Syst. Biol. 71(3):630–648, 2022

DOI:10.1093/sysbio/syab073
Advance Access publication September 1, 2021

Measuring Phylogenetic Information of Incomplete Sequence Data

TAE-KUN SEO[1,2,*], OLIVIER GASCUEL[2,3], AND JEFFREY L. THORNE[4]

[1]*Division of Life Sciences, Korea Polar Research Institute, 26 Songdomirae-ro, Yeonsu-gu, Incheon 21990, Republic of Korea*
[2]*Unité Bioinformatique Evolutive, Institut Pasteur, Paris, France*
[3]*Institut de Systématique, Evolution, Biodiversité (ISYEB - UMR 7205, CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), Paris, France*
[4]*Departments of Biological Sciences and Statistics, North Carolina State University, Raleigh, NC 27695-7566, USA*
**Correspondence to be sent to: Division of Life Sciences, Korea Polar Research Institute, 26 Songdomirae-ro, Yeonsu-gu, Incheon 21990, Republic of Korea; E-mail: seo.taekun@gmail.com.*

Received 2 January 2021; reviews returned 26 August 2021; accepted 27 August 2021
Associate Editor: Simon Ho

Abstract.—Widely used approaches for extracting phylogenetic information from aligned sets of molecular sequences rely upon probabilistic models of nucleotide substitution or amino-acid replacement. The phylogenetic information that can be extracted depends on the number of columns in the sequence alignment and will be decreased when the alignment contains gaps due to insertion or deletion events. Motivated by the measurement of information loss, we suggest assessment of the effective sequence length (ESL) of an aligned data set. The ESL can differ from the actual number of columns in a sequence alignment because of the presence of alignment gaps. Furthermore, the estimation of phylogenetic information is affected by model misspecification. Inevitably, the actual process of molecular evolution differs from the probabilistic models employed to describe this process. This disparity means the amount of phylogenetic information in an actual sequence alignment will differ from the amount in a simulated data set of equal size, which motivated us to develop a new test for model adequacy. Via theory and empirical data analysis, we show how to disentangle the effects of gaps and model misspecification. By comparing the Fisher information of actual and simulated sequences, we identify which alignment sites and tree branches are most affected by gaps and model misspecification. [Fisher information; gaps; insertion; deletion; indel; model adequacy; goodness-of-fit test; sequence alignment.]

Conventional information criteria such as the AIC (Akaike Information Criterion; Akaike 1974) and the BIC (Bayesian Information Criterion; Schwarz 1978) can be used to compare models of sequence change (e.g., Posada and Crandall 1998; Posada and Buckley 2004; Seo and Kishino 2008, 2009). Relative to nucleotide substitution or amino acid replacement, less attention has been devoted to the effects of insertion and deletion when applying information criteria. One option for treating insertion and deletion is to explicitly include them in probabilistic models of sequence change (e.g., Thorne et al. 1991, 1992; Hein et al. 2000; Metzler 2003; Redelings and Suchard 2005; Fleissner et al. 2005; Bouchard-Côté and Jordan 2013; Holmes 2020; De Maio 2021). While explicit treatment is biologically and statistically appealing, it can be accompanied by daunting computational challenges.

A conventional and computationally convenient alternative to explicit probabilistic insertion-deletion models is to assume that the alignment between sequences is known with certainty. This alternative treats alignment gaps as data that are missing at random (e.g., see Felsenstein 2004) but can be especially problematic when there is substantial alignment uncertainty. Methods exist for identifying alignment columns that are prone to alignment error so that these columns need not be included in downstream analyses (e.g., Talavera and Castresana 2007; Dress et al. 2008; Capella-Gutierrez et al. 2009). However, some studies have questioned the value of filtering alignment columns in this way because removing some columns will reduce evolutionary information and may affect the reliability of downstream analyses (Dessimoz and Gil 2010; Tan et al. 2015).

In this study, we quantify the informativeness of gap-containing columns. Because gaps are being considered as missing data, a simple and intuitive set of summary statistics regarding informativeness would be the proportions of gap positions in each aligned column and in each aligned sequence. Higher proportions of gaps would represent more missing data. A limitation of these summary statistics is they do not incorporate correlations among aligned sequences that are due to common ancestry. The effect on informativeness due to the presence of a gap at one position in a single-aligned sequence will depend on which other sequences share the gap as well as on the phylogenetic relationships between the sequences.

Here, we rely upon the Fisher information to assess the impact of gaps. With an abundance of gaps, the curvature of the log-likelihood function at the maximum likelihood estimate (i.e., the Fisher information) becomes gradual relative to the more extreme curvature with an absence of gaps. By simulating sequence evolution and then introducing gaps where data should be missing, the information loss caused by gaps can be quantified. However, the Fisher information is affected both by presence-absence of gaps and model misspecification. These can be difficult to disentangle. Whereas the difference in Fisher information between

complete (i.e., ungapped) and incomplete (i.e., gap-containing) data is straightforward to assess via simulation, actual aligned data are generated according to an unknown process and ungapped versions of the actual gap-containing data are unavailable.

To quantify the impact of gaps, our approach contrasts simulated gap-containing data with the corresponding complete version of the simulated data. To assess model misspecification, the approach contrasts actual gap-containing data with simulated gap-containing data. Building upon previous work regarding model adequacy (Goldman 1993; Duchêne et al. 2018), we show that the ability to quantify model misspecification can form the basis for a goodness-of-fit test with the observed gap-containing data.

The ability to disentangle the gap and model misspecification effects permits us to compare physical sequence length (PSL) and effective sequence length (ESL). Whereas PSL is observable and is the number of columns in the sequence alignment, ESL represents the number of columns in an ungapped alignment that would be needed to match the informativeness of an alignment with the observed PSL and the observed gap locations.

After introducing our statistical approach, we characterize it via simulation and then apply it to data sets of protein sequences from eukaryotes, nucleotide sequences from ray-finned fish, and nucleotide sequences from mouse lemurs. Based on the simulations, our test of model adequacy has low power but our ESL estimates are relatively robust to model misspecification. We conclude by discussing refinements and extensions of our approach.

Theory

Basic Assumptions

Our approach has three key assumptions. First, it assumes that the alignment relating the sequence data is correct. Second, the aligned sequence columns are assumed to be independently and identically distributed random samples. Third, the approach assumes that no information is contained in gaps and this implies that the nucleotide substitution (or amino acid replacement) process is independent of the insertion–deletion processes. All three assumptions are standard in phylogenetics and the first two are widely acknowledged (e.g., see Felsenstein 2004).

The third assumption permits reliance upon likelihood-based treatments of aligned sequence data that include explicit models of nucleotide substitution (or amino acid replacement) but that do not include explicit models of insertion and deletion. Our likelihoods represent the probabilities of aligned sequence data conditional upon the substitution model, its parameter values, and the evolutionary tree that relates the aligned sequences. With this third assumption, likelihoods can be calculated by treating the alignment gaps as data that are missing at random. Rather than modeling the missing data process, the likelihoods condition upon which data are missing. In the Discussion section, we discuss this assumption of independence between substitution and insertion–deletion in more detail.

Measuring Fisher Information of Sequence Data

Consider the true and unknown data-generating mechanism $g(\cdot)$ and the adopted model $f(\cdot|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a d-dimensional vector. Although the Fisher information is represented by a matrix when there is more than one parameter, we discuss the univariate case for the convenience of explanation. A more general description can be found in the Appendix and key mathematical notation is summarized in Table 1.

