Syst. Biol. 72(3):544–558, 2023

https://doi.org/10.1093/sysbio/syac046
Advance Access Publication 06 July 2022

Assessment of Absolute Substitution Model Fit Accommodating Time-Reversible and Non-Time-Reversible Evolutionary Processes

VADIM GOREMYKIN

Research and Innovation Centre, Fondazione Edmund Mach, Via Mach 1, 38098 San Michele all'Adige (TN), Italy
Correspondence to be sent to: Research and Innovation Centre, Fondazione Edmund Mach, Via Mach 1, 38098 San Michele all'Adige (TN), Italy; E-mail: vadim.goremykin@fmach.it.

Received 14 June 2021; reviews returned 16 June 2022; accepted 24 June 2022
Associate Editor: Josef Uyeda

Abstract.—The loss of information accompanying assessment of absolute fit of substitution models to phylogenetic data negatively affects the discriminatory power of previous methods and can make them insensitive to lineage-specific changes in the substitution process. As an alternative, I propose evaluating absolute fit of substitution models based on a novel statistic which describes the observed data without information loss and which is unlikely to become zero-inflated with increasing numbers of taxa. This method can accommodate gaps and is sensitive to lineage-specific shifts in the substitution process. In simulation experiments, it exhibits greater discriminatory power than previous methods. The method can be implemented in both Bayesian and Maximum Likelihood phylogenetic analyses, and used to screen any set of models. Recently, it has been suggested that model selection may be an unnecessary step in phylogenetic inference. However, results presented here emphasize the importance of model fit assessment for reliable phylogenetic inference. [Absolute model fit; model misspecification; origin of plastids; phylogenomics.]

A key requirement for avoiding errors in phylogenetic reconstruction is that the substitution model chosen for inference should approximate the data-generating process. This requirement stems from methodological considerations, as it has been shown that Bayesian inference and maximum likelihood (ML) methods are consistent estimators of true tree under the correct model (Chang 1996; Steel 2013; RoyChoudhury et al. 2015; Truszkowski and Goldman 2016). In other words, given infinite site sampling, the tree that generated the data has the highest likelihood/posterior probability when the data-generating substitution model is used for inference. This cannot be generally expected under a wrong model (e.g., Felsenstein 2004).

Although it is impossible to sample an infinite number of sites, a number of simulation studies have identified model misspecification as a source of error in phylogeny reconstruction (Bruno and Halpern 1999; Lemmon and Moriarty 2004; Kolaczkowski and Thornton 2008; Wang et al. 2008; Nguyen et al. 2012; Chen et al. 2019). Consequently, model selection has become an essential part of the phylogenetic reconstruction protocol. The importance of the search for the best-fitting model has recently been challenged (Abadi et al. 2019) on the basis of similarity of trees built under the best-fitting and misspecified models. The validity of the conclusions presented by Abadi et al. (2019) was checked in this study.

Model selection typically involves ranking models according to the information theoretic criteria, of which AIC (Akaike 1974) and BIC (Schwarz 1978) are the best known. These criteria compare model likelihood scores penalized by model dimensionality to avoid selection of overparameterized models. Despite their wide use and popularity, the above methods are limited to ML models with fixed numbers of parameters (Wang et al. 2018). Recently, Crotty and Holland (2022) found that these criteria are also not suitable to discriminate among complex ML finite mixture models and partitioned models and suggested to focus on developing alternative approaches to model selection.

Tests of absolute model fit are free from above limitations. They learn model properties from a large set of simulated data ("training data set") and compare how these properties deviate from those of the observed data. In principle, absolute indices of fit assessing overall fit of a model can be employed to compare all substitution models and to identify the most appropriate substitution model available. The main challenge in development of such indices is coming up with the metric that captures the information relevant to phylogenetic reconstruction.

Given infinite site sampling, the site pattern probabilities expected under the true tree and model are equal to observed pattern frequencies (Yang 2006). This provides a theoretical criterion for construction of indices of overall absolute fit in phylogenetics. A proper index of overall absolute model fit should be constructed in such a way as to yield the best possible estimate of model fit only when the set of site patterns in the observed alignment is predicted by a model. Such an index would assess how far a model deviates from the conditions under which the true tree can be expected to have the highest likelihood/posterior probability if the number of sites is large enough. Compliance with the above criterion becomes increasingly important in the modern phylogenomic era, with the growing use of large alignments in phylogenetic studies. The previous methods do not meet this criterion. This is because they use the statistics describing the observed alignment that are not sufficient to reproduce it.

The relative merits of absolute fit indices still remain poorly documented or simply are unknown because they are very rarely used in phylogenetics. The presented

study discusses the data properties used by these tests and compares their practical performance. Evaluation of a novel method, termed T statistic, that meets the above theoretical criterion, has been included in the study.

Existing test statistics can be divided into those that use multinomial likelihoods (Goldman 1993; Bollback 2002), site pattern binning (Lewis et al. 2014; Chen et al. 2019), pairwise site pattern frequencies (Goremykin 2019), marginal character state frequencies (Foster 2004) and Bowker's test *P*-values (Dutheil and Boussau 2008).

Multinomial likelihood-based tests implicitly assume that the best fit is associated with correct prediction of the observed site pattern counts. However, because multinomial likelihoods depend only on the counts of site patterns and not the patterns themselves, extremely different alignments can have the same multinomial likelihood (Lewis et al. 2014). This potential problem only increases in severity with increasing numbers of taxa and associated exponential growth in the number of possible site patterns. For alignments of finite length, there can be scarcity of site patterns in alignment compared to the number of site patterns the underlying substitution process can potentially generate. This could lead to a situation in which an overwhelming majority of site patterns in observed and simulated alignments would be rare and would have counts equal to one. Under such a scenario, multinomial likelihood-based statistics cannot distinguish if such patterns are accurately predicted or not. Also, in this case, the constant and near-constant sites will predominate in model predictions. This can be expected to decrease the sensitivity of these tests to lineage-specific changes in substitution process.

Tests that rely on site pattern binning also result in a loss of information necessary to detect departures from time-reversibility of the substitution process in different lineages. Such departures can be visualized by Bowker's test of symmetry (Bowker 1948; reviewed in Jermiin et al. 2017 and Naser-Khdour et al. 2019) that checks the null hypothesis of equality of occurrences of the forward and reverse substitutions in a pairwise comparison of sequences (see Supplementary Appendix S1 available on Dryad at http://dx.doi.org/10.5061/dryad.4f4qrfjc8). Considering the above null hypothesis here helps to highlight deviations from time-reversible assumptions concerning forward and reverse substitutions among sequences. This is necessary because lineage-specific deviations from time-reversibility of the substitutions process, common in biological data, have been shown to lead to errors in phylogeny reconstruction (Ho and Jermiin 2004; Jermiin et al. 2004; Gruber et al. 2007; Blanquart and Lartillot 2008; Duchêne et al. 2017). A comprehensive test of data-model fit should be sensitive to such phenomena.

The statistic utilized in the tests based on composition-dependent binning of site patterns (counts of sites in a binned category) is not sensitive to direction of change between character states. For example, the replacement of all A character states by T character states and all T character states by A character states in a site pattern assigned to the A + T category would not affect the size of the category and, thus, would not affect the results of model fit assessment. The same would be true for the exact frequency-based bins comprising, for example 50% A + 50% T character states and thus, for any downstream clustering algorithm which reduces the number of such bins by merging some of them together. These and similar types of replacements change compositional bias among sequences, but they cannot be registered by any test which is based on site pattern binning.

The test for overall substitution model fit described in Goremykin (2019), section "Estimation of Substitution Model Fit," has the same drawback. The test utilizes the counts for the combinations of aligned character states (A + A, A + C, etc.) calculated in a pairwise alignment that is built by concatenating all possible pairwise alignments of sequences of dissimilar taxa in each multiple sequence alignment (MSA). The observed and predicted counts are compared using the Gelfand and Ghosh (1998) statistic. Because the ratio of forward to reverse substitutions does not affect the calculation of these counts, the test cannot visualize an ability of a model to account for lineage-heterogeneous substitution process.

By contrast, the tree and model-based composition-fit test (Foster 2004) answers a specific question whether a model can account for departures from stationarity of substitution process across the lineages (Foster 2004; Duchêne et al. 2017; Jermiin et al. 2020a). The test utilizes the χ^2 statistic to compare the counts of character states in observed MSA sequences with corresponding model-based predictions. The test cannot discriminate among alignments with any substitution rate changes. The reason for considering the test here is to illustrate a case in which a statistic that is known to be limited a priori to only certain aspects of models is used for their ranking in terms of fit.

