied other learning algorithms to our data, however, the Random Forest algorithm performed considerably better. We expect that the main reasons for this stem from the diversity among the features. The study of modeling molecular evolution is broad, and the alignment characteristics that affect phylogeny reconstruction are equivocal. In contrast to other machine learning algorithms, where the addition of an uninformative feature could severely affect the performance, the Random Forest algorithm is known to be robust to a noisy set of features. Moreover, the feature importance analysis that is inherently done during the training of the prediction model could be highly beneficial for identifying the main factors that affect phylogenetic modeling.

The data that were used to perform the initial analysis present low level of complexity and do not account for possible model violations because they were generated according to a single model which is also part of the set of candidate models. For this reason, we opted to train ModelTeller on simulation patterns that were more intricate than a single simplistic model. Both in cross-validation over this training set (T_1) and in validation on the validation set (V), ModelTeller selected better models compared with AIC and BIC and ranked the candidate models more accurately (fig. 3). For the validation set, the rankings of AIC and BIC presented a symmetric pattern whereby successful rankings (with a Spearman correlation coefficient approaching 1) were as frequent as totally inverse ones (with a Spearman correlation coefficient approaching -1; fig. 2). Furthermore, the models that were selected by AIC and BIC were the best or the worst ones at similar frequencies (fig. 3). Examination of the selections showed that AIC and BIC tended to select complex models for these data sets, thereby presenting a pattern that is similar to consistently selecting the most complex model, GTR+I+G, which was the best model and worst model at similar frequencies. These selections perhaps better describe the generating process, which was more complex than any other alternative, but did not lead to more accurate branch-length estimates.

Notably, a common practice in phylogenetics is to partition the sequence data to blocks of alignment sites and use a model selection criterion (e.g., AIC) to fit a model to each one (Lanfear et al. 2017). This procedure requires that the user partitions the sequence data into blocks prior to the application of model selection. This is possible when there is some knowledge on the sequence data, for example, when multiple markers are available or when partitioning a codon alignment according to first, second, and third positions. It should also be beneficial to define other data blocks, such as conserved and nonconserved sites, regulatory elements motifs, or according to the secondary structure of an encoded protein. Given such a partitioned alignment, ModelTeller can be applied to each block separately similar to other model selection criteria. Obviously, biological data often exhibit substantial variation in the underlying substitution process and in the functional constraints that are applied to each site and thus there could exist many unknown data blocks within a single sequence data. In this study, we demonstrated that ModelTeller outperforms other model selection criteria when applied to a single block of data, particularly when it was generated from a heterogeneous mixture of processes.

Ultimately, empirical sequence data would have been the best choice for a training set. Evidently, the true phylogenies of empirical data sets are unknown. An alternative would be to train the machine-learning on data that are simulated according to intricate cases that mimic realistic processes, similar to the simulations conducted in this study. Other realistic simulation scenarios were previously proposed. Philippe et al. (2005) performed simulations based on combinations of within-site rate variation (heterotachy) and variation of evolutionary rates across lineages to show that maximum-likelihood methods are more accurate than maximum-parsimony. Sipos et al. (2011) devised a simulator that uses the Gillespie algorithm to integrate the actions of many concurrent processes such as substitutions, insertions, and deletions rather than assuming a homogeneous substitution model. Nevertheless, no matter which simulation method is employed, it must extend over the large range of possible realistic processes so that the trained model is applicable for many empirical data sets.

Even though ModelTeller was at least as accurate as AIC and BIC, it still did not reveal the optimal model with high frequency. Within a ranking of 1 to 24, the average rank of the models predicted by ModelTeller was 8.75 on the training data and 10.28 on the validation data. Obviously, recording the rank of the selected models does not reflect the actual distances of the reconstructed trees, therefore this measure cannot indicate whether one model leads to substantially different inferences than another (e.g., see supplementary figs. S3 and S4, Supplementary Material online). However, since the tree sizes in the data vary, measuring the ranks of models enables to distinguish between better and worse selections regardless of the tree size. Due to these reasons, we incorporated the ranking in the learning phase, such that the prediction model was trained to predict the ranks of the different candidate models. In a simple classification task, the machine-learning algorithm could have been trained to identify the single most accurate model. Although in trials this procedure predicted the best model for 4% additional data sets (data not shown), when it selected an inaccurate model, this model was of a very low rank. One advantage of predicting the ranking of models is that it not only enabled learning which models are preferred but also which are unfavorable per data set. Furthermore, in case that only a subset of the candidate models is implemented in downstream application software, the user can use the first model that is available among the predicted ranking. In addition, we also accounted for an additional measure, the BLE, that is more sensitive to the magnitude of the error, but is also relatively unbiased. We show that even though the predicted rank is in many cases suboptimal, the magnitude of the errors is not large. Moreover, although the predictions of ModelTeller were significantly better than random, the same was not true for other model selection criteria.