Using the log-likelihood function $l(\cdot)$ at the maximum likelihood estimate (MLE; $\tilde{\boldsymbol{\theta}}$), we represent the estimate of Fisher information for the ith parameter θ_i of the adopted model $f(\cdot|\boldsymbol{\theta})$,

$$-\frac{1}{n}\frac{d^2}{d\theta_i^2}l(\tilde{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{j=1}^{n}\frac{d^2}{d\theta_i^2}\log f(\tilde{x}^{(j)}|\tilde{\boldsymbol{\theta}}) =: \widehat{\tilde{I}}_{gfii} \qquad (1)$$

$$\approx \mathrm{E}_g\left[-\frac{d^2}{d\theta_i^2}\log f(\tilde{X}|\tilde{\boldsymbol{\theta}})\right]$$

$$\approx \mathrm{E}_g\left[-\frac{d^2}{d\theta_i^2}\log f(\tilde{X}|\boldsymbol{\theta}_*)\right] =: \tilde{I}_{gfii}, \qquad (2)$$

where n is the sequence length and $\tilde{x}^{(j)}$ is the jth potentially gap-containing (i.e., incomplete) sequence column. The notation "=:" means the term on the right is defined as the term on the left. The "˜" sign over the data $\tilde{x}^{(j)}$ and over the Fisher information $\tilde{I}_{(\cdot)}$ implies the data may potentially contain gaps. The first and second subscript indices of $\widehat{\tilde{I}}_{gfii}$ and $\tilde{I}_{gfii}$ respectively represent the true and adopted models, while the third and fourth indices represent the row and column position in the Fisher Information matrix. Although only diagonal elements of the Fisher information are considered here, the theory can be generalized to incorporate the off-diagonal elements (see Appendix [C]). To emphasize this, we intentionally adopt double indices for the explanation of univariate Fisher information.

As n increases, the MLE $\tilde{\boldsymbol{\theta}}$ approaches an unknown value $\boldsymbol{\theta}_*$ that minimizes the Kullback–Leibler divergence (KLD) between $g(\cdot)$ and $f(\cdot|\boldsymbol{\theta})$ (White 1982; see Appendix [B]). Importantly, the Hessian of the KLD is the Fisher information and the KLD serves as a central connection between the fields of information theory and statistics. In the following, we describe our theory via Fisher information rather than KLD for the convenience of statistical description.

The empirically obtained $\widehat{\tilde{I}}_{gfii}$ in equation (1) can serve as estimates of $\tilde{I}_{gfii}$ in equation (2). In the definition of $\tilde{I}_{gfii}$ in equation (2), the expectation $\mathrm{E}_g[\cdot]$ is performed

TABLE 1. Summary of mathematical notation.

Notation	Equation	Definition
I_{gfii}		Fisher information of complete data Data were generated with unknown $g(\cdot)$ but analyzed with $f(\cdot\|\boldsymbol{\theta})$ Unidentifiable in general
$\tilde{I}_{gfii}$	(2)	Fisher information of gap-containing data Data were generated with unknown $g(\cdot)$ but analyzed with $f(\cdot\|\boldsymbol{\theta})$ Identifiable from given data
I_{ffii}	(3)	Fisher information of complete data Data were both generated and analyzed with known $f(\cdot\|\boldsymbol{\theta})$ Identifiable via simulation
$\tilde{I}_{ffii}$	(4)	Fisher information of gap-containing data Data were both generated and analyzed with known $f(\cdot\|\boldsymbol{\theta})$ Identifiable via simulation
G_i	(5)	Ratio of ESL with respect to PSL at branch i, G-Factor, $\tilde{I}_{ffii}/I_{ffii}$
M_i	(5)	Model misspecification factor at branch i, M-Factor, $\tilde{I}_{gfii}/\tilde{I}_{ffii}$
G	(6)	Weighted average of G_i, $\sum u_i G_i$
M	(7)	Weighted average of M_i, $\sum v_i M_i$
n		PSL (physical sequence length)
$\widehat{n}_e$	(16 – 18)	ESL (effective sequence length), defined as $n\widehat{G}$
$\widehat{n}_e^{(i,j)}$	(16)	sp-ESL (site-parameter-wise ESL), ESL for ith parameter at jth site
$\widehat{n}_e^{(i,\cdot)}$	(17)	p-ESL (parameter-wise ESL), ESL for ith parameter
$\widehat{n}_e^{(\cdot,j)}$	(18)	s-ESL (sitewise ESL), ESL at jth site

with respect to the true distribution $g(\cdot)$ because the $\tilde{x}^{(j)}$ in equation (1) were generated by $g(\cdot)$ rather than $f(\cdot|\boldsymbol{\theta})$. Paralleling the definitions of $\widehat{\tilde{I}}_{gfii}$ and $\tilde{I}_{gfii}$ for incomplete data in equations (1, 2), we define the Fisher information for complete data,

$$I_{gfii} := \mathrm{E}_g\left[-\frac{d^2}{d\theta_i^2}\log f(X|\boldsymbol{\theta}_*)\right],$$

where X is a random variable representing a complete sequence column. That is, I_{gfii} is the Fisher information of θ_i when there are no gaps.

The ratio $\tilde{I}_{gfii}/I_{gfii}$ is the relative amount of information from incomplete data when compared to complete data. This ratio is not identifiable. $\tilde{I}_{gfii}$ can be estimated from given incomplete data via equation (1), but I_{gfii} cannot be determined because complete (ungapped) data are unavailable. I_{gfii} cannot be estimated via simulation because the true process $g(\cdot)$ is unknown.

Using the adopted model $f(\cdot|\boldsymbol{\theta})$, I_{ffii} can be estimated via simulation,

$$I_{ffii} := \mathrm{E}_f\left[-\frac{d^2}{d\theta_i^2}\log f(Y|\boldsymbol{\theta}_*)\right]$$

$$\approx -\frac{1}{nm}\sum_{j=1}^{nm}\frac{d^2}{d\theta_i^2}\log f(y^{(j)}|\tilde{\boldsymbol{\theta}}) =: \widehat{I}_{ffii}, \quad (3)$$

where m is a large integer that can be arbitrarily determined based on the preference of estimation precision and computation time. For the convenience of calculation, we set the simulated data size to be exactly m times the original data size n.

Similar to the definition of $\tilde{I}_{gfii}$ in equation (2), $\tilde{I}_{ffii}$ and its estimate from simulated incomplete data can be defined,

$$\tilde{I}_{ffii} := \mathrm{E}_f\left[-\frac{d^2}{d\theta_i^2}\log f(\tilde{Y}|\boldsymbol{\theta}_*)\right]$$

$$\approx -\frac{1}{nm}\sum_{j=1}^{nm}\frac{d^2}{d\theta_i^2}\log f(\tilde{y}^{(j)}|\tilde{\boldsymbol{\theta}}) =: \widehat{\tilde{I}}_{ffii}, \quad (4)$$

where $\tilde{y}^{(j)}$ is generated by replacing some nucleotides in $y^{(j)}$ with gaps. The data size of equations (3, 4) is m times the original data size n, where the gaps in each column of $\tilde{x}$ are copied into m columns when generating $\tilde{y}$. Specifically, the gap pattern of $\tilde{x}^{(j)}(1\le j\le n)$ is copied into sites j, $\{n+j\}$, $\{2n+j\}$, $\cdots$, $\{(m-1)n+j\}$ of $\tilde{y}$. These sites are therefore correlated in terms of gap pattern and these sites are resampled simultaneously during our bootstrap procedure (see the following subsection).

Instead of the unidentifiable ratio $\tilde{I}_{gfii}/I_{gfii}$, consider the identifiable ratio $\tilde{I}_{gfii}/I_{ffii}$,

$$\frac{\tilde{I}_{gfii}}{I_{ffii}} = \frac{\tilde{I}_{ffii}}{I_{ffii}}\cdot\frac{\tilde{I}_{gfii}}{\tilde{I}_{ffii}} =: G_i\cdot M_i,$$

$$\approx \frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}}\cdot\frac{\widehat{\tilde{I}}_{gfii}}{\widehat{\tilde{I}}_{ffii}} =: \widehat{G}_i\cdot\widehat{M}_i \quad (5)$$

where G_i will be referred to as the "Gap factor" or "G-Factor" for the ith parameter of model $f(\cdot|\boldsymbol{\theta})$ and where M_i will be referred to as the "Model factor" or "M-Factor" for the ith parameter. The G-Factor G_i represents the proportion of information that remains after gaps are inserted. The range of the G-Factor is $0\le G_i\le 1$. Using identical simulated data in conjunction with gap copying as in equations (3, 4) is very likely to restrict the estimated ratio of $\widehat{G}_i$ to be equal to or less than 1. Furthermore, $\widehat{G}_i\equiv 0$ when only gaps are present and $\widehat{G}_i\equiv 1$ for complete data. Our empirical observation is that the G-Factor is robust for different choices of the model $f(\cdot|\boldsymbol{\theta})$ (see Results section). Because of this robustness, we expect that the G-Factor will be similar to the unidentifiable ratio $\tilde{I}_{gfii}/I_{gfii}$ when some care is taken in choosing $f(\cdot|\boldsymbol{\theta})$.

