Syst. Biol. 68(1):157–167, 2019
© The Author(s) 2018. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syy066
Advance Access publication October 17, 2018

How Well Does Your Phylogenetic Model Fit Your Data?

DAISY A. SHEPHERD[1,*] AND STEFFEN KLAERE[1,2]

[1]*Department of Statistics, The University of Auckland, Auckland, New Zealand;* [2]*School of Biological Sciences, The University of Auckland, Auckland, New Zealand*
[*]*Correspondence to be sent to: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand; E-mail: dshe078@aucklanduni.ac.nz.*

Received 11 December 2016; reviews returned 05 October 2018; accepted 11 October 2018
Associate Editor: Peter Foster

Abstract.—The test for model-to-data fitness is a fundamental principle within the statistical sciences. The purpose of such a test is to assess whether the selected best-fitting model adequately describes the behavior in the data. Despite their broad application across many areas of statistics, goodness of fit tests for phylogenetic models have received much less attention than model selection methods in the last decade. At present a number of approaches have been suggested. However, these are often flawed, with problems ranging from the presence of systematic error in the models themselves to the difficulties presented by the nature of phylogenetic data. Ultimately these problems lead to an inadequate choice of statistic. This is one of the main reasons why goodness of fit assessment is often a neglected step within phylogenetic analysis. We argue not only for the necessity of these goodness of fit measures to test how well the model reflects the data, but additionally for the need for "useful" tests that explain why the model-to-data fit may be inadequate. Such tests are a critical part of the model building process, allowing the model to be adapted to provide a better model-to-data fit or to reject a model class outright due to such an inadequate fit that the intended use of the class may be compromised. Proposed and existing methods in both the maximum likelihood and Bayesian framework will be discussed here, whilst highlighting their strengths and limitations for assessing goodness of fit. The final section discusses some critical open statistical problems in goodness of fit assessment for this field, with the hope of encouraging more research into such a fundamental yet underdeveloped area of phylogenetic inference. [Bayesian phylogenetics; Goodness of fit; maximum likelihood; molecular phylogenetics; outlier detection; residual diagnostics.]

Accurately inferring evolutionary relationships heavily depends on a number of factors. Two of these concern the quality of data employed and the use of an appropriate statistical model. Pivotal to this is ensuring the model not only fits our data, but fits our data well. Poorly fitting models are an inadequate approximation of the true underlying evolutionary processes and can lead to systematic error and unreliable inferences (Kelchner and Thomas 2006).

Assessing goodness of fit is essential when building any statistical model. Fit is defined in two forms—*relative* and *absolute*. The relative goodness of fit reflects the discrepancy between two alternative statistical models, whereas the absolute goodness of fit portrays the discrepancy between the selected model and the data of interest. Both types are fundamentally important to the statistical model-building process.

Tests to assess the relative goodness of fit have been discussed intensely in a phylogenetic context (Sullivan and Joyce 2005; Posada 2001). As a result, there is a strong and impressive body of work regarding the identification of the best model from among a set of given phylogenetic models. Most notably, the software package jModelTest (Posada 2008; Darriba et al. 2012) has been used extensively to select the best model class, with over 9000 citations in the phylogenetic literature. Other, more recent, methods focus on fitting adequate mixtures of phylogenetic models to address particular characteristics of the data, for example, edge rate variation, gene history variation, or site heterogeneity. Examples of such methods are PartitionFinder (to select partition schemes Lanfear et al. 2012) or ModelFinder (novel ways to fit rate heterogeneity Kalyaanamoorthy et al. 2017).

Absolute goodness of fit, on the other hand, did not receive the amount of attention as relative goodness of fit. The purpose of such tests is to provide the possibility to reject the best model due to a lack of fit to the data (Navidi et al. 1991; Reeves 1992; Goldman 1993b; Waddell et al. 2009). Just because our selected model has been deemed the "best" fit, does not necessarily imply that the model adequately describes the behavior in the data, or in the colorful words of Gatesy (2007):

> Given the simplicity of most models, it is possible that model selection in modern systematics is analogous to an overweight man shopping in the petites department of a women's clothing store. A particular garment might fit the portly man best, but this does not imply a good overall fit.

However, the ability to test the adequacy of a model is only half the problem. Given that we have seen a poor fit between model and data, it would be of great use to understand why such a discrepancy has occurred. Or relating back to the example of Gatesy, given we have

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

found the garment does not fit the overweight man well, can we determine where exactly the garment is poorly fitting? In the context of phylogenetic models, this would allow specific insight into reasons behind the lack of fit to data, such as whether the poor fit is due to a nuisance parameter (parameter not of immediate interest), or whether the selected topology is flawed (definitely of interest). Such statistical tests are not unfamiliar in the statistical sciences, but difficult to provide within phylogenetic practice.

Bayesian inference approaches offer a variety of statistics to assess model-to-data fitness, including the so-called posterior predictive (PP) tests (Bollback 2002; Foster 2004; Brown 2014b). These tests draw samples from the posterior distribution and assess the distribution of a fitness statistic to the data. We will discuss some of the proposed methods throughout this review.

Within the maximum likelihood framework the development of solutions to the same challenge has been limited. Methods have been proposed before (Goldman 1993a, 1993b), but have been discounted due to lack of power (Waddell et al. 2009), or due to the lack of residual analysis tools in the event of rejection (pp. 495–496 Swofford et al. 1996). At present a number of promising approaches remain. Here we will present available tools in both the maximum likelihood and Bayesian frameworks, discussing their pros and cons, before suggesting some potential paths for improvement. For the sake of this discussion, we have restricted our focus to nucleotide models. Considering the reduced complexity and state space of nucleotide models, it will enable a clearer understanding of the mechanisms involved in goodness of fit tests. However, with suitable development these approaches can then be extended to other data types.

The need for goodness of fit tools is not a new idea and has previously been brought to light in reviews by Sullivan and Joyce (2005) and Kelchner and Thomas (2006). In this discussion, we wish to extend their ideas by providing an updated and more detailed insight into the current leading methods for assessing model adequacy.

This review will address two approaches for assessing model-to-data fitness. First we will discuss omnibus tests, tests that assess the overall adequacy of a model in explaining the data. Such tests usually state whether a model fits well or not. Next we will discuss residual diagnostics tools, which may diagnose the type of violation once an omnibus test has identified lack of fitness.