The learning procedure allowed us to measure the contribution of features that have previously been inspected with respect to determining a substitution model for tree reconstruction. To this end, we included the alignment size, which is used in the penalty function of BIC and AICc, and several test statistics that were proposed for testing model adequacy, such as the percentage of fully conserved sites, the distinct site-patterns, and parallel site-patterns suggested by Goldman (1993b), and the multinomial test statistic suggested by Bollback (2002). All of these features resulted in similar contribution to the prediction model in the feature importance analysis (between 1.2% and 3.3%). Notably, the effect of each of these features was obscured by other features. For example, the number of parallel site-patterns, that is, sites that correspond to a similar pattern of evolution regardless of the identity of the nucleotide (e.g., ACCCAA and AGGGAA correspond to the pattern XYYYXX) were highly correlated with the number of distinct site-patterns (in which ACCCAA and AGGGAA are counted as two patterns; Pearson $r^2 = 0.99$), and together, these features obtained an importance of 4.4%. Perhaps surprisingly, the importance of these correlated features was higher than that of the two features that represent the alignment size most intuitively; the importance of the number of sequences was only 1%, whereas that of the alignment length was 1.6%. This suggests that computing the information criteria for phylogenetic model selection might be improved by addressing the data size as the number of independent sites instead of the prevailing approach for addressing the data size as the number of alignment sites. Furthermore, we exploited the machine-learning framework to detect the features that distinguish between the two target inferences: the “true” model that represents the underlying processes and the model for optimal branch-length estimates. Namely, aside for ModelTeller, we trained three classifiers to predict the generating model, the model selected by AIC, and by BIC and compared the features that were most influential for each inference task. This analysis revealed that AIC and BIC, like the generating model, highly rely on attributes related to model parameters whereas the predictions of ModelTeller more strongly depend on the amount of sequence divergence and attributes related to the tree shape.

A benefit of the machine-learning approach for model selection is the possibility to change the target function for any desired inferential task. In this study, we used the BS distance proposed by Kuhner and Felsenstein (1994), a distance between unrooted trees that is most sensitive to the accuracy of branch-length estimates. However, different distances could be employed as well. For example, the RF distance (Robinson and Foulds 1981) could have been employed if topological accuracy had been the focus. This distance counts the number of branch partitions that appear in one tree but not the other, scoring 1 for each unmatched partition. Since it was previously shown that the inferred topology is nearly robust to the chosen model, performing model selection for this goal is not expected to be beneficial (Buckley et al. 2001; Posada and Crandall 2001; Sullivan and Swofford 2001; Abdo et al. 2005; Sullivan et al. 2005; Kelchner and Thomas 2007; Ripplinger and Sullivan 2008; Arbiza et al. 2011; Hoff et al. 2016; Abadi et al. 2019). Kuhner and Yamato (2015) examined the performance of nine tree comparison methods, and concluded that branch length based distances are more suitable for comparison of similar trees (as are the true tree and inferred one) than those based on topological difference only, even if topology is the focus. Among those, they found that a length-based variation of the Robinson–Foulds (RFL) distance (Robinson and Foulds 1979) is the most accurate for rooted trees, which is quite similar to the BS distance we used. Both RFL and BS increment the distance by the difference between corresponding branches, however, the absolute value of this difference is used in the former and the squared value in the latter. Using the BS distance grants higher weights to large differences between branches and lower weights to small differences, but since we compared trees inferred from the same data but with different models, we do not expect this subtle difference to change the results to a large extent. In the abovementioned study, Kuhner and Yamato studied additional metrics but most of them are restricted to clocklike rooted trees, which puts additional constraints on the data the prediction model is applicable to. Other restrictions on the data and the utility could be to analyze only coding or noncoding sequences, thus refining the accuracy of the learning procedure for specific niches. A similar prediction model could be trained over models of codons or proteins or a combination of different models for partitions of the sequence data. For such variations, the current framework of ModelTeller could serve as a foundation, but new data must be generated, relevant features should be extracted, and a new learning model should be trained.