It should be noted that the statistics for model fit based on Bowker's test *P*-values (Dutheil and Boussau 2008) are poorly suited for detecting the similarity between observed and predicted sequence alignments (Supplementary Appendix S1 available on Dryad). This is because pairwise alignments with different marginal base composition and different counts of individual substitutions can have the same Bowker's test *P*-value. Therefore, these statistics cannot serve as indices of overall model fit (Supplementary Appendix S1 available on Dryad).

In order to address the shortcomings of the previous methods, a novel T statistic is suggested here. The underlying principle of this novel test statistic is that a model should be evaluated based on how much it deviates from the observed data and not from any other alignment of the same size.

The T statistic checks how exactly a model can reproduce the structure of the pairwise alignment subsets characterized by different overall substitution rates. The statistic is sensitive to modeling of site-specific substitution rates and disparity among forward and reverse substitutions, which allows to visualize ability of a model to accommodate departures from

time-reversibility of the substitution process in different lineages. The proposed statistic is also sensitive to phylogenetic information contained in pairwise aligned sequences. Saturation due to multiple substitutions per site results in a loss of phylogenetic information and makes model prediction of the pairwise site patterns less exact. Taking that into account, the proposed statistic prioritizes modeling accuracy of the observed substitutions that are not affected by a degradation of phylogenetic signal due to saturation.

The novel test shows superior predictive power on unseen data in comparison to all other methods of absolute model-data fit assessment tested. Because the test statistic does not become zero-inflated with increase in number of taxa, it is suitable for assessment of fit of models to multitaxon data sets. The statistic can be used to compare lineage-heterogeneous and lineage-homogeneous substitution models used in ML and Bayesian analyses. Another important advantage of the T statistic is that it can also be applied to gapped alignments, which are used in a great majority of phylogenetic studies today.

Materials and Methods

The Novel T Statistic for Assessment of Substitution Model Fit

For each MSA included in comparisons of model-data fit, estimates of relative substitution rate in each site are conducted employing the method of Pesole and Saccone (2001):

$$R_s = \sum_{i=1}^{i=0.5z(z-1)} \frac{\delta_{is}}{D_i} \quad (1)$$

wherein D_i is a genetic distance calculated in the i^{th} pairwise comparison of sequences from a MSA with z taxa and $\delta_{is}=1$ if a substitution is observed in the above sequence comparison at a MSA site s or $\delta_{is}=0$ otherwise. Because the observed alignment used in this study contained heterogeneously evolved taxa, D_i was calculated employing the Tamura-Nei distance for heterogeneous substitution patterns (Tamura and Kumar 2002). For each pairwise sequence comparison, the distances are computed excluding sites with gaps and ambiguous characters.

The R_s values are used to partition MSA sites according to their relative substitution rate. In working out the test, I considered that rare pairwise patterns are likely to be affected by stochastic variation (i.e., noise). The number of such patterns increases with the number of partitions. If the test extracts noise in the training data, it negatively affects its discriminatory power on unseen data. By contrast, pairwise substitution patterns comprising numerous substitutions are more likely to represent the underlying data structure. The partitioning scheme, outlined below, helps to avoid rare pairwise substitution patterns in partitions and was chosen because it resulted in high discriminatory power of the test. If rare pairwise substitution patterns are unavoidable, the proposed method has another level of protection against the influence of noise on model fit estimates, as explained below (formula 10).

A sum of the R_s values (S) is calculated over all MSA sites, and the sites are sorted in ascending order of the R_s values. Starting from the first position in the sorted site set, the sites are consecutively added to the first MSA partition and their R_s values are added to each other. When the resulting sum exceeds 0, the sites are assigned to the second partition and when the sum exceeds 0.5*S, all the remaining sites are assigned to the third partition. This operation results in the variable MSA partitions with equal number of substitutions. The operation corresponds to the Step 1 in the flowchart presenting the main steps of the proposed method (Fig. 1).

In order to evaluate over all pairwise comparisons of taxa, names of taxa in MSA are sorted in alphanumeric order and each taxon is assigned a rank in the sorted list. A square matrix is formed with entries in rows and columns represent ranks of taxa in ascending order (Step 2 in Fig. 1). All pairwise sequence comparisons are iteratively performed among taxa represented by their ranks in either upper or lower triangular part of the resulting square matrix. The significance of it is that each sequence comparison in each MSA compared is performed maintaining the same direction of character state substitutions, for example from the taxon assigned Rank 1 to the taxon assigned Rank 2. Choice of upper/lower triangular part of the rank matrix does not affect the results.

For an alignment between sequences i and j, the counts of pairwise site patterns containing alphabet-specific character states (e.g., A–A, A–T, T–A, etc.) are separately computed in three pairwise alignment subsets, each containing sites assigned to the first, second or third MSA partitions (Step 3 in Fig. 1):

$$C_{xy(z)} = \sum_{b=1}^{b=k(z)} 1_{N_{ib}=x} 1_{N_{jb}=y} \text{ if } R_i < R_j \quad (2)$$

wherein x ∈ {*A*,*C*,*G*,*T*}, y ∈ {*A*,*C*,*G*,*T*}, z is a subset of sites assigned to a partition, $k_{(z)}$ is the length of z, N_{ib} and N_{jb} are the character states at site b for sequences i and j, respectively, $\mathbf{1}_v=1$ if v is true and 0 otherwise and R_i and R_j are the ranks for sequences i and j.

The counts are transformed into relative frequencies:

$$f1_{xy(z)} = \frac{C_{xy(z)}}{L} \quad (3)$$

wherein x ∈ {A,C,G,T}, y ∈ {A,C,G,T}, z is a subset of sites, and L is the length of pairwise alignment (PWA) among sequences i and j calculated excluding positions with gaps and ambiguous characters.

The relative frequencies of pairwise site patterns *in each PWA subset z* are also calculated for the observed

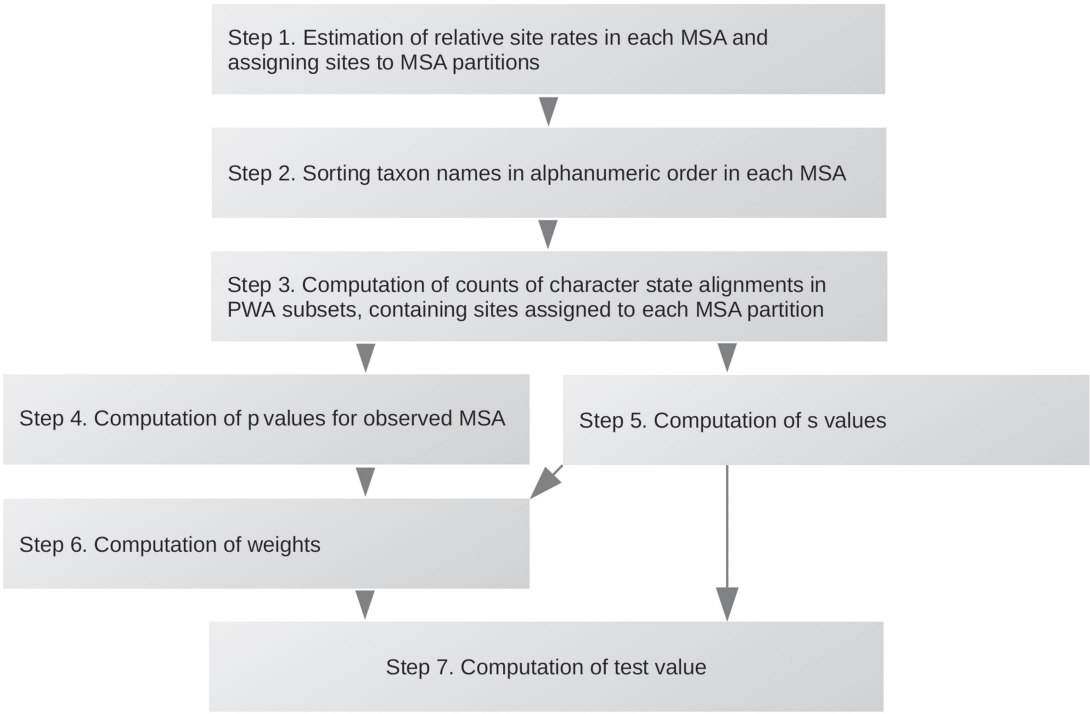

FIGURE 1. Flowchart of the main steps in calculation of the T statistic (formula 8). Step 1 provides sensitivity to modeling of the site rate distribution. Step 2 provides sensitivity to modeling lineage-heterogeneous substitution processes. The counts calculated at Step 3 are used to calculate (i) p values for the observed alignment (Step 4) and (ii) the observed and predicted data points that are compared by the method (s values, Step 5). The method incorporates p values to assign relatively higher weights (Step 6) to the squared Z-scores for the observed substitutions that are not affected by a degradation of phylogenetic signal due to saturation.

data:

$$f2_{xy(z)} = \frac{C_{xy(z)}}{L_{(z)}} \quad (4)$$

wherein $x \in \{A,C,G,T\}$, $y \in \{A,C,G,T\}$, z is a subset of sites, and $L_{(z)}$ is the length of z calculated excluding positions with gaps and ambiguous characters.