Goodness of Fit Statistics

Tests assessing the absolute goodness of fit generally aim to answer whether it is likely or not that an inferred phylogenetic model $\boldsymbol{\theta}$ gave rise to a set of observations $\boldsymbol{X}$. In this work we will use $\boldsymbol{\theta}$ to denote the phylogenetic model used for inference. We assume that $\boldsymbol{\theta}$ represents the model acquired after the initial model selection analysis (see Sullivan and Joyce 2005 for a good review of models and the model selection methods).

An essential part of the model is the tree topology. On the one hand, the topology can be fixed and thus a phylogenetic model is a set of parameters on a fixed topology. In this case, a goodness-of-fit test tests the hypothesis that the data come from this specific topology. On the other hand, we could consider the topology a free parameter of the model, in which case a goodness-of-fit test more generally tests the hypothesis that the data arise on a tree-like topology. In this work, we choose the latter convention.

The observations are represented as a $k \times n$ data matrix $\boldsymbol{X}$, where each of the k taxa are represented by a homologous sequence of length n. Each column in the data matrix determines an observed character $\boldsymbol{a}$ on the leaves of the phylogenetic tree. The state set of nucleotides is denoted as $\mathcal{N} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, with $\mathcal{N}^k$ being the character set, that is the set of possible realizations one can observe in a column on $\boldsymbol{X}$.

Absolute goodness of fit tests provide the opportunity to reject the "best" model (or rather find evidence against the choice), given that we observe a lack of fit to the data. This requires the construction of a suitable hypothesis test. Driven by specific questions, the hypotheses propose alternative descriptions of the behavior of the data. To perform such tests, a comparative statistic is used to quantify the behavior of the data, and is then used to assess the hypotheses.

General Statistics

In order to assess the goodness of fit, Cressie and Read (1984) formalized the family of power-divergence statistics. The two most famous statistics from this family are the deviance (G) and the Pearson X^2 statistic (X_P^2). These statistics have slightly different behavior towards deviations of individual observations with Pearson being more sensitive than the deviance (Read and Cressie 1988, Chapter 6). Therefore, X_P^2 has the ability to highlight local lack-of-fitness, while G may be seen as assessing more global lack of fitness (e.g., Waddell 1995). As the number of observed characters $n_{\boldsymbol{a}}$ approaches infinity, however, G will approach X_P^2 (Cressie and Read 1984). For our phylogenetic data, these are defined as

$$G = F_0(\boldsymbol{X}, \theta) = 2 \sum_{\boldsymbol{a} \in \mathcal{N}^k} n_{\boldsymbol{a}} \left(\log n_{\boldsymbol{a}} - \log m_{\boldsymbol{a}} \right), \tag{1}$$

$$X_P^2 = F_1(\boldsymbol{X}, \theta) = \sum_{\boldsymbol{a} \in \mathcal{N}^k} \frac{\left(n_{\boldsymbol{a}} - m_{\boldsymbol{a}}\right)^2}{m_{\boldsymbol{a}}} = \sum_{\boldsymbol{a} \in (\boldsymbol{X} \cap \mathcal{N}^k)} \frac{n_{\boldsymbol{a}}(n_{\boldsymbol{a}} - m_{\boldsymbol{a}})}{m_{\boldsymbol{a}}}, \tag{2}$$

where $m_{\boldsymbol{a}}$ is the expected frequency of character $\boldsymbol{a} \in \mathcal{N}^k$ under model $\boldsymbol{\theta}$. Under ideal conditions, both statistics follow a χ^2_{df} distribution, with the df corresponding

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

to the residual degree of freedom (the difference in degrees between the fitted model $\boldsymbol{\theta}$ and the unrestricted multinomial model).

The deviance statistic has been studied extensively in the context of phylogenetic analysis (Navidi et al. 1991; Reeves 1992; Goldman 1993b). In the standard hypothesis test setting, it is used in a nested hypothesis which states:

H_0 : The data have been generated under the selected phylogenetic model $\boldsymbol{\theta}$.

H_1 : There is no restriction on the model other than $\sum p_i = 1$.

The statistic measures the cost of fit by selecting a particular model. In this case, it measures a phylogenetic model $\boldsymbol{\theta}$ over the model of highest entropy, that is the assumption that the data come from a multinomial model only showing the observed site patterns $\boldsymbol{a} \in \boldsymbol{X}$. Under the multinomial assumption, the associated maximum likelihood estimates $\hat{p}_{\boldsymbol{a}} = n_{\boldsymbol{a}}/n$ are used to calculate the likelihood.

Both statistics are widely applied across the statistical sciences, making them a popular tool for assessing quality of model fit. Due to their popularity, the majority of statistical software provides an implementation of these statistics when fitting a model (e.g., the `deviance()` function when fitting a linear model within R).

The most common model-to-data fitness approach in Bayesian phylogenetic inference is PP testing. The PP approach creates samples from the posterior distribution to estimate the distribution of a fitness statistic conditional on said distribution. Model-to-data adequacy is then assessed by looking at the location of the statistic for the original data relative to this distribution (for an overview of methods Brown 2014b). PP approaches have been implemented in most Bayesian phylogenetic packages thus providing the means to test the fit between model and data. We highlight a few interesting goodness-of-fit measures here.

Bollback (2001) suggested using the multinomial test statistic T for the PP assessment, which was later studied more thoroughly by Bollback (2002). The multinomial test statistic is given as

$$T(\boldsymbol{X}) = \left(\sum_{\boldsymbol{a} \in \boldsymbol{X} \cap N^k} n_{\boldsymbol{a}} \log n_{\boldsymbol{a}} \right) - n \log n.$$

Essentially, this statistic assesses the entropy of the sample space. To assess model-to-data adequacy one then calculates the number of times the entropy of the posterior samples exceeded the entropy of the original data. However, this approach has been criticised for its preference of distributions with high posterior variation (e.g., Lewis et al. 2014). Foster (2004) suggested to use a Pearson-type statistic instead, and use PP to simulate a distribution of the X^2 statistic, thus extending the work of Goldman (1993b).