Materials and Methods

Data Assembly

The data used in this study for learning were simulated using characteristics and rates derived from empirical alignments. The alignments were obtained from four databases that differ in the evolutionary relationships between the sequences: 1) OrthoMam (Ranwez et al. 2007; Douzery et al. 2014), a database of orthologous mammalian markers; 2) Selectome (Moretti et al. 2014), which includes codon alignments of species within one of four groups (Euteleostomi, Drosophila, Primates, and Glires; to avoid overlap with OrthoMam, the last two groups were excluded); 3) PANDIT (Whelan et al. 2003), which includes alignments of the nucleotide coding sequences of protein domains; 4) ProtDB (Carroll et al. 2007), which includes genomic sequences that were aligned according to the tertiary structure alignments of the encoded proteins published in BALIBASE (Thompson et al. 2005), SMART (Ponting et al. 1999), OXBench (Raghava et al. 2003), and Prefab (Edgar 2004). From these four databases, alignments were sampled among those with 5–500 taxa and 50–5,000 alignment sites. From OrthoMam, Selectome, and PANDIT, 3,000 alignments were sampled to represent a broad spectrum of data size. Specifically, each database was divided to 20 bins according to the number of sequences, and 150 alignments were randomly sampled from each bin. From ProtDB, 1,270 data sets that comply with this range of data size were included. Together, these data included 10,270 data sets that range in sequence divergence, number of taxa, and alignment length. In a preliminary examination, we observed that a large fraction of the sampled data sets contain sequences of low divergence but only few with high divergence (reflected by the summation of branch lengths in the reconstructed trees). Therefore, we included additional data sets whose sequence divergence is large. To this end, for each data set that was not sampled from the four databases above, we computed the BioNJ tree using PhyML and added those for which the summation of branch lengths was >10, resulting in 563 additional data sets. Altogether, the studied set contained 10,832 data sets.

Single-Model Simulations (T_0 *Set)*

For each data set, we performed simulations according to 1 of 24 models whose rates were derived from the empirical data set. To this end, PhyML was executed for each empirical data set using a randomly selected model to infer the phylogenetic tree (established as the true tree) and the model parameters. The random selection was among a set of 24 substitution models, that is, JC, F81, K2P, HKY, SYM, and GTR—each one with the presence/absence of the +I (proportion of invariant sites) and +G (heterogeneity of rates among sites following the discrete gamma distribution with four categories) options. Given the selected model, the inferred tree, and the model parameters, an alignment was simulated using INDELible (Fletcher and Yang 2009).

Complex-Model Simulations (T_1 *and V Sets)*

To enhance the complexity of the simulations, we simulated additional data sets with heterogeneity of models and rates across the alignment sites rather than simulating according to a single substitution model for the entire alignment. For the first layer of complexity, variation in the substitution pattern across the alignment sites, each empirical data set was divided to partitions of 50 sites (data sets were trimmed such that the alignment length is divisible by 50). jModelTest (Guindon and Gascuel 2003; Darriba et al. 2012) was executed for each partition to obtain the best-fitted model and its inferred free parameters. To obtain the best-fitted model but avoid a bias toward a particular criterion, one maximum-likelihood criterion, that is, AIC or BIC, was randomly selected per partition. The second layer of complexity was obtained by providing the simulator site-specific rates that were drawn from an empirical distribution, fitted to each data set by Rate4site

(Mayrose et al. 2004), rather than being drawn from the invariance + gamma (I+G) distribution. To combine these two layers of complexity, INDELible (Fletcher and Yang 2009) was used to simulate each site given its respective model and rate, and the simulated sites were concatenated to form a single alignment. The input trees in these simulations were reconstructed from the respective empirical alignments using a BioNJ tree (as implemented in PhyML) with the distance matrix computed using the JC model. These trees were regarded as the true trees for the relevant comparisons. For 16 data sets jModelTest or Rate4site failed and therefore they were excluded from further analysis.