The $f2_{xy(z)}$ values are transformed into odds:

$$o_{xy(z)} = \frac{f_{xi} f_{yj}}{f2_{xy(z)}} \text{ if } R_i < R_j \quad (5)$$

wherein $x \in \{A,C,G,T\}$, $y \in \{A,C,G,T\}$, z is a subset of sites, f_{xi} is a frequency of x in a part of sequence i which contains alphabet-specific characters and is included in alignment with alphabet-specific characters in sequence j, f_{yj} is a frequency of y in a part of sequence j which contains alphabet-specific characters and is included in alignment with such characters in sequence i, and R_i and, R_j are the ranks for sequences i and j.

The odds are converted into probabilities (Step 4 in Fig. 1):

$$p_{xy(z)} = \frac{o_{xy(z)}}{1 + o_{xy(z)}} \quad (6)$$

The $p_{xy(z)}$ values (distributed on the Scale 0–1) are used to assign different weights to the quality of the model prediction of the individual observed data points, which is assessed based on squared Z-scores. The method incorporates $p_{xy(z)}$ values to assign relatively higher weights to the squared Z-scores for the observed data points (s values, described below) corresponding to the observed substitutions that are not affected by a degradation of phylogenetic signal due to saturation. A $p_{xy(z)}$ value for a substitution approximating 1 indicate that the effect of saturation is negligible.

The goodness of fit is assessed by comparing the actual and simulated values calculated as follows:

$$s_{xy(z)} = \frac{f_{xi} f_{yj}}{f1_{xy(z)}} \text{ if } R_i < R_j \quad (7)$$

wherein $x \in \{A,C,G,T\}$, $y \in \{A,C,G,T\}$, z is a subset of sites, $f1_{xy(z)}$ is as in formula 3, f_{xi} and f_{yj} are as in formula 5, and R_i and, R_j are the ranks for sequences i and j. The model-predicted $s_{xy(z)}$ values are inferred for each of the 16 possible pairwise site patterns. If the model-predicted $f1_{xy(z)}$ value is equal to zero, and $s_{xy(z)}$ cannot be computed, it is set to zero. These values are excluded from the subsequent calculation of Z-scores. This stage corresponds to Step 5 in Figure 1.

It should be noted that the computed s values can be used to build a 4×4 asymmetric square matrix of odds for pairwise site patterns in a PWA, where each matrix member representing a substitution can be calculated as $x = 1/(1/s_{xy(z2)} + 1/s_{xy(z3)})$ and other members as $x =$

$1/(1/s_{xy(z1)}+1/s_{xy(z2)}+1/s_{xy(z3)})$. The above matrix of odds corresponds uniquely to a PWA-specific set of frequencies of pairwise site patterns (e.g., Chao and Zhang 2008, p. 156). A frequency of a pairwise site pattern is the sum of all the frequencies of site patterns in a MSA that contains it. Following the rationale in Rogers (1997), a set of frequencies of pairwise site patterns, which can be sampled from a MSA therefore, should uniquely correspond to a set of site patterns in the MSA. The validity of this assumption was checked here in preliminary experiments involving random site pattern generation for a 4-taxon MSA. These indicated that a set of frequencies of the pairwise site patterns containing alphabet-specific character states uniquely corresponds to a set of site patterns in a MSA which does not contain (i) ambiguous character states, (ii) columns consisting exclusively of gaps, and (iii) columns with a single alphabet-specific character state. Such site patterns do not contain phylogenetic signal and/or are not encountered in simulated replicates and can be removed from the observed alignment. Considering above, assessment of model fit based on accuracy of prediction of s values can identify a model that is able to generate a set of site patterns in the observed MSA. An advantage of this statistic is that it is sensitive to heterogeneity of the substitution process across sites. Another advantage related to sensitivity to phylogenetic signal which is contained in pairwise aligned sequences is explained in Supplementary Appendix S2 available on Dryad.

Model fit is assessed by computation of a weighted average of squared Z-scores (numbers of standard deviations the realized s value lies from the mean over the corresponding s values calculated from simulated replicates) and taking a square root of the resulting value in order to restore the original scale of measurement (Step 7 in Fig. 1):

$$T=\sqrt{\sum_{n=1}^{n=t} w_n \left(\frac{s_n-\bar{m}_n}{\frac{1}{r}\sum_{i=1}^{i=r}(s_r-\bar{m}_n)^2} \right)^2} \tag{8}$$

In the above formula, t is the total number of s values calculated for the observed alignment, s_n is the n^{th}observed s value, $\bar{m}_n$ is a mean over corresponding replicate-specific (s_r) s values predicted by a model (formula 7) which are not equal to zero, r is the number of non-zero s_r values, and w_n is the weight assigned to each term in the sum. The smaller is the T value, the better is the data-model fit. A numerical example of calculation of T value is provided in Supplementary Appendix S3 available on Dryad.

The weighting scheme in formula 8 was designed such that the test score was influenced more by the squared Z-scores calculated for the observed s values corresponding to unsaturated substitutions. Calculation of weights (Step 6 in Fig. 1) used in formula 8 is performed as follows:

$$w_n=\frac{p_n{}^2 d_n}{\sum_{n=1}^{n=t} p_n{}^2 d_n} \tag{9}$$

wherein t is as in formula 8, p_n is as calculated in formula 6, and d_n is a coefficient which is introduced to correct for a small sample size:

$$d_n=\left(\frac{r}{h}\right)^c \tag{10}$$

In the above formula, h is the number of simulated replicates, r is as in the formula 8 (the number of non-zero s_r values), and c is a positive number. If h equals r, then d_n equals 1 and does not affect the results. However, if the observed pairwise site pattern is rare and the corresponding predicted pairwise site pattern is encountered in just a few replicates, then its impact on the total test result would be minimized. Rare substitutions are more likely to be affected by stochastic variation which does not represent the properties of the data-generating process. The d_n coefficient helps to avoid influence of stochastic variation on model fit estimates. In the preliminary experiments involving different partitioning methods and alphabets and resulting in rare alignments of character states in MSA partitions, introduction of the d_n coefficient was found to improve discriminatory power of the corresponding tests (results not shown). As the rate of improvement flattens out at high c values, the default value of c was set to 100.

Use of weights calculated as in formula 9 but without squaring p_n values was found to improve discriminatory power of the test more than twofold compared to the unweighted test version. Introduction of weights calculated as shown in formula 9 was found to improve the discriminatory power more than 12-fold in simulation experiments compared to the unweighted test version.

The time necessary to perform the test as implemented in test.pl script (available as Supplementary Material available on Dryad) with a data matrix of 86 taxa with 42,141 aligned positions, used as an observed alignment, and 500 simulated replicates of the same size representing model predictions is about 11 hours on a single core of Intel Xeon Gold 6242 Processor.

Preparation of a Gap-Less Alignment

Sequences of chloroplast protein-coding genes common to the plastomes of five Glaucocystophyta and eight Rhodophyta algae were fetched from GenBank. Translated sequences of *Porphyridium purpureum* genes, which were present in all above plastomes, were Blasted against a local database of translated sequences of cyanobacterial genes fetched from 74 cyanobacterial genomes. The cyanobacterial coding gene sequences corresponding to the best-scoring hit for each taxon were aligned with the plastid gene sequences using the MACSE program (Ranwez et al. 2011). The resulting set of codon-based alignments was inspected to discard (i)

alignments without gap-less sites shared by all species, (ii) alignments wherein homology of sequences was dubious due to short lengths of aligned regions and/or high sequence divergence, and (iii) alignments which produced trees under the GTR + G model characterized by a mixture of short and long cyanobacterial branches, which could be attributed to presence of paralogs.

The alignments of the 51 genes which passed the inspection were concatenated to produce a gapped alignment. The list of these genes is provided as Supplementary Material available on Dryad. An array of vertical gap-less blocks in the gapped alignment was sampled by the Gblocks program (Castresana 2000) embedded in the SeaView alignment editor (Gouy et al. 2010). Gblocks was run based on the protein alignment version, selected by toggling protein view mode on in SeaView. The resulting blocks were inspected and manually edited in SeaView. The final selection of blocks of codons was saved to produce a 42,141 positions long gap-less codon-based alignment of 86 taxa (henceforth referred as to the observed alignment, available as Supplementary Material available on Dryad). The alignment contains only DNA alphabet-specific character states.