Brown (2014a) used PP approaches to test the impact of bias on the inference. In this work, the test statistic has been defined as the difference in entropy of the prior and the posterior distribution. The size is then assessed using the PP p-value which is defined as the proportion of samples in the PP distribution with a test statistic value less or equal to the observed value. Brown called his approach *phylogenetic plausibility*. A result was that the tree topology was surprisingly robust to even quite severe biases in branch length estimation.

A Problem of Phylogenetic Data

Aside from the lack of implementation, the use of the deviance and Pearson X^2 within phylogenetics is not without limitations. A number of studies found both statistics often suffered from the form of data usually encountered in phylogenetics, with the data having a strong effect on their performance. Under ideal data conditions, both G and X_P^2 are approximately χ^2_{df} distributed. For the χ^2 distribution to be appropriate, a number of assumptions need to be satisfied - a large sample size, independence between sites, and at least 80% of characters having an expected count of at least five, with all characters having an expected count greater than one (Cochran 1952). However, within the phylogenetic context most of these assumptions are often violated.

We know that independence of sites is rarely given considering the nature and function of DNA. For instance, RNA and protein-encoding genes provide an inherent structure, and even noncoding regions may be subject to secondary structure restrictions. The assumption of expected counts rarely holds. For example, given that we have a tree with 10 leaves, we automatically have $4^{10} = 1,048,576$ potential characters. Under the assumption that 80% of these characters need an expected count of five or more, we can begin to see how large our data sets are required to be. In practice, acquiring such a sizeable data set is not always realistic or feasible. This observation also leads to the question of adequacy of the multinomial state space.

A study by Goldman (1993b) confirmed the inadequacy of the χ^2 distribution. Goldman compared the χ^2 distribution with a simulation approach (the Goldman–Cox test, GC) to evaluate the goodness of fit by G. The GC test alleviated the dependency on a known χ^2 distribution, by simulating samples under the null models whilst re-estimating parameters for both models. This obtained a range of plausible values for G under the null hypothesis (model-to-data fit is adequate), which was used to determine whether the observed G from the data indicated an adequate model-to-data fit. The results using the GC test confirmed the inadequacy of the distribution, with

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

the χ^2 statistic grossly overestimating the value of G for most phylogenetic data.

Waddell (1995) highlights a number of possible reasons for this overestimation. The χ^2 approximation requires the necessary degrees of freedom (df) to be known. However, estimating the df for a given phylogenetic model can be problematic, with currently no confirmed approach to accurately approximate the df associated with a given topology. For most conventional tests of relative goodness of fit estimating the df is not a problem. Model comparison tests share a common topology, allowing the associated tree df to cancel, and the numbers of free parameters being simple to count. However, when dealing with tests for the absolute goodness of fit, this is not the case. The topology df remains unknown, causing the df for the fitted model and null hypothesis to be ambiguous. As the expectation of the χ^2 distribution is equivalent to the associated df, without a suitable estimation of this value, the estimation for G could be completely inaccurate. This poor estimation of the df results in a lack of power for the G statistic when assessed using a χ^2 distribution (Waddell 1995).

Waddell (1995) further highlighted the problem of sparseness in nucleotide data. Alignment data $\boldsymbol{X}$ frequently include characters that are rarely observed or not observed at all, with the character space $\mathcal{N}^k$ only sparsely sampled. The few characters sampled may have a strong influence upon the inference, with the sparseness of the phylogenetic data matrix implying a lack of power of the test employed (Waddell et al. 2009). In fact, the low sampling of states means that we actually deal with zero inflation (lots of unobserved characters in $\mathcal{N}^k$) instead of just sparseness. Taking both factors into account might improve the performance of these statistics.

The GC test acknowledges the fact that classic distributions might not be a good fit if the state space is difficult to sample, and does not require a known χ^2 distribution. However, it needs to be stressed that not re-estimating all the parameters when sampling under the null models can have serious issues with regards to statistical power, so should be applied with caution (Goldman 1993b). As it stands, p4 (Foster 2004) is the only available implementation of the GC test. PAUP* (Swofford 2011) does allow the user to calculate both the multinomial likelihood and the model-based likelihood, so with suitable adaptation the GC test could be performed. In addition, PhyML (Guindon et al. 2010) calculates G but provides no assessment of effect strength.

A handful of studies have used the GC test alongside G or X^2_P to assess model adequacy for differing models (Foster 2004; Lanfear and Bromham 2008; Duchêne et al. 2010). A number of these applications interestingly concern the test for violations of compositional homogeneity in alignment data. Foster (2004) used the Pearson X^2 statistic alongside the GC test to investigate whether the composition of the model fits the composition of the data (model-to-data fitness). Later work by Ababneh et al. (2006) adapted the Bowker (1948) test statistic (a measure for matrix symmetry applied to nucleotides in the data matrix) to use the data in guiding the choice of a suitable substitution model. This area of research is relatively unique in its development of model-fitness tools and provides a useful example of how such tests are being applied and developed.

Temporal heterogeneity which accommodates changes in the rate matrix across branches of a tree was used in Foster (2009), similar to the model allowing a change in composition over the tree in Foster (2004). In this case, each branch is assigned its own rate matrix, reflecting a change in transition patterns across time. Such models potentially lead to a large increase in parameters. The more heterogeneity our models permit, the more the issue of overfitting comes into play. To address overparameterization when estimating branch-specific rate matrices, Jayaswal et al. (2011) suggested a branch classification approach where branches with similar behavior are grouped and a common rate matrix estimated. Foster (2004, 2009) restricted the number of parameters by placing a minimum number of composition vectors or rate matrices on the tree, with each vector or rate matrix having the potential to be assigned to more than one branch. It should also be noted that overparameterization can be less of a problem for Bayesian inference, by selecting appropriate priors for the amount of heterogeneity.

Ripplinger and Sullivan (2010) applied the GC test and the PP approach proposed by Bollback (2002) as a model selection feature. Ripplinger and Sullivan calculated the test statistics (G and multinomial likelihood, respectively) for a number of substitution models, and then selected the model with the least parameters that had not been rejected by the tests. They found these approaches tend to favor simpler models than established statistics like AICc or BIC. They also found that for empirical data, PP is more likely to reject a model than the GC test, while for simulated data PP is less likely to reject a model than GC as long as the substitution model includes the rates-across-sites component. Note, that the simulated data were generated under the GTR+I+Γ model thus giving an impressive picture of the impact of the Γ component in phylogenetic models.