The M-Factor M_i represents the "goodness of fit" of data to the model. When the adopted model is correct ($f=g$), $M_i\equiv 1$. Whereas $\widehat{G}_i$ is very unlikely to exceed 1, $\widehat{M}_i$ varies around 1 when the adopted model is correct.

If $\widehat{M}_i$ significantly deviates from 1, this indicates model misspecification.

After the "parameter-wise' G-Factors (i.e., the G_i's) and M-Factors (i.e., the M_i's) are estimated, an "overall" G-Factor G and an "overall" M-Factor M for the phylogeny can be inferred via weighted averages of parameter-wise values,

$$G := \sum_{i=1}^{d} u_i G_i \approx \sum_{i=1}^{d} \widehat{u}_i \widehat{G}_i =: \widehat{G} \tag{6}$$

$$M := \sum_{i=1}^{d} v_i M_i \approx \sum_{i=1}^{d} \widehat{v}_i \widehat{M}_i =: \widehat{M}, \tag{7}$$

where the u_i and v_i represent weights. As explained in the Appendix [B], these weights are

$$u_i := \frac{I_{ffii}^2}{\sum_i I_{ffii}^2} \approx \frac{\widehat{I}_{ffii}^2}{\sum_i \widehat{I}_{ffii}^2} =: \widehat{u}_i \tag{8}$$

$$v_i := \frac{I_{ffii}\tilde{I}_{ffii}}{\sum_i I_{ffii}\tilde{I}_{ffii}} \approx \frac{\widehat{I}_{ffii}\widehat{\tilde{I}}_{ffii}}{\sum_i \widehat{I}_{ffii}\widehat{\tilde{I}}_{ffii}} =: \widehat{v}_i. \tag{9}$$

Bootstrap Procedure to Measure Uncertainty of G-Factors and M-Factors

To assess the uncertainty of the overall G-Factor $\widehat{G}$ (equation (6)) and the overall M-Factor $\widehat{M}$ (equation (7)), we develop a bootstrap procedure. A hierarchical structure yields $\widehat{G}$ and $\widehat{M}$ via first obtaining the MLE ($\tilde{\boldsymbol{\theta}}$) from the original sequence data as in equation (1) and then generating extremely long sequence data by using $\tilde{\boldsymbol{\theta}}$ as in equations (3, 4). Therefore, our bootstrap resampling procedure needs to reflect this hierarchical structure. Our resampling approach is similar to the resampling of estimated log-likelihood (RELL; Kishino and Hasegawa 1989) approach. Instead of resampling sequence columns and then re-optimizing MLEs as would be done with a full bootstrap procedure, our RELL-like procedure calculates sitewise second derivatives once and saves computation by resampling these second derivatives with replacement. When the sample size n is small, the RELL-like procedure is prone to poor performance and can result in dissatisfying estimates such as negative G-Factors and M-Factors. In this case, the more computationally demanding full bootstrap is required. When n is large enough, the RELL-like procedure is asymptotically similar to the full bootstrap.

Following convention, we employ the "$*$" superscript to indicate a bootstrap-resampled random quantity. For notational convenience, the negative second derivatives at the rth site for $\widehat{\tilde{I}}_{gfii}$, $\widehat{\tilde{I}}_{ffii}$, and $\widehat{I}_{ffii}$ are respectively denoted x_r, y_r, and z_r. In a similar way, the sitewise resampled negative second derivatives for $\widehat{\tilde{I}}^*_{gfii}$, $\widehat{\tilde{I}}^*_{ffii}$, and $\widehat{I}^*_{ffii}$ are respectively denoted x_r^*, y_r^*, and z_r^*. With this notation,

$$\widehat{\tilde{I}}_{gfii} = \frac{1}{n}\sum_{r=1}^{n} x_r, \quad \widehat{\tilde{I}}^*_{gfii} = \frac{1}{n}\sum_{r=1}^{n} x_r^*$$

$$\widehat{\tilde{I}}_{ffii} = \frac{1}{nm}\sum_{r=1}^{nm} y_r, \quad \widehat{\tilde{I}}^*_{ffii} = \frac{1}{nm}\sum_{r=1}^{nm} y_r^*$$

$$\widehat{I}_{ffii} = \frac{1}{nm}\sum_{i=r}^{nm} z_r, \quad \widehat{I}^*_{ffii} = \frac{1}{nm}\sum_{i=r}^{nm} z_r^*. \tag{10}$$

Resampling x_r^* follows a simple and conventional RELL-like procedure with

$$x_r^* := x_{p(r)} \ (r=1,\ldots,n), \tag{11}$$

where $p(r)$ is a uniformly and randomly selected integer from 1 to n. After obtaining the x_r^*, we reuse their $p(r)$ indices for generating the y_r^* and z_r^*. Because the gap pattern of x_j is copied to columns j, $\{n+j\}$, $\{2n+j\}$, …, $\{(m-1)n+j\}$ of y, we mimic this dependency during bootstrapping. To do this, we use the stored $p(r)$ indices and define the resampled $y^*_{(\cdot)}$ as

$$y^*_{r+(k-1)n} := y_{p(r)+(k-1)n} + \left\{\widehat{\tilde{I}}^*_{gfii} - \widehat{\tilde{I}}_{gfii}\right\} \ (k=1,\ldots,m), \tag{12}$$

where $r=1,\ldots,n$. The translocation factor $\left\{\widehat{\tilde{I}}^*_{gfii} - \widehat{\tilde{I}}_{gfii}\right\}$ of equation (12) is necessary because we generate $\widehat{\tilde{I}}_{ffii}$ with $\tilde{\boldsymbol{\theta}}$. Therefore, $\widehat{\tilde{I}}_{ffii}$ and $\widehat{\tilde{I}}_{gfii}$ are correlated and this correlation needs to be preserved for generating $\widehat{\tilde{I}}^*_{ffii}$ and $\widehat{\tilde{I}}^*_{gfii}$ (see Supplementary material [A] available on Dryad at http://dx.doi.org/10.5061/dryad.zs7h44j9f).

Similar to the y_r translocations, we translocate the z_r to resample z_r^*. Using the position of $\widehat{\tilde{I}}^*_{ffii}$ and the stored $p(r)$ indices, we define $z^*_{(\cdot)}$ as

$$z^*_{r+(k-1)n} := z_{p(r)+(k-1)n} + \left\{\widehat{\tilde{I}}^*_{ffii}\widehat{I}_{ffii}/\widehat{\tilde{I}}_{ffii} - \widehat{I}_{ffii}\right\} \ (k=1,\ldots,m). \tag{13}$$

The translocation factor $\left\{\widehat{\tilde{I}}^*_{ffii}\widehat{I}_{ffii}/\widehat{\tilde{I}}_{ffii} - \widehat{I}_{ffii}\right\}$ of equation (13) is necessary because $\widehat{\tilde{I}}_{ffii}$ and $\widehat{I}_{ffii}$ are correlated and this correlation needs to be preserved for generating $\widehat{\tilde{I}}^*_{ffii}$ and $\widehat{I}^*_{ffii}$ (see Supplementary material [A] available on Dryad).

By applying equations (11, 12, 13), we generate the $x^*_{(\cdot)}$, $y^*_{(\cdot)}$, and $z^*_{(\cdot)}$. From these, we derive the overall G-Factor $\widehat{G}^*$ and the overall M-Factor $\widehat{M}^*$. Via iteration of bootstrapping, we thereby approximate the distribution of $\widehat{G}^*$ and $\widehat{M}^*$. These distributions can be used to estimate the variances of $\widehat{G}$ and $\widehat{M}$. The distribution of $\widehat{M}^*$ can be further used to test model adequacy (see next subsection).