Generation of Replicates

Some Bayesian and ML models used to generate replicates assumed the same substitution model schemes. In order to distinguish Bayesian and ML models, the italicized names for the ML models are henceforth used throughout the text. Model parameters for generation of replicates under *GTR + I + G*, *GTR + G*, *GTR + I*, *GTR*, *TIM + I + G*, *TIM + G*, *TIM + I*, *TIM*, *TVM + I+G*, *TVM + G*, *TVM + I*, *TVM*, *TRN + I + G*, *TRN + G*, *TRN + I*, *TRN*, *HKY + I + G*, *HKY + G*, *HKY + I*, *HKY*, *F81 + I + G*, *F81 + G*, *F81 + I*, *F81*, *JC + I + G*, *JC + G*, *JC + I*, and *JC* models were obtained from IqTree v. 1.6.12 (Nguyen et al. 2015) unconstrained tree searches performed based on the observed alignment. The rates among sites in *+G* models were modeled via a discrete gamma distribution with four categories. Model parameters and the optimal number of rate categories (8) for a *GTR+R* model with across sites rate heterogeneity modeled via FreeRate model (Soubrier et al. 2012) were determined employing ModelFinder pipeline (Kalyaanamoorthy et al. 2017) implemented in IqTree. The simulations were conducted with Seq-Gen (Rambaut and Grassly 1997) to sample sets of 500 parametric replicates, each 42,141 pos. long, under each model.

Replicates were also sampled under the default PhyloBayes v. 4.1 (Lartillot et al. 2009) parameters (drawing site-specific rates and site-specific frequency profiles from the corresponding conditional posterior distributions and the nucleotide state at the root from the conditional prior distribution) from the last 500 cycles of the PhyloBayes chains run for 3000 cycles based on the observed alignment. All Bayesian models used assumed GTR and f81 rate matrices. The Bayesian site-heterogeneous models (CAT + GTR and CAT + f81) assumed a mixture of equilibrium frequency profiles (CAT, Lartillot and Philippe 2004) and mixtures of distinct GTR-based rate matrices and equilibrium frequencies over alignment sites (QMM, Wang et al. 2008). Rates among sites for the Bayesian models were modeled via (i) a discrete gamma distribution with four categories (+G), (ii) a continuous gamma distribution (+Gc), and (iii) a Dirichlet process (+D). Heterotachy was modeled via (i) Mixture of Branch Lengths model (MBL, Kolaczkowski and Thornton 2008; Zhou et al. 2007) (+MBL), (ii) Tuffley and Steel's covarion model (Tuffley and Steel 1998) (+TS), and (iii) a modified version of the above model (–covext option in PhyloBayes) (+TSm). Combinations of some of the above components in Markov chain Monte Carlo mixtures resulted in errors in replicate sampling. The final selection of 36 PhyloBayes models (Table 2) used here was determined by the ability to sample replicates.

The default PhyloBayes parameters were applied to sample replicates from the last 500 cycles of a nhPhyloBayes (Blanquart and Lartillot 2008) chain run under the CAT-BP model based on the same data. At the moment of writing, 1314 cycles were sampled under the model, which took 200 days. A fixed value for the number of components in the mixture of equilibrium frequency profiles (350) was used for the run as previously determined under CAT + f81 + G model. Distinct sets of 500 replicates per chain were also sampled based on the last cycle of all above-mentioned chains under the default PhyloBayes parameters.

Assessment of Discriminatory Power

The discriminatory power of different tests was measured as the percentage of times a correct model showed better fit to 100 replicates simulated under the correct model in comparison to a misspecified model (termed "model separation value" [MS]). For each pairwise model comparison, the first 100 replicates in each 500-member replicate set sampled under the correct model were chosen to represent the "unseen data." Each of these 100 replicates was taken as "observed alignment" in turn, and was compared to (i) the last 400 replicates sampled under the correct model and (ii) the last 400 replicates sampled under a misspecified model to calculate two test values under a given method. The above 400-member replicate sets served in these experiments as "training data" to learn properties of the models. If the correct model showed better fit to each of the 100 replicates under a given method, MS value was at maximum (100%). These experiments show if a test statistic captures the properties of the data-generating process poorly or too closely. In both cases, the corresponding estimator can be expected to fail to predict the correct model type reliably in the unseen data.

In the Bayesian framework, the model parameters change from one chain cycle to another under a fixed substitution model scheme. However, the MS method

employed here can function properly only if the replicates in the unseen data set were sampled under the model parameters used to generate the training data for the correct model. Otherwise, the expectation of the best theoretically possible fit of a correct model to each of 100 replicates cannot be justified. Therefore all MS values for comparisons involving Bayesian models were calculated based on the posterior predictive replicates sampled from the last chain cycles. In this case, the above expectation should hold true and all the errors to identify the correct models could be attributed to the drawbacks of the methods of model fit assessment.

Assessment of Discriminatory Power in the Presence of Missing Data

Indel-Seq-Gen v. 2.1.0 (Strope et al. 2009) was employed to simulate two 42,141 pos. long replicates under $GTR + I + G$ model parameters estimated previously based on the observed alignment. The simulations assumed the default indel model, maximum indel size of 10 positions, zero probability of insertions, 0.1 deletion per substitution ratio for the first replicate and an analogous value set to 0.2 for the second replicate. The first and the second replicates (provided as Supplementary Material available on Dryad) are henceforth referred as to the "mask file 1" and "mask file 2," respectively.

The mask file 1 had 503,603 gaps distributed over 66% of alignment sites. The percentage of gaps per sequence in the file ranged from 23% to 1.4% with an average value of 14%. The mask file 2 had 906,686 gaps distributed over 70% of alignment sites. A percentage of gaps per sequence in the file ranged from 38% to 2.7% with an average of 25%.

In the first series of the experiments with missing data each character state in each taxon sequence in each replicate encountered at the position where a gap was inserted in the corresponding sequence in the mask file 1 was treated as a gap in calculation of the relevant test statistics. The second series of these experiments was conducted analogously using the mask file 2.

Alternative Statistics

Bollback's test (2002) involves calculation of the multinomial likelihood statistic for the observed alignment and each replicate generated under a model and calculation of a right-tailed *P*-value, representing a frequency of the test statistics calculated from replicates which are larger or equal to the realized test statistic. An absolute value of a Z-score (a number of standard deviations the realized test statistic lies from the mean over analogous values for replicates) was also tried here as an alternative way to summarize the results of the test. The corresponding test is henceforth referred to as "multinomial-Z test."

The tree- and model-based composition-fit test (Foster 2004), henceforth referred to as to "TMCF test," utilizes a Chi-square statistic, $X^2 = \sum[(observed - expected)^2/expected]$, which is individually calculated for the observed alignment and for each replicate. The *observed* values for the observed alignment and for any replicate represent corresponding counts of taxon-specific character states in these multiple sequence alignments. In experiments conducted here, each *expected* value was calculated as a taxon-specific mean value of counts of each character state in the distribution of replicates. Such empirical estimation of the *expected* values can be applied for all the models used in the present study (personal communication from the author). The significance of the test result is indicated by a *P*-value representing the frequency of encountering χ^2 values calculated from replicates, which are larger or equal to the realized test statistic (Foster 2004). An alternative Z-score-based statistic described above was also employed here to summarize test's results. The corresponding test is henceforth referred to as "TMCF-Z test".

The test based on binning of site patterns into 15 categories with distinct combinations of the character states (A, C, G, T, AC, AG, AT, CG, CT, GT, ACG, ACT, AGT, CGT, and ACGT) proposed in Lewis et al. (2014) was conducted forcefully setting the expression $(0 \times \log 0)$ to 0 in calculation of Gelfand and Ghosh statistics, employed to compare observed and predicted categories, considering the limit of the expression. The implementation of the Gelfand and Ghosh method in Lewis et al. does not address missing data explicitly. Following the suggestion made by the authors, analyses with gapped alignments were performed employing fractional distribution of each gapped site patterns to compatible bins. The binning scheme used here (Supplementary Appendix S4 available on Dryad) assumes that each missing character in a gapped site pattern can represent any of the four nucleotides with equal probability.

Recovery of the True Trees from Replicates

Fifty gap-less replicates (termed "initial replicates") were sampled from 3000th cycles of the 50 PhyloBayes chains run based on the observed alignment under 50 distinct full topological constraints (provided as Supplementary Material available on Dryad) and a QMM + D model. Unconstrained tree searches based on each replicate were run under QMM + D and CAT + GTR + D models in PhyloBayes and under $GTR + R$, $GTR + I + G$, $GTR + I$, and *JC* models in IqTree. Bayesian trees were built from the last 500 cycles of the unconstrained chains run for 3000 cycles. Resulting tree topologies were compared to the true trees. The fit of QMM + D, CAT + GTR + D, $GTR + R$, $GTR + I + G$, $GTR + I$, and *JC* models to the above 50 initial replicates was assessed under the proposed test. In these experiments, each model was represented by 500 gap-less replicates generated as described in the section "*Generation of replicates*", but using the "initial replicates" as the observed data. Posterior predictive replicates were sampled from the last 500 chain cycles of the chains.