Marginalized Tests

Waddell et al. (2009) suggested the use of marginalized tests as another approach to assess the absolute goodness of fit. Marginalized tests group similar characters to increase the observed frequency of pooled characters. The marginalization may regard looking at subsets of taxa or at a simplified version of the data matrix $\boldsymbol{X}$, thus decreasing the state space and reducing the sparseness of the data. The deviance and Pearson X^2 statistics are then applied to the subsets before summing

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

(as opposed to each individual character under the traditional approach).

A study by Waddell et al. (2009) compared an application of the standard deviance statistic with a marginalized test. The results indicated that whilst the deviance is considered to be the uniformly most powerful test (given the data meets certain conditions as previously discussed), the marginalized test outperformed it, due to it better accounting for the sparseness in the data. For a data set knowingly not suited to the model in question, the deviance test did not reject the fit of data to the model, whereas the marginalized test consistently did. In fact, marginalization enabled the deviance statistic to regain power (Waddell et al. 2009). The marginalization even provided an additional strength, with different marginalizations detecting different deviations from the fit (Waddell et al. 2009). However, binning data can often correspond to a loss of information. This is generally viewed as a negative aspect for a statistical tool. Nonetheless, Waddell et al.'s study did indicate the potential behind this approach, but the lack of any implementation in software prevents an accessible application of such tests.

Dealing with Ambiguous or Gappy Sites

All of the above methods assume a gap-free and ambiguity-free alignment, that is all the observed characters do not have ambiguous nucleotides and there are no gaps or unsequenced regions. However, this is seldom the case in practice. The effect of gappy sites on the phylogenetic inference is not a well understood problem (Hartmann 2008). The popular protocol is to simply remove the gappy or ambiguous sites from the analysis. However, a number of times this has been found to be problematic for the inference (Loytynoja 2009; Dessimoz and Gil 2010).

How to handle such data is also a key concern when assessing model adequacy. As Waddell (2005) pinpoints, removing the data before applying tests (such as G or X_P^2) is often undesirable or impractical, with much of the data being discarded. As an alternative approach, Waddell suggested a fit in which missing or ambiguous characters are considered to share the likelihood of those characters which are observed in the alignment. For example, if we observe the ambiguous character `AANC` and the alignment also contains the observed characters `AATC` and `AACC`, then the likelihood of the ambiguous character is simply the sum of these two characters' likelihoods. The modified observed and expected frequencies are then used with the G or X_P^2 statistics to assess the absolute fit of data to model. While this approach is practical it also treats missing and ambiguous characters as the same, which is problematic. The latter is generally a sequencing artifact while the former could also be attributed to a loss event in the evolutionary history.

The implemented tests have different ways in dealing with this. P4 will not do the test if the alignment has gappy or ambiguous sites while PhyML only includes fully resolved sites in the calculation of the statistics.

Methods for Residual Diagnostics

Rejecting a model is relatively uninformative unless there is supporting information that can direct the following inference steps (pp. 495–496 Swofford et al. 1996). Given that we have seen a violation in the model-to-data fit, it would be useful to gain more information about the nature of this failure. The common approach in statistical methods is to use *residual diagnostic tools* as a second step after the initial model adequacy check. The term 'residual diagnostics' is used in regression analysis (estimating relationships among variables) and calculates the difference between the observed and fitted values. Here we can apply the same logic to our framework, by looking at specific sites that violate the model fit. Residual diagnostic tools offer a more detailed assessment of the model-to-data fit and enable the user to understand *where* and *how* the model shows poor fit to the data. This can then be used to direct future steps in the model-fitting process.

We discuss a number of residual diagnostic approaches that have been proposed for phylogenetic practice. These use a mixture of quantitative and qualitative assessment to pinpoint model-to-data fit violations. It should be noted, that most of the methods discussed here were introduced to filter data and did not deal with model fitness directly. The filtering methods assign a fitness statistic to every site in an alignment and filters out those sites that exceed a threshold which corresponds to stability in the inference. For residual diagnostics the filtering step is inadequate. However, the fitness statistics employed permit a visualization of the variability of alignment sites and can highlight sites and areas within the alignment which may be difficult to model. The inclusion of filtering approaches here should be seen from this point of view rather than as applied.

Likelihood Scores

One way to address areas of poor model-to-data fit is to identify those columns of the data matrix (observed characters) that contribute the most to the deviance. The inadequacy of a selected model is often reflected by large deviations between the observed and expected (under $\boldsymbol{\theta}$) character frequencies.

Nguyen et al. (2011) developed a simple approach to visualize this deviance using essentially a weighting system. The Misfits approach aims to evaluate the goodness of fit, through computation of an associated biologically motivated score relating to these deviations. The deviations are simply the deviance statistic

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

calculated for an individual character $\boldsymbol{a}$ (individual contribution from Equation 1).

The Misfits method is built upon the acceptance that the model-to-data fit is never completely accurate; there is naturally a difference between the expected and observed character frequencies (residuals). To account for variation in the data, Nguyen et al. (2011) computed the simultaneous 95% Gold confidence region for multivariate proportions (Gold 1963) as

$$CI_{\text{Gold}} = p_{\boldsymbol{a}} \pm \sqrt{\frac{\kappa p_{\boldsymbol{a}}(1-p_{\boldsymbol{a}})}{\ell}}, \tag{3}$$

where κ is the $0.05/\ell$-quantile of the χ^2_{df} distribution, with the degrees of freedom corresponding to the number of estimated variables in the inferred model $\boldsymbol{\theta}$ (substitution model parameters and number of branches). The confidence interval corresponds to the likelihood for character $\boldsymbol{a}$ under the model $\boldsymbol{\theta}$, and creates a region of acceptable deviations.

Characters are classified as *over-represented* ($\mathcal{D}^+$) when the observed frequency exceeds the upper bound of the confidence interval, and as *under-represented* (D^-) when the observed frequency is below the lower bound of the confidence region. The Misfits score is then computed by considering the number of mutations it takes on the tree to transform instances of over-represented characters into under-represented characters.

Lower Misfits scores correspond to a situation in which few additional substitutions are required, suggesting a more accurate model to data fit. By classifying each character, the Misfits approach allows specific areas of model violation to be pinpointed. The method considers the underlying biology behind the evolutionary process when attempting to identify outliers. This provides a strength of the Misfits approach, drawing on both the biological and statistical theory combined in phylogenetics.