Validation Set

As a validation set, we used data different from those used for training ModelTeller. To this end, the data characteristics for the validation set were obtained from TreeBASE (Piel et al. 2009), a repository of user-submitted phylogenies. All available nucleotide alignments were downloaded from the repository. Data sets with fewer than 50 alignment sites, invalid nucleotide characters, multiple identical sequences, sequences that are all gaps or ambiguous characters, or >50% gaps were excluded. The final validation set comprised 3,429 data sets. The data characteristics were extracted from the data sets and used to simulate new alignments using the complex-model simulation procedure.

Tree Inference, Distance Metrics, and Ranking of Models

To determine which model is best for phylogeny reconstruction for each simulated data set, the maximum-likelihood tree of each of the 24 candidate models was reconstructed using PhyML (Guindon et al. 2010), their distances from the true tree were measured, and the model that yielded the minimal distance was considered as best. For branch-length estimation accuracy, the BS distance (Kuhner and Felsenstein 1994) was computed, that is, the sum of the squared differences between corresponding branches of the trees, as implemented in Treedist (Felsenstein 2008). Branches that were not found in one tree due to different topologies were treated as having a length of 0. To compare across data sets of different tree sizes, we computed the branch-length error (BLE), that is, the square-rooted BS divided by the sum of branch lengths of the corresponding true tree. For topological accuracy, the RF distance (Robinson and Foulds 1981) as implemented in TreeCmp (Bogdanowicz et al. 2012) was computed. In accordance with the BLE, the TE was computed as the RF divided by the number of inner nodes in the trees ($n-3$, where n is the number of tree tips). The rankings of models were determined according to the error measures from lowest to highest (ranks 1 to 24). Ties were assigned the same rank with no skips between increasing ranks. Since different models almost conclusively lead to different branch lengths, the BLE rankings resulted in 24 distinct ranks for all data sets in our analysis. In contrast, different models may lead to identical topologies, and so the topologies ranking resulted in fewer distinct ranks. For analysis of a fixed GTR+I+G topology (i.e., ModelTeller$_G$, AIC$_G$, BIC$_G$), the maximum-likelihood phylogeny was first reconstructed in PhyML using the GTR+I+G model, and then the 24 models were optimized along with the branch-length estimates while fixing the GTR+I+G topology.

Machine-Learning Training and Cross-Validation

ModelTeller was implemented as a Regression task where the objective is to learn the ranking of the candidate models according to the BLEs. To incorporate all models within the training phase, each data set, represented by the extracted features, was replicated 24 times corresponding to the 24 models. Each replicate was assigned with additional features that indicate the respective model (see the "Predictive Features" section). The target value for each replicate was the rank of the corresponding model from 1 to 24 (the models that produce minimal and maximal BLE, respectively). Then, the Random Forest for Regression algorithm, implemented in the scikit-learn python module (Pedregosa et al. 2011) was trained over this extended training set of 10,832 X 24 samples using 50 decision trees. The predicted ranks of the test set were determined by applying the trained algorithm to the test samples replicates and ranking the predicted values from low to high (1 to 24, being best to worst; termed "the predicted ranking"). For the cross-validation procedure, in each iteration, one decile of the data sets, including all 24 replicates of each, was reserved as a test set, whereas the others were used for training. In each such iteration, the algorithm was trained over the respective training set and the test set was used for prediction of the ranks. We used two metrics for performance evaluation. First, for each data set, we computed the Spearman correlation coefficient between the true ranking of the models (i.e., according to the error measures) and the predicted ranking (either as produced by ModelTeller or by AIC/BIC) and averaged the coefficients across all. Additionally, we averaged the true rank of the model that was ranked first by the model selection strategy.

To examine if the training set size is sufficient, we examined the performance of growing sizes of the training set on a test set. To this end, we sampled 10% of the data and fixed it as a test set. From the other 90% of the data, we sampled subsets of increasing sizes. The machine learning was trained on each subset and predictions were made for the fixed test set. To avoid biases caused by random sampling, we repeated this procedure ten times such that in each iteration a test set was resampled. The results show that beginning from ~60% of the data, the results were consistent, implying that addition of data samples is not likely to improve the performance of ModelTeller (supplementary fig. S2, Supplementary Material online).