RESULTS

Comparison of the Discriminatory Power of the Tests

The discriminatory power of the tests was assessed in the comparison of (i) gap-less replicates (comparison series A), (ii) replicates with 503,603 alphabet-specific characters per replicate masked by gaps (comparison series B), and (iii) replicates with 906,686 alphabet-specific characters per replicate masked by gaps (comparison series C). The discriminatory power of Bollback's and multinomial-Z tests was assessed in the comparison of gap-less alignments only. Each model was taken as correct in turn, and was iteratively compared to 65 misspecified candidate models to calculate 65 MS values in $65 \times 100 = 6500$ individual comparisons of model pairs in terms of fit. These comparisons were performed for each of 66 correct models and involved calculation of $66 \times 65 = 4290$ MS values in 429,000 individual comparisons of model fit in total for each test in each comparison series.

Figure 2 provides an overview of discriminatory power of different tests in the comparison series A. When each correct model was iteratively compared to other 65 models, the lowest MS value was recorded for each test (henceforth referred to as "LMS value"). The LMS values in Figure 2 are sorted in ascending order for each test. Each colored line with squares, diamonds, and triangles represents 66 LMS values, obtained for 66 correct models under each test (shown to the right of the graph). Because a vast majority of LMS values (65 out of 66) registered in comparison series A employing Bollback's test were 0%, these results are not shown in the figure.

Table 1 shows the (i) the lowest LMS values (out of 66 calculated for each test), (ii) the total numbers of failures to identify correct models, and (iii) the total percentages of failures to identify correct models registered in comparison series A, B, and C for the tests compared. The results indicate that the discriminatory power of Bollback's test approximates a random draw (evidenced by 50.7% of failures to identify correct models). The discriminatory power of the test assessed in a separate experiment involving the Bayesian models was almost the same (results not shown). Further experiments with the test were not conducted.

In each comparison series, the proposed T statistic showed the lowest number of failures to identify correct models compared to previous methods. Most of these failures (shown in Supplementary Table S1 available on Dryad) were registered in pairwise model comparisons of the Bayesian CAT + f81 models with different mixture components. These experiments suggest that the new test generalizes on unseen data far better than previous methods.

Assessments of the Fit of Substitution Models to the Observed Data

The observed alignment is characterized by compositional heterogeneity among lineages, which is evidenced by a majority (99.73%) of pairwise sequence comparisons failing Bowker's test at 0.05 *P*-value (performed using the SymTest program [Ababneh et al. 2006]). The percentages of such failed tests was also calculated for replicates (one replicate per model). The replicates representing Bayesian models were sampled from the last chain cycles. The percentage of failed tests was 98.17% for the CAT-BP-based replicate. Other values

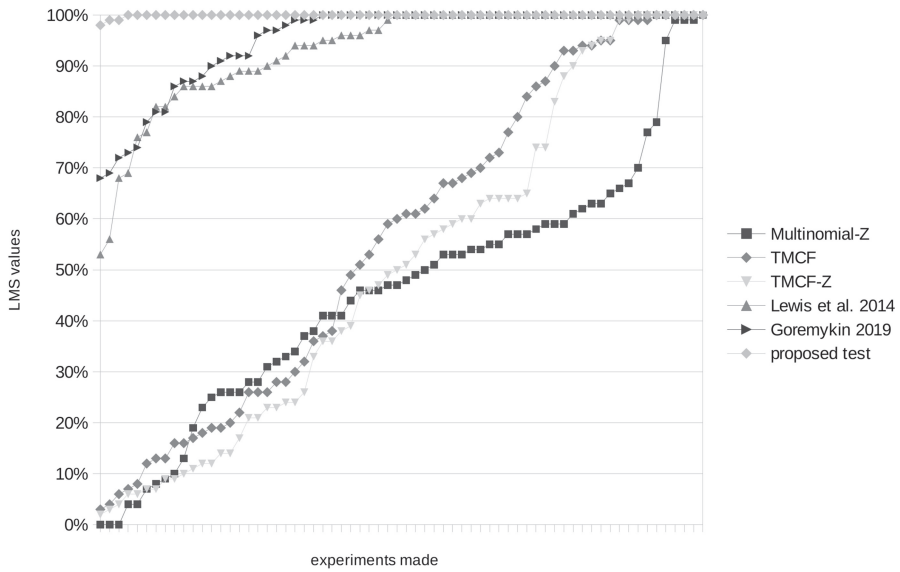

FIGURE 2. Discriminatory power of the tests compared shows that the proposed test has the highest levels of discrimination among models. The graph shows, for each test, 66 LMS values calculated for 66 correct models in the comparison series A, sorted in ascending order. A LMS value is the lowest model separation (MS) value among all values calculated for a correct model. The names of the tests are shown in the legend to the right of the graph. Each colored sign (diamond, square, and triangle) in the graph corresponds to a LMS value, shown in the *Y*-axis, obtained in each experiment aimed at calculating the value, shown in the *X*axis.

TABLE 1. Indices of performance of different tests in comparison series A, B, and C

Tests compared	1[a]	2[b]	3[c]	4[d]	5[e]	6[f]	7[g]	8[h]	9[i]
Bolback (2002)	0	–	–	217,458	–	–	50.7	–	–
Multinomial-Z	0	–	–	23,250	–	–	5.4	–	–
TMCF	3	2	3	34,679	35,241	34,385	8.1	8.2	8
TMCF-Z test	2	2	3	33,969	33,821	34,175	7.9	7.9	8
Lewis et al. (2014)	53	46	47	507	1140	1417	0.12	0.27	0.34
Goremykin (2019)	68	65	65	644	772	860	0.15	0.18	0.2
T statistic	98	98	96	5	7	13	0.0012	0.0016	0.003

Note: The Bollback (2002) and multinomial-Z tests are not suitable for gapped alignments. Dashes indicate missing values for these tests in the corresponding experiments.

[a]The lowest MS value registered in comparison series A.

[b]The lowest MS value registered in comparison series B.

[c]The lowest MS value registered in comparison series C.

[d]The number of failures to identify correct models in comparison series A.

[e]The number of failures to identify correct models in comparison series B.

[f]The number of failures to identify correct models in comparison series C.

[g]The percentage of failures to identify correct models in comparison series A.

[h]The percentage of failures to identify correct models in comparison series B.

[i]The percentage of failures to identify correct models in comparison series C.

ranged from 1.89% to 10.89%. These observations suggest that, with the likely exception of the CAT-BP model, the models compared failed to model lineage-heterogeneous substitutions processes, which are characteristic of the observed data.

Comparison of the absolute fit of the models (each represented by 500 replicates) to the observed alignment was performed using various goodness of fit indices (summarized in Table 2 and shown in Supplementary Table S2 available on Dryad). The results obtained with Bowker's test-based methods are given in Supplementary Appendix S1 available on Dryad. All the replicates for Bayesian models used in these experiments were sampled from different chain cycles, following Bollback (2002). The test proposed here and the TMCF-Z test demonstrated better fit of the CAT-BP model. The TMCF test is not indicated in the presented comparison because all the test values obtained were zero. The TMCF-Z test yielded poor estimates of fit for the majority of site-heterogeneous models (e.g., CAT + GTR + G showed worse fit compared to a *JC* model). By contrast, the proposed test revealed a better fit of all site-heterogeneous models compared to site-homogeneous models.

The multinomial-Z test identified a GTR + D model as providing the best and almost perfect fit ($|Z| = 0.09$) to the observed data (Supplementary Table S2 available on Dryad). The corresponding value for the similar GTR + G model was 54 (Supplementary Table S2 available on Dryad). These variations indicate poor reliability of the test results. The percentage of the site patterns which are shared with the observed alignment was calculated for each replicate. The mean percentage values over each set of replicates generated under each model ranged from 2.88% to 0.003% with an overall mean of 1.84%. Visual inspection of the shared site patterns showed that they were constant or near-constant. This observation reveals that only a small proportion of the information in site patterns is available for the test to assess model-data fit. The above features of the test make it not well suited for assessing overall model fit.

The statistics presented in Lewis et al. (2014) and Goremykin (2019) (section "Estimation of Substitution Model Fit") are insensitive to the ratios of forward to reverse substitutions in different lineages. The results of model ranking obtained with these methods confirm the expectation outlined in the Introduction section that this drawback does not allow detection of the better fit of lineage-heterogeneous models to heterogeneously evolved data.

The results obtained under the proposed test when posterior predictive replicates were sampled across different cycles (i.e., under the conditions that allow changes in model parameters, leading to broader distributions of predicted s values) indicated a better overall fit of the Bayesian models as compared to their ML counterparts. In order to compare model performance under the same conditions (assuming fixed model parameters) for all the models compared, a separate evaluation of fit for the Bayesian models under the proposed test was conducted based on the replicates sampled from the last chain cycles. Lower T values (formula 8) for the Bayesian models were also obtained in these experiments (Supplementary Table S3 available on Dryad).