Model Adequacy: Hypothesis Test of M-Factors

In equation (5), $M=1$ if $g=f$. Thus, large absolute values of $\{\widehat{M}-1\}$ suggest model misspecification. We develop a hypothesis test in which the null hypothesis is $g=f$ and the test statistic is $\left|\widehat{M}-1\right|$. Here, we describe a test for the overall M-Factor that has a null hypothesis of $M=1$, but the approach also can be applied to parameter-wise M-Factors (i.e., M_i's).

To test for a deviation of $\widehat{M}$ from 1, we use the following approximation of distributions,

$$\left\{\widehat{M}-1\right\} \overset{d}{\approx} \left\{\widehat{M}^{*(i)}-\overline{\widehat{M}^{*}}\right\}, \tag{14}$$

where $\widehat{M}^{*(i)}$ is the overall M-Factor estimate from the ith resampled data, and $\overline{\widehat{M}^{*}}$ is the average $\widehat{M}^{*(i)}$. Following the guideline of "bootstrap centering" (Hall and Wilson 1991), we measure the significance of $\left|\widehat{M}-1\right|$ via the distribution of $\left|\widehat{M}^{*(i)}-\overline{\widehat{M}^{*}}\right|$. For our two-tailed test of $\widehat{M}=1$, the P value of $\left|\widehat{M}-1\right|$ is estimated as

$$\frac{1}{B}\sum_{i=1}^{B} I\left(\left|\widehat{M}^{*(i)}-\overline{\widehat{M}^{*}}\right| > \left|\widehat{M}-1\right|\right), \tag{15}$$

where B is the number of bootstrap samples and the indicator variable $I(\cdot)$ is 1 if the condition within the parentheses is satisfied and is 0 otherwise.

ESL versus PSL

We refer to the number of alignment columns n in the observed data as the "Physical Sequence Length (PSL)". The information about parameter i in a simulated incomplete data set of size n would be $n\tilde{I}_{ffii}$. We let $\widehat{n}_e^{(i,\cdot)}$ be the number of simulated alignment columns needed in a complete (gapless) data set to have the same amount of information about parameter i as the amount of information $n\tilde{I}_{ffii}$ in the simulated data with gaps, $n\tilde{I}_{ffii}=n_e^{(i,\cdot)}I_{ffii}$. This leads to $n_e^{(i,\cdot)}=nG_i$. We refer to $n_e^{(i,\cdot)}$ as the "parameter-wise Effective Sequence Length" ("p-ESL") of incomplete data with respect to parameter i.

Paralleling the derivation of the overall G-Factor (G) from individual G-Factors (i.e., G_i's) in equation (6), an overall ESL can be derived from individual p-ESL terms. The overall ESL (n_e) is $n_e:=nG$ where G is given in equation (6). For a given incomplete sequence data set, PSL can be directly observed whereas ESL is not directly observable but can be estimated via Fisher information and PSL.

We can also consider "sitewise ESL (s-ESL)" and "site-parameter-wise ESL (sp-ESL)." The basic idea is to separate the total ESL into components for each site or into components that represent each combination of site and parameter. To simplify notation, define $l''^{(j)}_{ii}$ as the second derivative at the jth site of equation (1),

$$l''^{(j)}_{ii} := \frac{d^2}{d\theta_i^2}\log f(\tilde{x}^{(j)}|\tilde{\boldsymbol{\theta}}).$$

By using equations (5–9), we represent $\widehat{n}_e$ as

$$\widehat{n}_e := n\widehat{G} = n\sum_{i=1}^{d}\widehat{u}_i\widehat{G}_i$$

$$= \sum_{i=1}^{d}\left[\widehat{u}_i\sum_{j=1}^{n}\left\{\frac{-l''^{(j)}_{ii}}{\widehat{I}_{ffii}\widehat{M}_i}\right\}\right] =: \sum_{i=1}^{d}\left[\widehat{u}_i\sum_{j=1}^{n}\left\{\widehat{n}_e^{(i,j)}\right\}\right] \tag{16}$$

$$= \sum_{i=1}^{d}\left[\widehat{u}_i\left\{\sum_{j=1}^{n}\frac{-l''^{(j)}_{ii}}{\widehat{I}_{ffii}\widehat{M}_i}\right\}\right] =: \sum_{i=1}^{d}\left[\widehat{u}_i\left\{\widehat{n}_e^{(i,\cdot)}\right\}\right] \tag{17}$$

$$= \sum_{j=1}^{n}\left\{\sum_{i=1}^{d}\frac{-\widehat{u}_i l''^{(j)}_{ii}}{\widehat{I}_{ffii}\widehat{M}_i}\right\} = \sum_{j=1}^{n}\left\{\sum_{i=1}^{d}\widehat{u}_i\widehat{n}_e^{(i,j)}\right\} =: \sum_{j=1}^{n}\left\{\widehat{n}_e^{(\cdot,j)}\right\}, \tag{18}$$

where $\widehat{n}_e^{(i,j)}$, $\widehat{n}_e^{(i,\cdot)}$, and $\widehat{n}_e^{(\cdot,j)}$ will be respectively referred to as estimators of the site-parameter-wise ESL (sp-ESL), parameter-wise ESL (p-ESL), and sitewise ESL (s-ESL). The sp-ESL, p-ESL, and s-ESL, respectively represent informativeness of each parameter at each sequence column, each parameter and each sequence column. Although $n\widehat{G}$ is positive, an sp-ESL or s-ESL can be negative if it conflicts with the rest of the data. When sp-ESL and s-ESL coincide with the information in the overall data set, their values will be positive (and potentially large). We note that the p-ESL $\widehat{n}_e^{(i,\cdot)}$ of equation (17) is the simple summation over sites of the sp-ESL $\widehat{n}_e^{(i,j)}$ of equation (16) but that the s-ESL $\widehat{n}_e^{(\cdot,j)}$ of equation (18) is the weighted average of $\widehat{n}_e^{(i,j)}$.

Results

We studied our approach with simulations and applied it to both DNA and protein sequence data. As discussed in more detail in the Appendix [C], our implementation only considers the diagonal elements of the Fisher Information matrix and branch lengths are the only parameters represented in our Fisher Information estimates. We demonstrate with our empirical data analyses that model violations can be detected via their effects on Fisher information related to branch-length parameter estimates.

Simulation Studies

We performed three-step simulations to evaluate our procedure for estimating the G-Factor (G) and M-Factor (M):

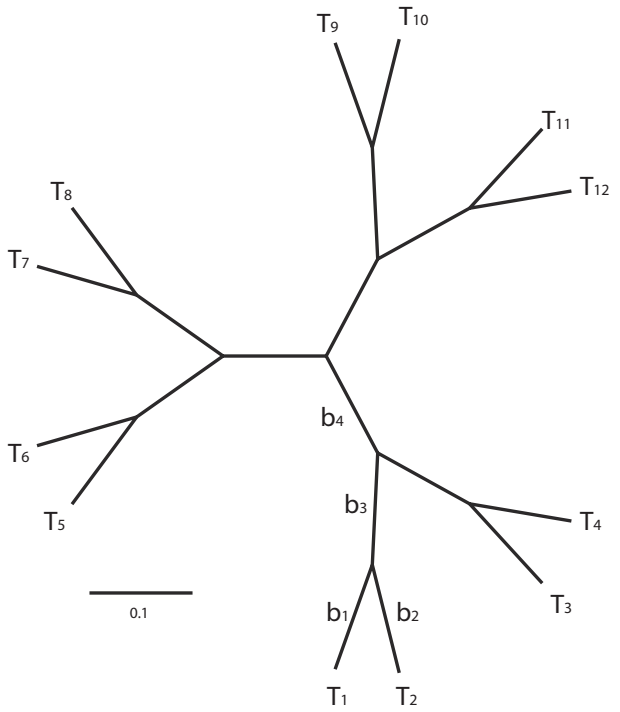

FIGURE 1. Phylogeny of 12 taxa for simulation. All branch lengths are 0.1 substitutions per nucleotide site.