Predictive Features

Each data set was represented by a set of 54 explanatory attributes extracted from it. These features can be classified into three sets of features. The first set are those extracted solely from the sequence alignment. These include the number of sequences, the length of the alignment, the frequency of each type of nucleotide, the alignment entropy, the multinomial statistic, and the sum-of-pairs score. The second set of features represent tree features extracted from a quick approximation of the phylogeny together with the +G and +I parameters approximated from this tree. That is, a basic BioNJ (Gascuel 1997) tree was reconstructed using PhyML (Guindon et al. 2010) and the rate parameters of the GTR+I+G model were optimized on a fixed topology and branch lengths to avoid long computation time (for ModelTeller$_G$, those features were computed over the optimized GTR+I+G phylogeny that was computed as a first step). The values of the estimated parameters including the substitution rates, the shape parameter of the gamma distribution (heterogeneity across sites), and the proportion of invariant sites, as well as the sum of branch lengths of the reconstructed phylogeny were used as features. The third set of features describes the similarity within a subset of the sequences, considering that a distant group may affect the prediction accuracy. To select these sequences, the BioNJ tree was pruned at the longest branch, and the sequences of the larger subtree were used to extract the induced MSA from the original one. Then, some of the MSA features (set 1) were computed over this reduced MSA. For a full description of the features and their extraction procedures, see supplementary table S1, Supplementary Material online.

Implementation and Availability

ModelTeller was implemented in python using the scikit-learn module implemented in python (Pedregosa et al. 2011). The score provided by the algorithm, essentially represents the ranking of the models, such that the first one is predicted to yield the best branch-length estimation. The ModelTeller utility is available for online use and for download at https://ModelTeller.tau.ac.il/ and offers three running modes: 1) predicting a single model for optimization of the model parameters, the branch lengths, and the topology (ModelTeller); 2) predicting a model for branch-length optimization over a fixed GTR+I+G topology (ModelTeller$_G$); and 3) predicting a model for branch-length optimization over a fixed topology given by the user. The online tool provides the predicted ranking of models as well as the output phylogeny for the best model, computed by PhyML (Guindon et al. 2010). If a single model is desired, the first configuration should be used, since it consists of rapid feature extraction and a quickly inferred BioNJ phylogeny and thus its running time is very short. If desired, maximum-likelihood computation of the phylogenetic tree can be followed. When the best phylogeny is desired, we recommend using the second configuration, ModelTeller$_G$. According to our results, this configuration yields more accurate phylogenies in terms of both topology and branch lengths. In this configuration, the feature extraction phase is longer due to the optimization of the GTR+I+G model. However, once the best model is found, its optimization over the precomputed GTR+I+G topology is considerably faster, and thus an accurate phylogeny is obtained without substantial increase in running time. When the true topology is (assumed to be) known, we recommend using the third option. In that case, the computation of ModelTeller does not require tree reconstruction for feature extraction, thus, it is very rapid. Maximum-likelihood computation of the best model and branch lengths over the user-defined topology will be executed if desired.

In order to gain further insights into the features that lead to the prediction of a given sequence data, we also provide the features contribution following the model selection procedure in the webserver. The feature contribution is computed according to values at the inner nodes of the decision trees that can be interpreted as weights that represent the quality of splitting the data by the corresponding features during the training of ModelTeller. To compute the contribution scores for the prediction process, the weights of features along the paths that lead to a certain prediction are averaged across the decision trees in the forest.

Data Availability

The data sets contained within the T_0, T_1, and V sets have been deposited in Open Source Framework (OSF) with the identifier doi:10.17605/osf.io/2ws3a.

Supplementary Material

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

Acknowledgments

S.A. is supported by PhD fellowships provided by the Rothschild Caesarea Foundation and the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University, and travel fellowships provided by Joan and Jaime Constantiner Institute for Molecular Genetics and the Society of Systematic Biology. O.A. is supported by PhD fellowships provided by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University, the Dalia and Eli Hurvits Foundation LTD, and the Ministry of Science, Technology & Space of Israel. T.P. is supported by an Israeli Science Foundation number 802/16. I.M. is supported by an Israeli Science Foundation number 961/17.