The rankings of models were compared with mainstream model comparison under BIC (as calculated in IqTree). To quantify the similarity in ranking, each ML model was assigned a number according to its rank in descending order of fit as estimated under BIC (shown in Supplementary Table S2 available on Dryad). In the lists of ML models ranked in terms of fit under other methods model names were substituted by these numbers. The strength of association between the resulting arrays of numbers was assessed by the Spearman's rank correlation. The correlation coefficients calculated in the comparisons of the array corresponding to BIC with those obtained for the proposed test, Lewis et al.

TABLE 2. The ranking of models in terms of absolute fit to the observed alignment obtained with different tests

Proposed test[a]	Values[b]	Multinomial-Z[c]	TMCF-Z[d]	Goremykin 2019[e]	Lewis et al. 2014[f]
CAT-BP	3.778	GTR + D	CAT-BP	CAT + F81 + Gc + MBL	CAT + F81 + G
QMM + D+MBL	9.049	CAT + GTR + D + MBL	QMM + D + MBL	CAT + F81 + D + MBL	CAT + F81 + Gc + MBL
QMM + G	9.560	QMM + D + MBL	CAT + GTR + G + MBL	CAT + F81 + G + MBL	CAT + F81 + G + MBL
QMM + D	9.749	QMM + G	QMM + D	CAT + F81 + Gc	CAT + F81 + Gc
QMM + G+ MBL	9.839	F81 + G + MBL	QMM + G	CAT + F81 + D	CAT + F81 + D + MBL
CAT + GTR + G + MBL	10.523	CAT + GTR + G + MBL	***GTR + R***	CAT + F81 + G	CAT + F81 + D
CAT + GTR + D	11.322	QMM + D	***HKY + I + G***	CAT + F81 + Gc + TSm	CAT + GTR + G
CAT + GTR + D + MBL	11.365	QMM + G + MBL	***TRN + I + G***	CAT + F81 + D + TSm	CAT + GTR + Gc
CAT + GTR + Gc + MBL	11.879	CAT + GTR + Gc + MBL	***GTR + I + G***	CAT + F81 + G + TSm	CAT + F81 + G + Ts
CAT + GTR + G	12.064	CAT + GTR + Gc	***TVM + I + G***	CAT + F81 + G + TS	QMM + G + MBL
CAT + GTR + Gc	12.156	CAT + GTR + D	***F81 + I + G***	***GTR + R***	CAT + GTR + Gc + MBL
CAT + F81 + D	13.303	CAT + F81 + Gc	***TRN + G***	***GTR + I + G***	QMM + D
CAT + F81 + D + MBL	13.345	CAT + GTR + G	***F81 + G***	***TIM + I + G***	CAT + F81 + D + TSm
CAT + F81 + Gc	13.483	CAT + F81 + D	***TIM + I + G***	***TRN + I + G***	CAT + GTR + G + MBL
CAT + F81 + G	13.516	CAT + F81 + G + MBL	***TIM + I***	***TVM + I + G***	CAT + F81 + Gc + TSm
CAT + F81 + G + MBL	13.519	CAT + F81 + D + MBL	***JC + G***	***HKY + I + G***	CAT + GTR + D
CAT + F81 + Gc + MBL	13.555	CAT + F81 + G	***JC + I+G***	CAT + GTR + G + MBL	QMM + D + MBL
CAT + F81 + G + TS	13.769	CAT + F81 + Gc + MBL	***GTR + G***	CAT + GTR + Gc	CAT + F81 + G + TSm
CAT + F81 + D + TSm	13.826	CAT + F81 + D + TSm	***HKY + I***	CAT + GTR + G	QMM + G
CAT + F81 + Gc + TSm	13.925	CAT + F81 + Gc + TSm	***TIM + G***	CAT + GTR + Gc + MBL	CAT + GTR + D + MBL
CAT + F81 + G + TSm	14.228	CAT-BP	***GTR + I***	CAT + GTR + D	CAT-BP
GTR + G + MBL	17.007	CAT + F81 + G + TSm	***GTR***	CAT-BP	GTR + G
GTR + G	17.180	F81 + D + TSm	***HKY + G***	***GTR + G***	***GTR + R***
GTR + Gc	17.330	CAT + F81 + G + TS	***TVM + G***	***GTR + I***	GTR + Gc
GTR + Gc + MBL	17.349	F81 + D	***TVM + I***	***TRN + I***	GTR + Gc + MBL
GTR + D	17.356	F81+ D + MBL	***TRN + I***	***TIM + I***	GTR + D
GTR + R	17.373	GTR + D + MBL	***TRN***	***TIM + G***	GTR + G +MBL
GTR + D + MBL	17.805	GTR + G + MBL	***TVM***	***TRN + G***	GTR + D + MBL
GTR + I + G	18.699	***TVM + I + G***	QMM + G + MBL	CAT + GTR + D + MBL	***GTR + I + G***
TVM + I + G	19.461	***HKY + I + G***	***F81 + I***	***TVM + I***	***TVM + I + G***
TIM + I + G	19.968	***TRN + I + G***	***JC***	***TVM + G***	F81 + G
TRN + I + G	20.093	***GTR + I + G***	***JC + I***	***HKY + I***	***TRN + I + G***
HKY + I + G	20.774	***TIM + I + G***	***HKY***	***F81 + I + G***	***TIM + I + G***
GTR + G	27.113	***F81 + I + G***	***TIM***	GTR + G	***HKY + I + G***
TVM + G	28.475	***JC + I + G***	***F81***	***HKY + G***	F81 + G + TS
TIM + G	28.912	***GTR + I***	F81 + D + MBL	***JC + I + G***	F81 + Gc + MBL
TRN + G	29.069	***HKY + I***	F81 + Gc + TSm	***F81 + G***	F81 + Gc
F81 + G + TS	29.466	***F81 + I***	F81 + Gc	QMM + G + MBL	F81 + D + MBL
HKY + G	30.046	***JC + I***	F81 + G + TSm	QMM + D + MBL	F81 + D
F81 + G + MBL	30.872	***TIM + I***	F81 + G	QMM + D	F81 + G + MBL
F81 + D + MBL	31.159	***TRN + I***	GTR + G	***JC + G***	***GTR + G***
F81 + G + TSm	31.204	***TVM + I***	F81 + G + TS	QMM + G	F81 + Gc + TSm
F81 + D + TSm	31.452	GTR + Gc + MBL	F81 + D	GTR + Gc	***TVM + G***
F81 + Gc + MBL	31.700	GTR + Gc	GTR + G + MBL	GTR +Gc + MBL	F81 + G + TSm
F81 + Gc + TSm	31.833	F81 + G + TS	F81 + Gc + MBL	GTR + D	***JC + G***
F81 + D	31.964	F81 + G + TSm	F81 + D + TSm	F81 + G + TS	***TRN + G***
JC + I + G	31.999	***GTR + R***	GTR + Gc	F81 + Gc + MBL	***F81 + G***
F81 + G	32.166	F81 + Gc + TSm	F81 + G + MBL	F81 + Gc + TSm	***TIM + G***
F81 + I + G	32.905	F81 + Gc + MBL	GTR + D	F81 + G + TSm	***HKY + G***
F81 + Gc	32.943	F81 + Gc	CAT + F81 + D + TSm	F81 + D + TSm	F81 + D + TSm
JC + G	35.646	F81 + G	GTR + Gc + MBL	F81 + D	***F81 + I + G***
F81 + G	36.890	GTR + G	CAT + GTR + D	F81 + Gc	***JC + I + G***
GTR + I	37.181	***JC + G***	CAT + GTR + D + MBL	F81 + G + MBL	***GTR + I***
TIM + I	38.040	***F81 + G***	CAT + F81 + Gc + TSm	F81 + D + MBL	***TVM + I***
TRN + I	38.412	***TIM + G***	CAT + F81 +G + TSm	F81 + G	***TIM + I***
TVM + I	40.329	***TRN + G***	CAT + F81 + G+ TS	***F81 + I***	***TRN + I***
HKY + I	40.775	***TVM + G***	GTR + D + MBL	***JC + I***	***HKY + I***
JC + I	54.663	***HKY + G***	CAT + F81 + D + MBL	***GTR***	***F81 + I***
F81 + I	55.819	***GTR + G***	CAT + F81 + D	***TRN***	***JC + I***
GTR	64.557	***GTR***	CAT + GTR + Gc + MBL	***TIM***	***GTR***
TIM	66.506	***TRN***	CAT + GTR + G	GTR + G + MBL	***TVM***
TRN	67.047	***TIM***	CAT + F81 + Gc	***TVM***	***TIM***
TVM	73.054	***TVM***	CAT + GTR + Gc	***HKY***	***TRN***
HKY	75.140	***HKY***	CAT + F81 + G	GTR + D + MBL	***HKY***
JC	82.453	***F81***	CAT + F81 + G + MBL	***JC***	***F81***
F81	87.798	***JC***	CAT + F81 + Gc + MBL	***F81***	***JC***

Note: Maximum likelihood models are shown in bold font.