Step 1: Sequence generation
Figure 1 shows a phylogeny of 12 taxa. All branch lengths on this phylogeny are 0.1 nucleotide substitutions per site. The PSL (n) of the simulated original data is 1000. The odd-numbered taxa in Figure 1 have exclusively gaps in their final $n/2=500$ sites. The ungapped nucleotide data were generated with the GTR+Gamma model (Tavaré 1986; Yang 1994a, 1994b) and with nucleotide frequencies of A, C, G, and T respectively being 0.2, 0.3, 0.2, and 0.3. The rate matrix parameters were set to 0.7 (A↔C), 0.8 (A↔T), 10.0 (A↔G), 5.0 (C↔T), 0.9 (C↔G), and 1.0 (T↔G). A discrete-gamma model with five categories and $\alpha=1.0$ (Yang 1994b) was used to incorporate rate heterogeneity among sites.

Step 2: Estimation
Each simulated data set was analyzed with the GTR+Gamma, TN93+Gamma (Tamura and Nei 1993), and JC+Gamma (Jukes and Cantor 1969) models. Using each adopted model, we generated an extremely long sequence data set with $m=100$ (see equations (3, 4)) and then estimated the G_i's, M_i's, G, and M.

Step 3: Uncertainty assessment
After estimating the G-Factors and M-Factors, we applied the RELL-like bootstrap procedure with B = 500 replicates. These replicates allowed precise estimation of the variability of G-Factors and M-Factors as well as testing of $M=1$.

Repetition: Steps 1–3 were repeated 500 times to yield 500 P values for the model adequacy test of $M=1$ as well as 500 sets of $\widehat{M}$, $\widehat{G}$, $\widehat{M}_i$'s, and $\widehat{G}_i$'s.

The simulated alignments and the underlying phylogeny (see Fig. 1) were designed to have three-fold symmetry between the taxon subsets $T_1 - T_4$, $T_5 - T_8$, and $T_9 - T_{12}$. Because of the symmetry and because all odd-numbered taxa have gaps in their final 500 sites, we focus only on the branches $b_1 - b_4$ (see Fig. 1). For each of these branches, we obtained a $\widehat{G}_i$ value for each of the 500 simulated data sets.

In the analysis with the GTR+Gamma model, the average and standard deviations of $\widehat{G}_i$ for $b_1 - b_4$ are respectively 0.500 ($\pm 3.00\times10^{-4}$), 0.673 ($\pm 1.35\times10^{-3}$), 0.742 ($\pm 1.66\times10^{-3}$), and 0.882 ($\pm 6.92\times10^{-4}$). Because T_1 has gaps in half of its sites, $\overline{\widehat{G}}_1=0.500$ is consistent with the simulation setting. Although T_2 is complete (i.e., ungapped), $\overline{\widehat{G}}_2$ is less than 1 because of the effects of gaps in other sequences. This illustrates the point made in the Introduction section that a simple gap proportion is a flawed measure for information loss. For the internal branch b_3, the average G-Factor $\overline{\widehat{G}}_3$ is less than 1 but greater than $\overline{\widehat{G}}_2$. This suggests gaps have a stronger effect on information in terminal than interior branches, presumably because information about interior branches is more evenly distributed among sequences. The average G-Factor $\overline{\widehat{G}}_4$ is presumably greater than $\overline{\widehat{G}}_3$ because internal branch b_4 is farther from the gaps of T_1 than is internal branch b_3. Among the 500 simulated data sets, the average overall $\widehat{G}$ is 0.631 ($\pm 6.68\times10^{-4}$) which is less than the 0.75 proportion of alignment positions that are ungapped.

For each of the 500 simulated data sets and for each of the three substitution models, we tested model adequacy via a null hypothesis of $M=1$. If the null hypothesis is true and the model adequacy test functions as intended, the distribution of $\widehat{M}-1$ should be well approximated by $\widehat{M}^*-\overline{\widehat{M}^*}$ so that $P(\widehat{M}^*-\overline{\widehat{M}^*}<\widehat{M}-1.0)$ has a uniform distribution between 0 and 1. First, we explored the case where the null hypothesis was true because the GTR+Gamma model was used for both simulating and analyzing the data. For each of the simulated data sets, $\widehat{M}$ was estimated and then 500 bootstrap replicates were employed to approximate $P(\widehat{M}^*-\overline{\widehat{M}^*}<\widehat{M}-1.0)$. The concentration around 0.5 in Figure 2a differs from a uniform distribution and indicates that the $\{\widehat{M}^*-\overline{\widehat{M}^*}\}$ test statistic underestimates the tails of the $\{\widehat{M}-1.0\}$ distribution when the null hypothesis is true. This suggests that our model adequacy test is conservative. For a significance level of 0.05, our model adequacy test rejects the null hypothesis of $M=1$ for 0.022 (11 of 500) simulated data sets.

This conservative nature of the model adequacy test is presumably because it relies upon "plug-in" parameter estimates rather than actual values of parameters when approximating the test statistic distribution. This "plug-in" nature of our approach can produce conservative or anti-conservative hypothesis tests (Robins et al. 2000).

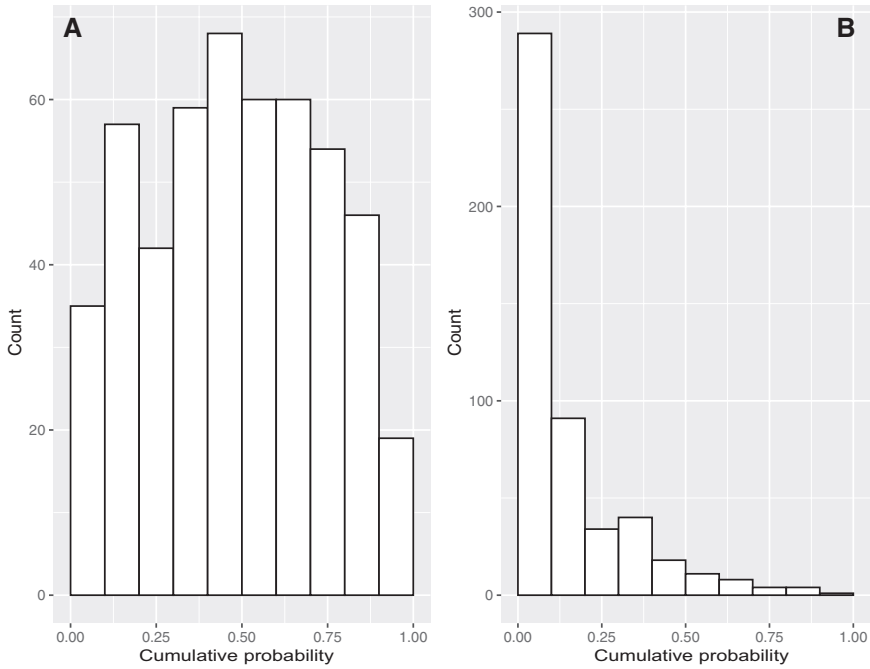

FIGURE 2. Histogram of the bootstrap approximation to the distribution of cumulative probabilities of the test statistic $\widehat{M}-1.0$. Sequence data were simulated with GTR+Gamma and the bootstrap approximation was applied to estimate $P(\widehat{M}^*-\overline{\widehat{M}^*}<\widehat{M}-1.0)$. a) Data were analyzed with GTR+Gamma. b) Data were analyzed with JC+Gamma.

However, the conservativeness will decrease as n increases. Although a derivation is omitted, the difference between second derivatives at the true $\boldsymbol{\theta}$ and at the MLE $\tilde{\boldsymbol{\theta}}$ is bounded in probability with the order of $n^{-1/2}$ (for a definition of "bounded in probability," see Bishop et al. 2007). Because the variances of the original and bootstrapped M-Factor estimates are asymptotically equal (see Supplementary material [B] available on Dryad), the conservative nature of our model adequacy test should be diminished when n is large.