However, the problem of the Misfits approach lies in the great disparity between available data and possible characters. This leads to the common observation that many more characters are over- rather than under-represented. In extreme instances all characters may be over-represented, thus providing no under-represented characters to transform them to, including characters not observed. The method would benefit from exploring how to better handle the unequal under- and over-representation of characters.

The Misfits approach is undoubtedly an interesting tool. However, Holland (2013) discussed how the seemingly "back-to-front" nature of the method, provides the opposite to how a statistician would typically assess the goodness of fit. The method firstly highlights areas of poor fit, before exploring how to *change the data* to fit the model. However, such methods should rather investigate how to *change the model* to better fit the data.

In addition, it should be stressed that in its current form, the Misfits score does not represent an effect size. The statistic purely is a reflection of the number of additional substitutions required and is dependent upon the number of taxa within the inference. In order to be an effect size, the Misfits score should be standardized. This enables us to understand the magnitude of the effect on a general scale that is no longer specific to this data set only, which is more informative about the scale of violation between data and model.

Influence Scores

Influence functions have been used extensively throughout regression analysis to identify outliers of a fitted model (Rousseeuw and Leroy 1987). Bar-Hen et al. (2008) adapted this method for phylogenetic models, to assess the impact of a single site on the likelihood for a model.

Define $\boldsymbol{X}$ to be the original alignment and $\boldsymbol{X}^{(i)}$ to be the alignment after removing site i. Let $\boldsymbol{\theta}$ denote the model inferred from $\boldsymbol{X}$ and $\boldsymbol{\theta}^{(i)}$ the model inferred from $\boldsymbol{X}^{(i)}$. To assess the influence of a site i on the model $\boldsymbol{\theta}$, the influence function $\mathbb{F}_{\boldsymbol{\theta}}(i)$ is calculated by

$$\mathbb{F}_{\boldsymbol{\theta}}(i) = (n-1)\left[\log L(\boldsymbol{\theta}|\boldsymbol{X}) - \log L\left(\boldsymbol{\theta}^{(i)}|\boldsymbol{X}^{(i)}\right)\right], \tag{4}$$

where L denotes the likelihood. The influence score represents the change in average likelihood from removing site i. If $\mathbb{F}_{\boldsymbol{\theta}}(i)$ is positive then site i does not support $\boldsymbol{\theta}$, whilst a negative influence score indicates that the site does support $\boldsymbol{\theta}$. The method considers sites with the highest influence scores as *outlier sites*, due to their lack of support for $\boldsymbol{\theta}$. The authors further suggested the use of topological influence indices like interior node changes as a means to test the topological adequacy of each site.

Similar to the Misfits approach, the method's strengths lie with the ability to identify regions along the alignment that display poor model fit. Yet unlike the approach discussed above, the normalized influence scores do provide a suitable representation of the effect size. Likelihood influence scores highlight sites that may disrupt stable inference while topological influence scores may identify topological stability within an alignment.

Lewis et al. (2014) introduced a similar approach that can be considered as a Bayesian analogue to the influence score—the conditional posterior ordination (CPO) scores. This cross-validation type approach calculates the CPO score for each alignment site and sums the scores to get the log pseudomarginal likelihood. Similarly to the influence scores, the sitewise scores give an indication of individual suitability of sites. In particular, Lewis et al. demonstrated the potential of this index in showing partition structure in the data. The authors also proposed to combine insights from multiple indices to assess model fitness. This stressed the point that each statistic measures a particular type of goodness of fit (e.g., phylogenetic signal, site fitness), and therefore a range of test statistics can assess fitness more generally.

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

Filtering Methods

In addition to the methods mentioned above, there are also a number of measures which were originally introduced to filter the data. These methods do not deal with model-to-data fit directly, but do highlight a particular form of behavior in sites which may be affecting the fit. The ability to identify such classes of sites might be useful in developing future residual diagnostic tools to be applied once a lack of overall model-to-data fit has been detected. Note that contrary to the filtering methods, for residual diagnostics such statistics provide a grouping of all sites permitting to diagnose behavior in classes of sites rather than only the most "abnormal."

Phylogenetic models often accommodate variation in the rate of evolution across sites (Yang 1994; Lockhart et al. 1996). Sites with a high rate have been linked to saturation issues, one of the causes of misleading phylogenetic signal and long branch attraction (LBA; Rodríguez-Ezpeleta et al. 2007). LBA incorrectly infers distantly related taxa to be closely related due to convergent and parallel evolution, and may result in estimating an incorrect tree.

Methods of identifying and removing fast-evolving sites have been proposed and used a number of times for phylogenetic data (Hirt et al. 1999; Ruiz-Trillo et al. 1999; Burleigh and Mathews 2004; Rodríguez-Ezpeleta et al. 2007). These studies used the substitution rate scores directly as a metric to rank the sites. Results indicated a reduction in nonphylogenetic signal after removing the fast-evolving sites, leading to a more stable inference. However, the approach is hindered by the importance of the topology used. The topology affects the rate estimation, and thus can heavily influence which sites are identified as outliers. Using an inaccurate topology could result in removing sites critical to the inference.

Rodríguez-Ezpeleta et al. (2007) confirmed the strong influence of topology selection, suggesting the rates should be averaged over a few best possible topologies, prior to identifying the fast-evolving sites. A further suggestion involved grouping taxa and investigating the within and across group rate variation before removing the saturated sites (Brinkmann and Philippe 1999; Lopez et al. 1999). However, it should be noted that caution needs to be taken when specifying the groups to prevent misspecification within the alignment data. Both suggestions have shown to alleviate some of the dependence on the topology. However, the former relies on the knowledge of a few best possible topologies, which may not necessarily be known, whilst the latter increases the computational burden.