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.

Abdo Z, Minin VN, Joyce P, Sullivan J. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Mol Biol Evol*. 22(3):691–703.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. International Symposium on Information Theory. Budapest: Akademiai Kiado. p. 267–281.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 19(6):716–723.

Arbiza L, Patricio M, Dopazo H, Posada D. 2011. Genome-wide heterogeneity of nucleotide substitution model fit. *Genome Biol. Evol.* 3:896–908.

Bogdanowicz D, Giaro K, Wróbel B. 2012. TreeCmp: comparison of trees in polynomial time. *Evol Bioinforma*. 2012:475–487.

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

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

Buckley TR, Cunningham CW. 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Mol Biol Evol*. 19(4):394–405.

Buckley TR, Simon C, Chambers GK. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst Biol*. 50(1):67–86.

Bzdok D. 2017. Classical statistics and statistical learning in imaging neuroscience. *Front Neurosci*. 11:543.

Carroll H, Beckstead W, O'Connor T, Ebbert M, Clement M, Snell Q, McClellan D. 2007. DNA reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics* 23(19):2648–2649.

Caruana R, Niculescu-Mizil A. 2006. An empirical comparison of supervised learning algorithms. In: ACM International Conference Proceeding Series. Vol. 148. New York: ACM Press. p. 161–168.

Churchill GA, von Haeseler A, Navidi WC. 1992. Sample size for a phylogenetic inference. *Mol Biol Evol*. 9(4):753–769.

Cowan J. 1984. Some mathematical questions in biology. *Neurobiol Math Biosci*. 70(2):265–267.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol Biol Evol*. 37(1):291–294.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. JModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 9(8):772.

Douzery EJP, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol Biol Evol*. 31(7):1923–1928.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.

Felsenstein J. 2008. Treedist – distances between trees. Univ. Washingt. [Internet]. Available from: http://evolution.genetics.washington.edu/phylip/doc/treedist.html.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 26(8):1879–1888.

Foutz RV, Srivastava RC. 1977. The performance of the likelihood ratio test when the model is incorrect. *Ann Stat*. 5(6):1183–1194.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14(7):685–695.

Golden RM. 1995. Making correct statistical inferences using a wrong probability model. *J Math Psychol*. 39(1):3–20.

Goldman N. 1993a. Statistical tests of models of DNA substitution. *J Mol Evol*. 36(2):182–198.

Goldman N. 1993b. Simple diagnostic statistical tests of models for DNA substitution. *J Mol Evol*. 37(6):650–661.

Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. *Proc R Soc Lond B*. 265(1407):1779–1786.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52(5):696–704.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Mach Learn*. 46(1/3):389–422.

Hasegawa M, Kishino H, Yano T. A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22(2):160–174.

Hoff M, Orf S, Riehm B, Darriba D, Stamatakis A. 2016. Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics* 17(1):143.

Huelsenbeck JP, Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Ecol Syst*. 28(1):437–466.

Huelsenbeck JP, Rannala B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276(5310):227–232.

Hurvich C, Tsai C. 1989. Regression and time series model selection in small samples. *Biometrika* 76(2):297–307.

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

Kelchner SA, Thomas MA. 2007. Model use in phylogenetics: nine key questions. *Trends Ecol Evol*. 22(2):87–94.

Kent JT. 1982. Robust properties of likelihood ratio tests. *Biometrika* 69(1):19–27.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 16(2):111–120.

Kotsiantis SB. 2007. Supervised machine learning: a review of classification techniques. *Informatica* 31:249–268.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*. 11(3):459–468.

Kuhner MK, Yamato J. 2015. Practical performance of tree comparison metrics. *Syst Biol*. 64(2):205–214.

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.

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.

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst Biol*. 52(5):674–683.

Moretti S, Laurenczy B, Gharib WH, Castella B, Kuzniar A, Schabauer H, Studer RA, Valle M, Salamin N, Stockinger H, et al. 2014. Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res*. 42(D1):D917–D921.

Morozov P, Sitnikova T, Churchill G, Ayala FJ, Rzhetsky A. 2000. A new method for characterizing replacement rate variation in molecular sequences. Application of the Fourier and wavelet models to Drosophila and mammalian proteins. *Genetics* 154(1):381–395.