When the null hypothesis is wrong, $P(\widehat{M}^*-\overline{\widehat{M}^*}<\widehat{M}-1.0)$ will be concentrated around 0 or 1 if the model adequacy test has power to reject the null hypothesis. For the case where JC+Gamma was the adopted model but the truth was GTR+Gamma, Figure 2b summarizes the histogram of $P(\widehat{M}^*-\overline{\widehat{M}^*}<\widehat{M}-1.0)$ from 500 simulated data sets and 500 bootstrap replicates per simulated data set. For this situation where the null hypothesis should be rejected, the null was rejected at a significance level of 0.05 in a proportion 0.326 of cases (163 out of 500). Whereas the model adequacy test often rejected the null when the JC+Gamma model was used, it had low power when the TN93+Gamma model was assumed. Specifically, the null was rejected at a significance level of 0.05 in a proportion 0.022 of cases (11 out of 500). Our model adequacy test with PSL = 1000 did not distinguish the TN93+Gamma and GTR+Gamma models, but conventional information criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978) as well as the likelihood ratio test do better. For example, the likelihood ratio test rejected the TN93+Gamma model at a significance level of 0.05 in a proportion 0.284 (142 of 500) of cases when the truth was GTR+Gamma.

Empirical Data Analysis 1: Eukaryote Protein Sequences

We analyzed amino acid sequences from 55 taxa, including 42 eukaryotes (Derelle and Lang 2012). Aligned sequences were obtained from TreeBASE (Piel et al. 2009; Vos et al. 2012; TreeBASE Matrix ID:M11012). For this data set, the PSL is 11,500 sites and the mean proportion of nongap characters per taxon is 0.8045 ($\pm$0.0237).

We inferred the maximum likelihood tree with the LG+Gamma model (Le and Gascuel 2008) by using RAxML software version 8.24 (Stamatakis 2014) along with a four-category discretized gamma distribution to incorporate rate heterogeneity among sites (Yang 1994b). For the estimation of G-Factors and M-Factors, we applied four amino acid models: LG+Gamma, WAG+Gamma (Whelan and Goldman 2001), JTT+Gamma (Jones et al. 1992), and Dayhoff+Gamma (Dayhoff et al. 1978). Assuming the maximum likelihood tree topology and these four substitution models, the maximum log-likelihood scores were respectively $-$490,769.53, $-$494,591.39, $-$500,473.62, and $-$500,900.11.

Overall G-Factors: For the LG+Gamma model, the overall G-Factor ($\widehat{G}$) is 0.8154 ($\pm 4.7\times10^{-3}$) and the ESL of the aligned data is 9377 ($\approx$0.8154 $\times$11,500). In this case, the G-Factor is slightly greater than the proportion 0.8045 of ungapped positions in the alignment. The overall G-Factor was robust to the amino acid replacement

model. The overall G-Factors for the WAG+Gamma, JTT+Gamma, and Dayhoff+Gamma models are respectively 0.8187 ($\pm 5.1 \times 10^{-3}$), 0.7902 ($\pm 4.6 \times 10^{-3}$), and 0.8152 ($\pm 4.9 \times 10^{-3}$).

Parameter-wise G-Factors: As with the overall G-Factors, our estimates of the ratios $\tilde{I}_{ffii}/I_{ffii}$ (i.e., the parameter-wise G-Factors G_i) were robust to model choice (data not shown). Figure 3 maps the parameter-wise G-Factors that were estimated under the LG+Gamma amino acid model onto branches of the phylogeny. Consistent with the simulation results, the $\widehat{G}_i$ estimates show a gradual change over branches. Some internal branches have high $\widehat{G}_i$ and these seem to be the ones that are far from terminal taxa with many gaps.

Overall M-Factors: Because the log-likelihood score with the LG+Gamma model is the highest among the four models that we explored, our discussion concentrates on results from it. The overall M-Factor ($\widehat{M}$) is 0.7595 ($\pm 5.5 \times 10^{-3}$) and is significantly different from 1 (P-value $\ll 0.01$). This implies that the adopted LG+Gamma model is significantly different from the unknown data-generating mechanism. The M-Factors for the WAG+Gamma, JTT+Gamma, and Dayhoff+Gamma models were respectively 0.7566 ($\pm 5.2 \times 10^{-3}$), 0.7492 ($\pm 5.6 \times 10^{-3}$), and 0.7714 ($\pm 5.1 \times 10^{-3}$). While these M-Factor estimates vary, all M-Factor estimates are approximately equally far from 1 and are significantly different from it. These M-Factor estimates are therefore consistent with the possibility that the difference between the true data-generating mechanism and any of these models is far bigger than the differences between these models.

Parameter-wise M-Factors: Even for individual branches on the phylogeny, the LG+Gamma model does not fit well. Out of 107 branches, 100 show significant rejection of $M_i = 1$ (two-tailed P value < 0.05) when assuming the LG+Gamma model. To confirm that the large number of significant parameter-wise M-Factors are not artifacts, we performed a simple simulation. Using the maximum likelihood phylogeny with the LG+Gamma model, we simulated a data set with a PSL that matches the 11,500 of the actual data and with a gap pattern that is identical to the original data. We then estimated G-Factors and M-Factors with the aforementioned amino acid replacement models. Consistent with our finding from the original data set, the G-Factors estimated from this simulated data set are robust to model choice (data not shown). However, the estimated M-Factors from the simulated data show a different pattern relative to the original data. When the correct model (LG+Gamma) is adopted for analyzing the simulated data, the parameter-wise M-Factors are distributed around 1 with a mean of 0.999 and a standard deviation of 0.022. In contrast, the M-Factors from the actual data tend to be substantially less than 1. Among the 107 branches on the eukaryotic tree, all yield M-Factors that are less than 1. These estimates from the actual data have a maximum of 0.947, a minimum of 0.626, a mean of 0.873, and a standard deviation of 0.0653. The contrast between the results from simulated and original data implies that the substantially smaller M-Factors from the original data are not artifacts.

Evaluation of filtering scheme: By using the Gblocks filtering program (Castresana 2000) with options that were slight modifications of the defaults, Derelle and Lang (2012) removed 2759 gap-containing sequence columns from the original sequence data. We measured the sitewise ESL (s-ESL) of positions that were removed by Derelle and Lang (2012). A high value of s-ESL implies consistency of the column with the reconstructed phylogeny. About 6% (167 of 2759) of the removed sites had an s-ESL that exceeded 5. In contrast, only about 2% (180 of 8741) retained sites exceeded 5. This suggests that Gblocks tends to remove sites with high information content. Figure 4 displays the s-ESL distributions among removed and retained sites. It has been suggested that Gblocks tends to remove too many gap-rich columns from data sets (e.g., Tan et al. 2015). The relatively high frequency of high s-ESL values among removed sites is consistent with the possibility that the removed sites actually contain substantial phylogenetic information.

The s-ESL distributions in Figure 4 are skewed because the distribution of sitewise second derivatives is highly skewed. Our experience is that a large proportion of sitewise Fisher information values (i.e., negative second derivatives) are distributed near zero and some are even negative. A relatively small proportion of sites show large positive values of sitewise Fisher information. For this reason, the resampled data with our RELL-like procedure often will not closely approximate the original skewed distribution for small sequence lengths n. This unsatisfactory behavior when n is small is characteristic of RELL-like procedures.

Empirical Data Analysis 2: Ray-finned Fish DNA Sequences

To illustrate the approach with aligned nucleotide sequences, we used a ray-finned fish data set (Li et al. 2008) in TreeBASE (Piel et al. 2009; Vos et al. 2012; TreeBASE Study ID S2045). It represents 52 ray-finned fish and 4 outgroup taxa. The PSL of the original aligned data is 7995 nucleotide sites and the mean proportion of nongap characters per taxon is 0.8155 (±0.0216). Although there is compelling justification for analyzing these fish data with more parameter-rich modeling frameworks (Li et al. 2008; Seo and Thorne 2018), we contrast three simple substitution models (GTR+Gamma, TN93+Gamma, and JC+Gamma) for the sake of illustration. With the GTR+Gamma model, the RAxML software (Stamatakis 2014) finds the topology depicted in Figure 5 that is used below.