Instead of using (group) rates to pinpoint fast-evolving sites, Pisani (2004) suggests a method built on the compatibility definition of Le Quesne (1969). As a brief explanation, "two characters are deemed compatible if they can be mapped onto the same topology without homoplasy." The method calculates an incompatibility score for each site in the alignment, which is the number of sites incompatible with i. High incompatibility scores indicate that the associated site could be saturated, with most of their phylogenetic information being lost. Compatibility methods are not dependent on the topology, providing an advantage over the previously discussed rate scores (Pisani 2004). For this reason, Cummins and McInerney (2011) used the compatibility work of Pisani to identify fast-evolving sites. The compatibility approach removes the potential problems of group misspecification within the alignment data, present in the method of Brinkmann and Philippe (1999) above. However, there is no clear definition as to when a site is no longer an outlier and could contain actual signal to be incorporated by the model.

Goremykin et al. (2010) suggested a different method to pinpoint saturated sites within an alignment. The premise was to identify an index that is topology-free and permits an ordering of sites for filtering. For each site the method calculates the observed variability distance (OV distance), which is the proportion of mismatched sequence pairs over all sequence pairs. The method then orders the sites according to their OV-distance in decreasing order and removes sites until a stable topology is obtained. A study by Zhong et al. (2011) used the approach to drop the most varied sites from a chloroplast genome alignment. General results indicated removing sites based on their OV scores was useful in pinpointing the fast-evolving sites, resulting in a more robust tree inference. Similarly to the other filtering methods, this approach also falls short from the lack of formal cut-off boundary to determine which sites are considered fast-evolving.

It should be stressed again that as both the OV distances and compatibility scores are computed without any knowledge of the inferred model, they cannot be used directly to assess model-to-data fitness. However, the indices could be useful in combination with other postinference indices (e.g., Misfits) to identify usual sites.

Selecting a suitable and optimal boundary for filtering methods is mostly subjective and potentially problematic. The work of Cummins and McInerney (2011) acknowledged the limits to stripping data and removing sites for this reason, but offered the approach as a useful tool in data exploration. This idea of exploring the data for specific behavior os sites aligns neatly with the aims underpinning residual diagnostic tools.

Taxon Sampling

The above methods mainly deal with identifying sequence sites or characters which are badly explained by an inferred model. However, model fit can also be improved by investigating the impact of a taxon on the fit between model and data. This moves away from considering the effect of a column in the data matrix $\boldsymbol{X}$ but rather assesses the impact of the row on the inference.

Mariadassou et al. (2012) defined a measure to assess the influence of each taxon on the phylogeny—the *taxon influence index* (TII). This measure was used

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

to detect influential taxa which strongly impacted the phylogenetic estimates. The taxon influence index quantifies the effect of removing a taxon on the stability of the tree inference. Using any inference method, we define T^* to be the tree inferred from the complete alignment. Let T_k be a smaller tree, inferred from the alignment lacking taxon k. Taxon k is then excluded from T^* to produce tree T_k^*. Thus, the TII is defined as the distance between trees T_k and T_k^*, such that

$$\mathrm{TII}(k) = d(T_k, T_k^*). \quad (5)$$

Small values for the TII indicate that taxon k does not change tree T, with larger scores correspond to more influential taxa. This influence can either highlight a node introducing or resolving bias.

Mariadassou et al. (2012) presented a case study to demonstrate the performance of the index. The majority of taxa showed weak influence on the phylogeny. However, a fraction of taxa were found to be highly influential, altering the phylogeny even in clades only loosely related to them. Such an observation can be seen as a sign of long branch attraction (LBA). Several methods to detect LBA have been proposed (e.g., Bergsten 2005), and it might be worthwhile studying the performance of TII with those methods. However, it also needs to be stressed that a common solution to LBA is to add taxa. Such taxa would most certainly be identified as influential since they resolve the observed LBA issue. Thus, one needs to be careful before treating high influence as detrimental to the inference.

Conclusions

Despite the substantial development of phylogenetic inference over the last few decades, one area still remains very much wanting—the test for goodness of fit between model and data.

Assessing model-to-data fitness is a critical protocol within any statistical model building process. It is imperative for the selected best model to reflect what the data is telling us, and to reflect it accurately. Or in the words of Gelman et al. (1995):

> We do not like to ask, "Is our model true or false?," since probability models in most data analyses will not be perfectly true. … The more relevant question is, "Do the model's deficiencies have a noticeable effect on the substantive inferences?"

This reinforces the importance of whether the model-to-data fit is indeed adequate or not. An inadequate model fit would affect any future inferences made based on this model. Furthermore, given that we determine an inadequate model-to-data fit, consequent steps need to be taken to assess where the violations occur, and thus how to adapt for these.

Combining these ideas, we have the following approach for assessing a model once it has been determined as the "best":

1. Is the overall model-to-data fit adequate or not?
2. Given the model fit is not adequate, where do the model-to-data fit violations occur?
3. Given that we have found the violations, how do we then deal with this?

The first stage concerns the use of goodness of fit statistics to assess the absolute fit between model and data. In the phylogenetic context a couple of approaches have been proposed in a maximum likelihood framework. However, due to the nature of phylogenetic data, the power-divergence statistics quickly become unsuitable for our problem. To alleviate some of these issues, marginalized tests were introduced. Unfortunately, the lone explicit proposal (Waddell et al. 2009) uses a complex grouping criteria, which has prevented any implementation in software, let alone an efficient and easily applied one. Due to the nature of phylogenetic data, the use of nonparametric statistics could alleviate the dependency on predefined distributions.

Goodness of fit assessment is an area in which Bayesian methods are ahead of their frequentist counterpart. Bayesian statistics employs the PP distribution of states to simulate a distribution of an adequacy statistic (e.g., Bollback 2002; Foster 2004; Brown 2014b). The statistics used to assess the fit resemble the likelihood statistics discussed above. However, the Bayesian approaches take the uncertainty in model parameters into account (including the topology and branch lengths, e.g., Bollback 2002). Foster (2004) found the multinomial test statistic to be less than satisfactory and not sensitive enough to detect inadequate model-to-data fit. In addition, the power of the multinomial test statistic can be affected by a number of factors. Most notably the statistic loses power if the model assumptions are violated and can lead to high confidence in potentially wrong topologies (Bollback 2002). PP tests have shown promise in addressing the issue and are constantly adapted and developed to provide robust results (e.g., Duchêne et al. 2016, 2018). Finding adequate analogs for maximum likelihood inference remains an open challenge.

From a practical perspective, rejecting the fit of a model is only an initial step. In fact, this is only useful if it provides an indication of how to continue the inference. In such a case, the need for a set of residual diagnostic tools becomes critical, leading us to the second stage when assessing model fit.