[a]Ranking of models under the proposed test.

[b]T values (Formula 8).

[c]Ranking of models under Multinomial-Z test.

[d]Ranking of models under TMCF-Z test.

[e]Ranking of models under the test for substitution model fit proposed in Goremykin 2019.

[f]Ranking of models under the binned test proposed in Lewis et al. (2014).

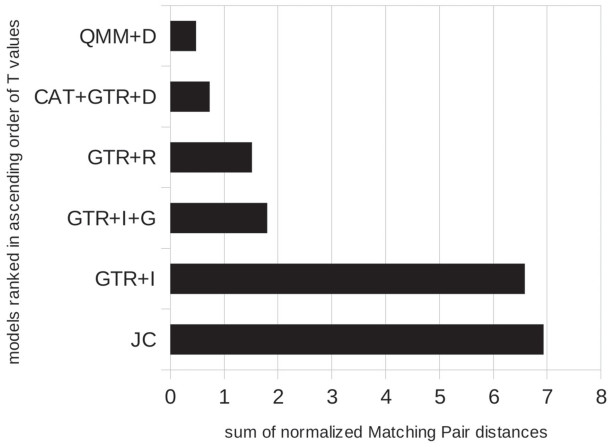

FIGURE 3. Relationship between the T statistic and accuracy of phylogeny reconstruction. The *X*-axis shows sums of normalized Matching Pair distances between the true tree topologies, used as full topological constraints to simulate replicates under a QMM + D model, and the tree topologies recovered under different models from these replicates. The *Y*-axis shows models ranked in ascending order of T values (formula 8) calculated in comparisons of the models to the above replicates.

test, Goremykin's (2019), TMCF-Z, and Multinomial-Z tests were 0.99704, 0.97537, 0.89507, 0.85566, and 0.64236, respectively.

The Effect of Model-Data Fit as Assessed with T Statistic on the Inference of Tree Topology

Models used to estimate trees from 50 replicates simulated under distinct topological constraints and a QMM + D model were ranked in descending order of fit to the corresponding replicates under the proposed test. The following model ranking was obtained in all experiments: QMM + D, CAT + GTR + D, *GTR + R*, *GTR + I + G*, *GTR + I*, and *JC* (Supplementary Table S4 available on Dryad). Similarity between each true tree topology and each topology recovered under above models from the corresponding replicate was quantified under the normalized Matching Pair distance, which is reported to be free from a number of drawbacks characteristic to previously published metrics (Bogdanowicz and Giaro 2017) employing the Visual TreeCmp web applet (Goluch et al. 2020). The distance value ranges from 0 to 1. For the purpose of comparison, all the trees compared here were rooted at *Gloeobacter violaceus*, a basally diverging cyanobacterion. Distances for the unconstrained trees recovered under each model (Supplementary Table S5 available on Dryad) were summed up. The results, shown in Figure 3, suggest that phylogenetic inference can be expected to become more reliable with decrease in T values (formula 8) and highlight the usefulness of absolute model fit assessment for phylogenetic practice.

The tree built under the best-fitting model identified under the proposed test (CAT-BP) is presented in Figure 4. In the tree, the branch supporting the clade of plastids plus *Gloeomargarita lithophora* appears at a basal position within the cyanobacterial radiation. The topology of the tree lends support to the conclusions presented by Ponce-Toledo et al. (2017), which were based on analyses of protein and re-coded protein data.

The tree topology was compared to those recovered under other models from the observed data as described above. The distances between the topology recovered under CAT-BP and the topologies recovered under the models ranked second to eighth in terms of fit (shown in the first column in the Table 2) under the proposed test were about 0.2 (Fig. 5). Analogous distances to the topologies recovered under the 14 worst-fitting models were larger than 0.59.

The branch supporting the sister group relationship between *G. lithophora* and plastids was present in the seven trees built under the site-heterogeneous models ranked second to eighth in terms of fit under the proposed test (Table 2). By contrast, the branch was not recovered under site-homogeneous models. These are the worst-fitting models considered here.

All 29 ML models and 9 site-homogeneous Bayesian models have recovered the branch supporting plastids as sister to *Trichodesmium erythraeum*, a crown-group nitrogen-fixing cyanobacterion. A BioNJ tree was built with FasTMe v.2.1.6.2 (Lefort et al. 2015) under the default options from the CFS compositional distances calculated from the observed alignment using Homo v. 2.0 (Jermiin et al. 2020a). The resulting tree also places plastids sister to *T. erythraeum*. Given that the CFS distance cannot inform about phylogenetic relationships (Jermiin et al. 2020a), this indicates that the above placement of plastids in the phylogenetic trees is likely due to compositional signal.

DISCUSSION

Theoretically, methods assessing how well simulated data resemble the observed can be used to compare any models regardless of their assumptions and dimensions and to identify the best-fitting model. The area of applicability of such methods is limited only by the ability to generate simulated data. This is not an intrinsic limitation of the methodology *per se* and with the development of new simulation tools, it can be expected to become less significant. As the number of sites in molecular phylogenetic data sets continues to grow, assessment of absolute model fit has the potential to become a very useful procedural addition to the phylogenetic protocol. With the increase in the number of sites, site pattern probabilities under the true tree and model approximate observed pattern frequencies (Yang 2006). Therefore, it can be expected that a good agreement between observed and predicted site pattern distributions can become an increasingly important indicator of correctness of phylogeny reconstruction.

Suggestions to include absolute model-data fit assessments into a standard phylogenetic practice have already been made (Jermiin et al. 2020b). However, absolute model fit assessment is still rarely used in phylogenetics and the need for a wider understanding

FIGURE 4. The tree recovered under CAT-BP model. All branches with PP support < 1 are labeled with diamonds. The plastid clade is highlighted in green. The branch leading to *Trichodesmium erythraeum* is highlighted in red.

of the advantages and disadvantages of the assessment methods persists (Jermiin et al. 2020b). It is important to note that similarity of the observed and predicted data features compared by previous methods does not indicate if a model predicts well the set of site patterns characteristic of the observed alignment. For instance, the observed alignment and other, extremely different alignments can have the same multinomial likelihoods, distributions of binned site pattern categories, marginal base compositions of taxon sequences, median Bowker's test values, etc.

To address shortcomings of previously proposed methods, the presented study proposes a novel T statistic. It was designed in such a way as to yield the best possible estimate of model fit only if a model is able to correctly reproduce the set of site patterns in the observed alignment. The statistic does not become zero-inflated with increase in number of taxa, which makes

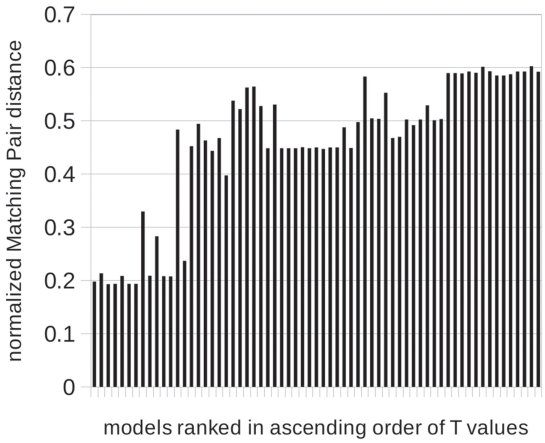

FIGURE 5. Dissimilarity between the tree topologies recovered from the observed alignment under different models. The figure presents normalized Matching Pair distances between the tree topology recovered under a CAT-BP model and phylogenetic tree topologies recovered under other models shown in ascending order of T values (formula 8) calculated in comparisons of these models to the observed data.

it suitable for evaluation of model fit to the multitaxon data sets.

The discriminatory power of the absolute fit indices was assessed here on the basis of their ability to generalize to unseen data. In designing this experiment, it was considered that if a method excessively learns the noise in the training data, it will negatively affect its predictive power on unseen data. It was also considered that poor predictive ability on unseen data can be expected if a method fails to capture important training data properties. The proposed T statistic showed by far the best out-of-sample prediction ability (Fig. 2).