G-Factors and M-Factors: As with the analysis of amino acid sequences, we observed robustness of the parameter-wise G-Factors ($\widehat{G}_i$) among the nucleotide models (data not shown). For the GTR+Gamma model,

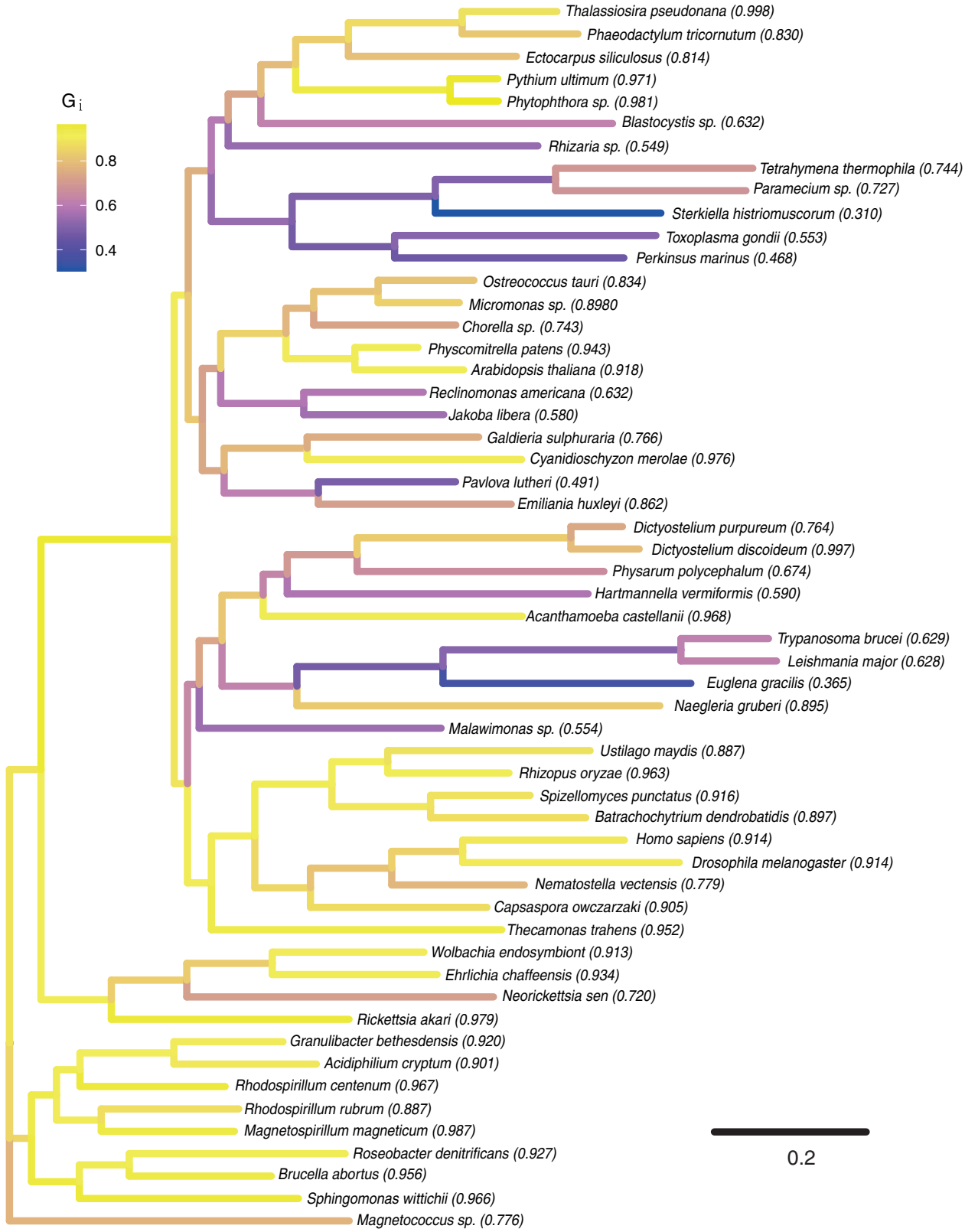

FIGURE 3. G-Factor (G_i) estimates for the eukaryote data with the LG+Gamma model. Each branch is shaded according to its G_i estimate. The proportion of nongaps for each taxon is shown in parentheses following its name.

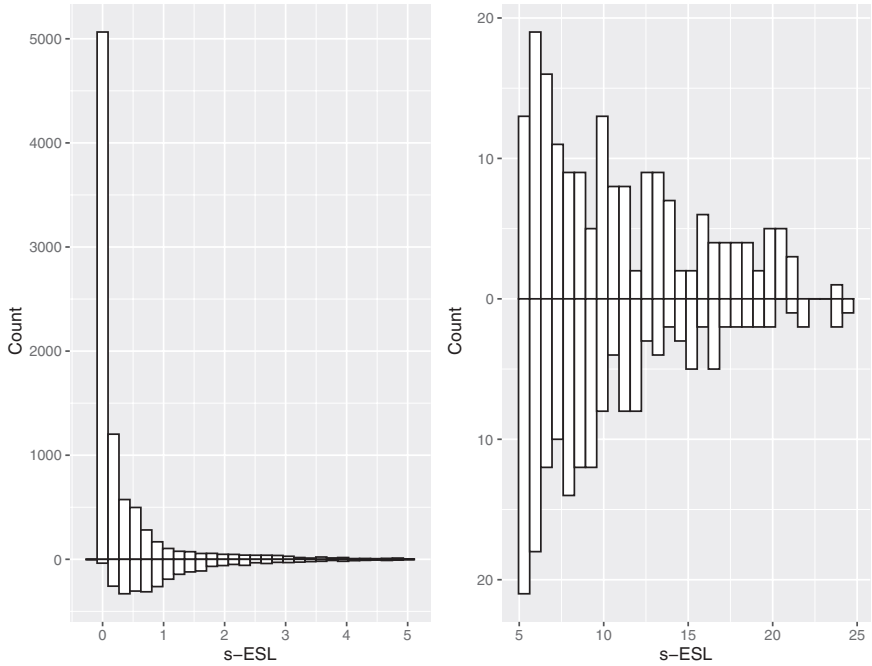

FIGURE 4. The s-ESL distributions among sites from the eukaryotic data. The left histograms show sites with s-ESL values that are <5. The right histograms show sites with s-ESL values that are ≥ 5. For both the left and the right, the histograms on the bottom show s-ESL values for removed sites whereas the histograms on the top show s-ESL values of retained sites. The upper left, upper right, lower left, and lower right histograms respectively summarize 8561, 180, 2592, and 167 sites.

the overall G-Factor ($\widehat{G}$) is 0.8547 ($\pm 3.2 \times 10^{-3}$). The overall G-Factors were again quite robust to model. The overall G-Factor was 0.8522 ($\pm 3.3 \times 10^{-3}$) for the TN93+Gamma model and 0.8494 ($\pm 3.4 \times 10^{-3}$) for the JC+Gamma model.

Relative to the parameter-wise G-Factor estimates, the parameter-wise M-Factor ($\widehat{M}_i$) estimates from the ray-finned fish data show more variability among models. The overall M-Factor is 0.8667 ($\pm 6.7 \times 10^{-3}$) for the GTR+Gamma model. This is significantly different from 1 (P value $\ll 0.01$). The overall M-Factor was 0.8711 ($\pm 6.7 \times 10^{-3}$) for the TN93+Gamma model and 0.8971 ($\pm 4.9 \times 10^{-3}$) for JC+Gamma. As was the pattern when applying amino acid replacement models to the eukaryote protein data, these M-Factor estimates are consistent with the possibility that the difference between the true data-generating mechanism and these substitution models is far greater than the differences between these substitution models. Others have also concluded that widely used models of sequence change provide poor fits to real data. For example, Chen et al. (2019) introduced a model adequacy test that strongly rejected the GTR+Gamma+"Invariant Sites" model for most of the data sets to which it was applied.

Interpretation of ESL: Because the PSL is 7995, the ESL is about 6833 ($\approx \widehat{G} \times 7995$). As shown in equations (16–18), the total ESL ($\widehat{n}_e := n\widehat{G}$) can be expressed as sums of terms that are sp-ESL (site-parameter-wise ESL), s-ESL (sitewise ESL), or p-ESL (parameter-wise ESL). Because the number of sp-ESL terms is the product of the numbers of s-ESL and p-ESL terms, individual sp-ESL terms are particularly subject to stochastic error. Whereas p-ESL values will always be positive, the s-ESL and sp-ESL can have negative values.