Over the years a number of promising approaches have been proposed; most notably the use of likelihood and influence scores (Bar-Hen et al. 2008; Nguyen et al. 2011). All the discussed methods (Misfits, OV distance, influence scores etc.) offer a number of ways to classify and highlight sites which display behavior deviating from that expected under the fitted model or deemed as "noisy."

However, one common observation is evident across all residual diagnostic methods—the inconsistent

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

perception of what a "noisy" site is. This issue might arise from the multifaceted idea of noise itself. For example, sites that have accumulated too many mutations over time, that is sites that accrue a lot of changes on any topology, are considered as noisy. Further, sites might have been subject to horizontal transfer, that is the model would infer a different splitting pattern for the site compared to the rest.

From this it becomes apparent that we firstly need to determine what a good diagnostic method should discover. In such a situation, it might be important to determine which noise type is identified by which of the currently proposed approaches. This could allow methods to also pinpoint not only which sites may be causing the inadequate model fit, but also why this might be the case.

This leads us to the third stage of our process—given we have found the sites that do not fit our model well, what do we do then? Standard regression practice asks for removal of influential sites from the inference but keeping track of the culprits for further assessment as they might provide an interesting exception to the story. In phylogenetics such methods are still needed to improve in part due to the nature of data and model. As in many complex models, pinpointing a model violation to a particular source can be incredibly difficult since parameters might interact, are hierarchically linked, or the violation is not modelled. Rather than pinpointing a source of violation to a single source why not provide a spectrum assessment (e.g., 70% chance that the violation is due to horizontal transfer, 20% chance that it is due to a long branch attraction, and 10% chance that it is simply due to noise).

In addition, we could think about not only pinpointing specific observations that display poor model-to-data fit, but rather the specific parameters. For example, if the parameter of interest was the topology, then investigating clades would be beneficial. This is something already done in bootstrapping approaches and would be a useful idea to explore when assessing goodness of fit.

However, the question still remains—how do we handle observations displaying poor model-to-data fit? If the sites are resulting in the wrong processes being determined, then removing them might be the best solution. However, if the sites are compounding processes that contain useful signal, these should not be removed, but rather have their potential behavior explored. Exploring such sites would allow better assessment of the behavior and be extremely useful in modelling the underlying biological processes. However, at present it is not obvious which steps to take next. Without an established approach on how to handle such sites, employing residual diagnostic tools quickly becomes redundant. Ideally, the development of informative statistics (both at the omnibus and site level) would help direct the next steps to take in the model-building process.

Whilst considering the open challenges in model adequacy assessment, we should also emphasize the need for an implementation of these tools. This step again seems trivial, but unfortunately is one of the major reasons behind the lack of application for these tests in phylogenetic practice. Implementation of these tools in software is fairly rare (both at the omnibus and site level), causing many methods (i.e., deviance statistic, Misfits, influence scores, OV distance etc.) to be often overlooked and ignored within practice. Combining with popular inference tools such as RAxML (Stamatakis 2014) would be advantageous. Creating an available implementation for users across a range of disciplines (statistics, biology, computer science etc.) would aid in encouraging their use.

This discussion would be incomplete without acknowledging we are in the era of genomics, with the increasing size of phylogenetic data sets. In such a situation, the feasibility of approaches pinpointing specific sites quickly comes into question: as the alignment increases in length, how much impact will a single site really have on the model fitted? Assessing the influence of a single site seems not only overkill for large data but quickly becomes computationally intensive. This is a consideration that may be useful when developing new tools in the field.

In the era of genomics, we may also consider that the unit of influence on model-to-data fit might not necessarily be a site, but in fact a gene or partition. Genomics presents an increasing amount of biologically processes, suggesting the impact of a single site may no longer be of key interest. A number of people are currently working on this idea, by considering wider scale mechanisms in the genome and a more realistic approach to pinpoint these (Delsuc et al. 2005; Lanfear et al. 2012).

Further, Zhong et al. (2011) indicate an issue of the central limit theorem. We have sufficient data to be very confident about our inference. But this confidence comes not from low variability in the model but from the sample size. In these cases, an evaluation of the model variability (through cross-validation approaches) should be used to test for systematic errors. Such approaches have been suggested in the Bayesian framework (e.g., Duchêne et al. 2016) based on the cross-validation work of Lartillot et al. (2007).

Finally, we need to consider that the more tests we perform, the more prone we are to find something. In such cases we could become victim to Type 3 errors, mistaking statistical significance for practical relevance. For instance, in Nguyen et al. (2011) the Misfits approach identifies four mutations to explain the deviation between a mitochondrial genome alignment and the model. This was shown to be statistically significant and spun into a story about functional differences. Still, four changes in a 16 kb alignment seems not that relevant. In such a situation it could be more useful to assess the impact of regions of sites along the alignment, as opposed to the individual effects. However, this seems almost like a "running before you walk philosophy," and still requires useful fitness tools to be developed (albeit at the site or region level).

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

When looking into the previously proposed methods, it becomes apparent that adapting goodness of fit tools to the phylogenetic framework is not an easy process. Nonetheless, this does not remove the need for such a critical step within any model building process. For instance, how useful is a model if we really have no understanding of how well that model fits our data? The development of powerful statistics and availability within an accessible implementation would hopefully increase the use of goodness of fit procedures in phylogenetics. An ideal set of such statistics will help identify missing or unnecessary parameters, visualize the variability within a sequence alignment, and provide an overall level of confidence in the aspects of interest within an inference. Any approach addressing these points should lead to an increased use of model-to-data fitness statistics.

Acknowledgments

The authors would like to thank Peter Foster, Cymon Cox, Ed Susko, and other anonymous referees for their constructive criticism and suggestions throughout the reviewing process.

References

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

Baele G., Lemey P., Suchard M.A. 2016. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Syst. Biol. 65:250–264.

Bar-Hen A., Mariadassou M., Poursat M.-A., Vandenkoornhuyse P. 2008. Influence function for robust phylogenetic reconstructions. Mol. Biol. Evol. 25:869–873.

Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163–194.

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.

Brinkmann H., Philippe H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Brown J.M. 2014a. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Brown J.M. 2014b. Predictive approaches to assessing the fit of evolutionary models. Syst. Biol. 63:289–292.