This is a desirable property of an index of overall absolute fit. Such indices assess how far a model deviates from the underlying process that generated the observed data. The ability to identify the data generation process in unseen replicates indicates that the method captures important properties of the process and, thus it estimates how much the model deviates from this process. By contrast, previous statistics which (intentionally or not) describe partial features of the data-generating mechanism can visualize only how strongly a model deviates from the correct description of these features. A model can describe other important features of the data-generating process well or not, but this will not be registered by these statistics. This leads to incorrect model ranking in terms of overall fit, and, by consequence, errors in in identification of the data-generating models.

The main reason to include assessment of model fit in empirical phylogenetic studies is based on the expectation that application of the best-fitting model can help to avoid errors in phylogeny reconstruction. The results obtained here allow a researcher to check the validity of this expectation, which is essential for the further development and use of the methods for assessment of the model-data fit in phylogenetics. This is all more important in light of the recent claim that the expectation is not justified (Abadi et al. 2019). The authors supported their recommendation to abandon the search for the best model by referring to similarity in the topologies of the trees built under the optimal model identified employing jModelTest (Darriba et al. 2012) and other models, including a simple *JC* model. The authors suggested to uniformly use a $GTR + I + G$ model in phylogenetic studies (Abadi et al. 2019).

It should be noted that the choice of models in Abadi et al. was determined by availability of models implemented in jModelTest. These models are all linked to simplified assumptions of molecular evolution such as across-site and across-tree homogeneity of substitution process. More realistic models that relax above assumptions are available. However, their effects on tree inference were not considered in Abadi et al. (2019) due to limitation of the model selection method employed, which is suitable only for models of fixed sizes. These effects can be quite noticeable. If the data are generated by a complex substitution process, which usually the case with biological alignments, use of a simplified models (such as $GTR + I + G$) can lead to errors in phylogeny reconstruction (Fig. 3). Considering absolute model-data fit can help to improve reliability of phylogenetic inference (Fig. 3).

Phylogenetic inference from biological sequence data reported here also contradict the hypothesis that the degree of model fit has no noticeable influence on the results. Here, all ML models implemented in jModelTest and ranging in complexity from *JC* to $GTR + I + G$ uniformly recovered the same phylogenetic artifact—the branch supporting plastids as sister to *T. erythraeum.* The branch was also recovered in the tree built from compositional distances, which indicates a probability of a systematic error in the placement of these lineages in the corresponding phylogenetic trees. The error was confirmed by application of a site- and lineage-heterogeneous model (CAT-BP), which provided the best fit to the observed data under the most powerful test considered here. In the tree recovered under the CAT-BP model (Fig. 4), plastids formed a sister group to the deep-branching cyanobacterium *G. lithophora*. In this tree, the branches leading to plastids and *Trichodesmium* were divided by seven internal branches. The above results indicate that the recommendation to abandon time-consuming model evaluations (Abadi et al. 2019) based on assumed similarity of the trees, which can be obtained with various models can only lead to confusion and unfounded expectations.

There is no reason to restrict assessment of fit to any group of models. Model selection methods that cannot explore available model space can only provide a limited perspective on the efficiency of better-fitting models in recovery of phylogenetic relationships. This drawback of the current model selection methodology can promote misconceptions. The methodological recommendations put forward by Abadi et al. (2019) highlight this point and illustrate the need to change the current approach to model fit assessment. The presented analyses indicate

that examination of the discrepancies between substitution models and data can be a useful procedural addition to a standard phylogenetic protocol. The observations reported here encourage that introduction of the proposed T test into a broad phylogenetic practice can help to avoid errors in phylogeny reconstruction.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.4f4qrfjc8.

References

Ababneh F., Jermiin L.S., Ma C., Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*. 22:1225–1231.

Abadi S., Azouri D., Pupko T., Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun.* 10:1–11.

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

Blanquart S., Lartillot N. 2008. A site-and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol*. 25: 842–858.

Bogdanowicz D., Giaro K. 2017. Comparing phylogenetic trees by matching nodes using the transfer distance between partitions. *J. Comput. Biol.* 24:422–435.

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

Bowker A.H. 1948. A test for symmetry in contingency tables. *J. Am. Stat. Assoc.* 43:572–574.

Bruno W.J., Halpern A.L. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* 16:564–566.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552.

Chao K.M., Zhang L. 2008. Sequence comparison: theory and methods. Vol. 7. London: Springer Science & Business Media.

Chang J.T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* 137:51–73.

Chen W., Kenney T., Bielawski J., Gu H. 2019. Testing adequacy for DNA substitution models. *BMC Bioinformatics.* 20:349.

Crotty S.M., Holland B.R. 2022. Comparing partitioned models to mixture models: do information criteria apply? *Syst. Biol*. (in press) doi: 10.1093/sysbio/syac003.

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

Duchêne D.A., Duchêne S., Ho S.Y. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. *Mol. Biol. Evol.* 34:1529–1534.

Dutheil J., Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* 8:1–12.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Foster P.G. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–495.

Gelfand A.E., Ghosh S. 1998. Model choice: a minimum posterior predictive loss approach. *Biometrika.* 85:1–11.

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

Goluch T., Bogdanowicz D., Giaro K. 2020. Visual TreeCmp: comprehensive comparison of phylogenetic trees on the web. *Methods Ecol. Evol.* 11:494–499.

Goremykin V. 2019. A novel test for absolute fit of evolutionary models provides a means to correctly identify the substitution model and the model tree. *Genome Biol. Evol.* 11:2403–2419.

Gouy M., Guindon S., Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27:221–224.

Gruber K.F., Voss R.S., Jansa S.A. 2007. Base-compositional heterogeneity in the RAG1 locus among didelphid marsupials: implications for phylogenetic inference and the evolution of GC content. *Syst. Biol.* 56:83–96.

Ho S.Y., Jermiin L.S. 2004. Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.* 53:623–637.

Jermiin L.S., Catullo R.A., Holland B.R. 2020b. A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. *NAR Genom. Bioinform.* 2:lqaa041.

Jermiin L.S., Ho S.Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol.* 53:638–643.

Jermiin L.S., Jayaswal V., Ababneh F.M., Robinson J. 2017. Identifying optimal models of evolution. In: Keith J., editor. Bioinformatics: data, sequence analysis, and evolution. Vol. 1., 2nd ed. New York: Humana Press. p. 379–420.

Jermiin L.S., Lovell D.R., Misof B., Foster G.F., Robinson J. 2020a. Detecting and visualising the impact of heterogeneous evolutionary processes on phylogenetic estimates. *bioRxiv*. (in press) doi:10.1101/2020.01.03.894097.

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

Kolaczkowski B., Thornton J.W. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol. Biol. Evol.* 25:1054–1066.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics.* 25:2286–2288.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21:1095–1109.

Lefort V., Desper R., Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32:2798–2800.

Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53:265–277.

Lewis P.O., Xie W., Chen M.H., Fan Y., Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.* 63:309–321.

Naser-Khdour S., Minh B.Q., Zhang W., Stone E.A., Lanfear R. 2019. The prevalence and impact of model violations in phylogenetic analysis. *Genome Biol. Evol.* 11:3341–3352.

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

Nguyen M.A.T., Gesell T., von Haeseler A. 2012. ImOSM: intermittent evolution and robustness of phylogenetic methods. *Mol. Biol. Evol.* 29:663–673.

Pesole G. and Saccone C. 2001. A novel method for estimating substitution rate variation among sites in a large data set of homologous DNA sequences. *Genetics.* 157:859–865.

Ponce-Toledo R.I., Deschamps P., López-García P., Zivanovic Y., Benzerara K., Moreira D. 2017. An early-branching freshwater cyanobacterium at the origin of plastids. *Curr. Biol.* 27: 386–391.

Ranwez V., Harispe S., Delsuc F., Douzery E.J. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One.* 6:e22594.

Rambaut A., Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13:235–8.

Rogers J.S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 46:354–357.

RoyChoudhury A., Willis A., Bunge J. 2015. Consistency of a phylogenetic tree maximum likelihood estimator. *J. Statist. Plann. Inference.* 161:73–80.

Schwarz GE. 1978. Estimating the dimension of a model. Ann Stat. 6(2):461–464.

Soubrier J., Steel M., Lee M.S., Der Sarkissian C., Guindon S., Ho S.Y., Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.* 29:3345–3358.

Steel M. 2013. Consistency of Bayesian inference of resolved phylogenetic trees. *J. Theor. Biol.* 336:246–249.

Strope C.L., Abel K., Scott S.D., Moriyama E.N. 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol. Biol. Evol.* 26:2581–2593.

Tamura K., Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.* 19:1727–1736.

Truszkowski J., Goldman N. 2016. Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. *Syst. Biol.* 65:328–333.

Tuffley C., Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147:63–91.

Wang H-C, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.* 8:331.

Wang H.C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67: 216–235.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Zhou Y., Rodrigue N., Lartillot N., Philippe H. 2007. Evaluation of models handling heterotachy in phylogenetic inference. *BMC Evol. Biol.* 7:206.