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*. 10(9):424–430.

Pedregosa FVaroquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. 2011. Scikit-learn: machine learning in {P}ython. *J Mach Learn Res*. 12:2825–2830.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol*. 5(1):50.

Piel WH, Chan L, Dominus MJ, Ruan J, Vos RA, Tannen V. 2009. TreeBASE v. 2: a database of phylogenetic knowledge. London: e-BioSphere.

Ponting CP, Schultz J, Milpetz F, Bork P. 1999. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res*. 27(1):229–232.

Posada D. 2001. The effect of branch length variation on the selection of models of molecular evolution. *J Mol Evol*. 52(5):434–444.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol*. 25(7):1253–1256.

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.

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

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Syst Biol*. 50(4):580–601.

Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M. 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol Biol Evol*. 19(12):2294–2307.

Raghava GPS, Searle SM, Audley PC, Barber JD, Barton GJ. 2003. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* 4(1):47.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol*. 7(1):241.

Rokach L, Maimon Z. 2008. Data mining with decision trees. Vol. 69. Singapore: World Scientific Publishing Co. Pte. Ltd.

Ripplinger J, Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? *Syst Biol*. 57(1):76–85.

Robinson DF, Foulds LR. 1979. Comparison of weighted labelled trees. In: Combinatorial mathematics VI. Lecture notes in mathematics. Heidelberg (Berlin): Springer. p. 119–126.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci*. 53(1–2):131–147.

Sanderson MJ, Doyle JA. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from rbcL and 18S rDNA data. *Am J Bot*. 88(8):1499–1516.

Sanderson MJ, Kim J. 2000. Parametric phylogenetics? *Syst Biol*. 49(4):817–829.

Schöniger M, Von Haeseler A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol Phylogenet Evol*. 3(3):240–247.

Schwarz G. 1978. Estimating the dimension of a model. *Ann Statist*. 6(2):461–464.

Sipos B, Massingham T, Jordan GE, Goldman N. 2011. PhyloSim – Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics* 12(1):104.

Spielman SJ. 2020. Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics. *Mol Biol Evol*. 37(7):2110–2123.

Stamatakis A, Ott M, Ludwig T. 2005. RAxML-OMP: an efficient program for phylogenetic inference on SMPs. *Parallel Comput Technol*. 3606:288–302.

Sugiura N. 1978. Further analysis of the data by Akaike's Information Criterion and the finite corrections. *Commun Stat Theory Methods*. 7(1):13–26.

Sullivan J, Abdo Z, Joyce P, Swofford DL. 2005. Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. *Mol Biol Evol*. 22(6):1386–1392.

Sullivan J, Swofford DL. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J Mamm Evol*. 4(2):77–86.

Sullivan J, Swofford DL. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst Biol*. 50(5):723–729.

Suvorov A, Hochuli J, Schrider DR. 2020. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Syst Biol*. 69(2):221–233.

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 9:678–687.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 10(3):512–526.

Tao Q, Barba-Montoya J, Huuki LA, Durnan MK, Kumar S. 2020. Relative efficiencies of simple and complex substitution models in estimating divergence times in phylogenomics. *Mol Biol Evol*. 37(6):1819–1831.

Tao Q, Tamura K, Battistuzzi FU, Kumar S. 2019. A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. *Mol Biol Evol*. 36(4):811–824.

Thompson JD, Koehl P, Ripp R, Poch O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* 61(1):127–136.

Van Noorden R, Maher B, Nuzzo R. 2014. The top 100 papers. *Nature* 514(7524):550–553.

Vos RA, Balhoff JP, Caravas JA, Holder MT, Lapp H, Maddison WP, Midford PE, Priyam A, Sukumaran J, Xia X, et al. 2012. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Syst Biol*. 61(4):675–689.

Whelan S, de Bakker PIW, Goldman N. 2003. Pandit: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics* 19(12):1556–1563.

Yang Z, Goldman N, Friday A. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst Biol*. 44(3):384–399.

Zhang J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol Biol Evol*. 16(6):868–875.

Zharkikh A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol*. 39(3):315–329.