Burleigh J.G., Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

Cochran W.G. 1952. The χ^2 test of goodness of fit. Ann. Math. Stat. 23:315–345.

Cressie N., Read T.R.C. 1984. Multinomial goodness-of-fit tests. J. R. Stat. Soc. Ser. B (Methodological) 46:440–464.

Cummins C.A., McInerney J.O. 2011. A Method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60:833–844.

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

Delsuc F., Binkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Dessimoz C., Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11(4):R37.

Doyle V.P., Young R.E., Naylor G.J., Brown J.M. 2015. Can we identify genes with increased phylogenetic reliability. Syst. Biol. 64:824–837.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2010. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol. 34:1529–1534.

Duchêne D.A., Duchêne S., Di Giallonardo F., Eden J-S., Geoghegan J.L., Holt K.E., Ho S.Y.W., Holmes E.C. 2016. Cross-validation to select Bayesian hierarchical models in phylogenetics. BMC Evol. Biol. 16:115.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2018 PhyloMAd: efficient assessment of phylogenomic model adequacy. Bioinformatics 34:2300-2301.

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

Foster P.G., Cox C.J., Embley T.M. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos. Trans. R Soc. Lond. B Biol. Sci. 364:2197–2207.

Gatesy J. 2007. A tenth crucial question regarding model use in phylogenetics. Trends Ecol. Evol. 22:509–510.

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

Gelman A., Robert C. Chopin N., Rousseau J. 1995. Bayesian data analysis. Chapman and Hall/CRC, New York:CRC Press.

Gold R.Z. 1963. Tests auxiliary to chi squared tests in a Markov chain. Ann. Math. Stat. 34:56–74.

Goldman N. 1993a. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. 37:650–661.

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

Goremykin V.V., Nikiforova S.V., Bininda-Emonds O.R.P. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71:319–331.

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

Hartmaan S., Vision T.J. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment?. BMC Evol. Biol. 8:95.

Hirt R.P., Logsdon Jr. J.M., Healy B., Dorey M.W., Doolittle W.F., Embley T.M. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. USA 96:580–585.

Holland B.R. 2013. The rise of statistical phylogenetics. Aust. N. Z. J. Stat. 55:205–220.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Jayaswal V., Ababneh F., Jermiin L.S., Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. Mol. Biol. Evol. 28:3045–3059.

Jermiin L.S., Jayaswal V., Ababneh F., Robinson J. 2008. Phylogenetic model evaluation. In: Keith J.M., editor. Methods in molecular biology™, vol. 452. Totowa, NJ: Humana Press. p. 331–364.

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

Kelchner S.A., Thomas M.A. 2006. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22:87–94.

Kostka M., Uzlikova M., Cepicka I., Flegr J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of blastocystis. BMC Bioinformatics 9:341.

Lanfear R., Bromham L. 2008. Statistical tests between competing hypotheses of hox cluster evolution. Syst. Biol. 57:708–718.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022

Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7:S4.
Le Quesne W.J. 1969. A method of selection of characters in numerical taxonomy. Syst. Zool. 18:201–205.
Lewis P.O., Xie W., Chen M., Fan Y., Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. Syst. Biol. 63:309–321.
Lockhart P.J., Larkum A.W., Steel M.A., Waddell P.J., Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.
Lopez P., Forterre P., Philippe H. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49:496–508.
Löytynoja A., Goldman N. 2009. Uniting alignments and trees. Science 324:1528–1529.
Mariadassou M., Bar-Hen A., Kishino H. 2012. Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference. Syst. Biol. 61:337–345.
Navidi W.C., Churchill G.A., von Haeseler A. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. 8:128–143.
Nguyen M.A.T., Klaere S., von Haeseler A. 2011. MISFITS: evaluating the goodness of fit between a phylogenetic model and an alignment. Mol. Biol. Evol. 28:143–152.
Pennell M.W., FitzJohn R.G., Cornwell W.K., Harmon L.J. 2015. Model adequacy and the macroevolution of angiosperm functional traits. Am. Nat. E33–E50:309–321.
Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53:978–989.
Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.
Posada D., Crandall K.A. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.
Read T.R.C., Cressie N.A.C. 1988. Goodness-of-fit statistics for discrete multivariate data. Springer Series in Statistics. Springer, New York. ISBN 978-1-4612-4578-0.
Reeves J.H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17–31.
Ripplinger J., Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. Mol. Biol. Evol. 27:2790–2803.
Rodríguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B.F., Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.
Rousseeuw P.J., Leroy A.M. 1987. Robust regression and outlier detection. New York, NY, USA: John Wiley & Sons, Inc.
Ruiz-Trillo I., Riutort M., Littlewood D.T.J., Herniou E.A., Baguna J. 1999. Acoel flatworms: earliest extant bilaterian metazoans, not members of platyhelminthes. Science 283:1919–1923.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.
Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Ann. Rev. Ecol. Evol. Syst. 6:445–466.
Swofford D.L. 2011. PAUP*: phylogenetic analysis using parsimony (* and other methods), version 4.0b10.
Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. 1996. Phylogenetic inference, chapter 11. In: Hillis, David M, Moritz, Craig and Mable, Barbara K., eds. Molecular systematics. Sunderland, Mass:Sinauer Associates. p. 407–514.
Waddell P.J. 1995. Statistical methods of phylogenetic analysis: including Hadamard conjugations, LogDet transforms, and maximum likelihood [PhD Thesis]. Massey University, Palmerston North. https://mro.massey.ac.nz/xmlui/bitstream/handle/10179/4127/02_whole.pdf
Waddell P.J. 2005. Measuring the fit of sequence data to phylogenetic model: allowing for missing data. Mol. Biol. Evol. 22:395–401.
Waddell P.J., Ota R., Penny D. 2009. Measuring fit of sequence data to phylogenetic model: gain of power using marginal tests. J. Mol. Evol. 69:289–299.
Woodhams M.D., Fernández-Sánchez J., Sumner J.G. 2015. A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. Syst. Biol. 64:638–650.
Xie W., Lewis P.O., Fan Y., Kuo L., Chen M-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.
Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3:1340–1348.

Downloaded from https://academic.oup.com/sysbio/article/68/1/157/5133546 by guest on 21 August 2022