The sp-ESL values are influenced by the estimated length of the branch to which they correspond and also by the strength of evidence that the site changed or did not change on the branch. Large positive sp-ESL values occur when there is strong evidence that a site changed on a short branch. When a site is unlikely to have changed on a short branch, sp-ESL values will tend to be slightly below zero. When evidence is weak regarding whether a site did or did not change on a branch, sp-ESL values will be close to 0. Weak evidence can stem from a combination of reasons including long branches, branches that are far from any tips of the tree, an abundance of gaps at a site, and changes at the site at multiple branches that are sufficiently nearby on the tree as to make the most parsimonious mapping of the site unreliable.

With our implementation, Fisher information concerning branch lengths is considered but Fisher information concerning rate and nucleotide frequency parameters is not. Therefore, a site must have at least two meaningful molecular characters to have a nonzero s-ESL value. This is because two characters correspond to a path in the phylogeny and thereby contain information for branch length estimation.

It is helpful to compare the ray-finned fish phylogeny of Figure 5 to the sites depicted in Table 2. Site #5242 of Group A in Table 2 has the highest s-ESL value

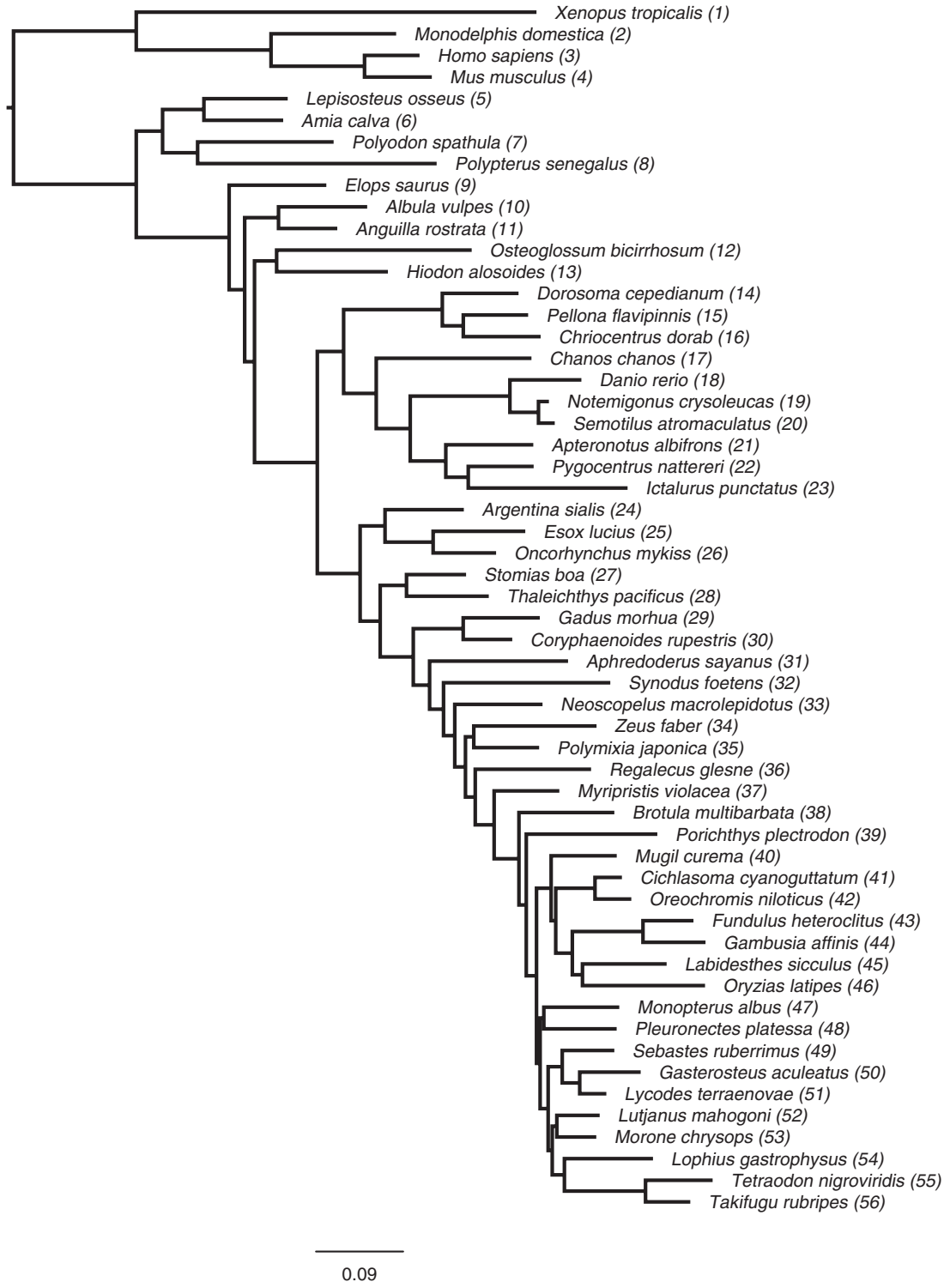

FIGURE 5. Maximum likelihood phylogeny of ray-finned fish taxa. Each taxon name is followed in the parentheses by the taxon number that is used in Table 2.

(168.4) among all 7995 sites. This site appears to have experienced a change from G to A on the shortest branch of the phylogeny (i.e., the branch that ends at the most recent common ancestor of Taxa #47–56 in Fig. 5). In fact, four of the five sites in Group A of Table 2 appear to have experienced a change on this shortest branch. These changes lead to large positive sp-ESL values that have a substantial influence on membership in Group A.

TABLE 2. Sitewise ESL (s-ESL) values of the ray-finned fish group. Taxon numbering is compatible with that of Figure 5. This table displays sites with the greatest (Group A) and smallest (Group F) s-ESL values. It also displays sites with relatively high s-ESL values (Group B), sites that yield the highest sp-ESL values for the moderately long branch on Figure 5 that ends with the most recent common ancestor of Taxa #9-56 (Group C), sites with s-ESL values that are slightly below zero (Group D), and sites with relatively low s-ESL values (Group E)

Group	Site number	Taxon number 1 2 3 4 5 12345678901234567890123456789012345678901234567890123456	Sitewise ESL (s-ESL)
A	5242	AGGG-A-G-----GGG-GGG--GGGGG-GGGGAGGGGGGGGGGG-GAAAAAGA-AG	168.4
	2934	GTTT-ACTC-C-C-G-CGGGCCCCCCCCGCCCCCCCCCCCCCCCCG-AGGAGGGTGA	165.0
	6516	AAAAGGTTG--GA-AGGGAAGGT-GG--G----G-GGTGGTTGG-GCCCCGCCTGG	153.0
	6470	CCCCGCGGG--GG-GGGGGGGAG-GG--G----G-GGGGGGGGA-GGGGGGCCCCC	109.9
	1431	----GA--GGG-GGGGGGGG-AAGGGGGGGGGGGGGGGGGAGGGAGG-AAAAGAAAAA	109.7
B	2265	TCTCTTTTC---TTCT-CTT-TGATA----CT-TC-CC-C-AC-CTTCTTTTT-CC	9.697
	7779	TCCCGGTCG--AGTGGGGGGTGGGGGCGG-GGGA-GGGGGAAGGGAAG-AGAAGAA	9.387
	7503	CCTTCCCTC--CCCGCCCCCCCCCCCCCC-CCGC-TCCGGAAAGACAG-GGGGCGT	9.079
	2430	GGAGTTCCG---GGGG-GGT-GGGCG----GG-CG-CG-G-GT-GTGACCCCG-TT	8.863
	5325	GATA-G-A-----AGG-GGG--GAGGG-GGGAAGAGGAAGGGAG-GGGAAAAA-GG	8.745
C	7109	CCCCCCCCG--GGGGGGGGGGGGGGGGGG-GGGG-GGGGGGGGGGGGG-GGGGGGG	0.002324
	2743	CCCC-CCCA-A-A-A-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	0.002320
	6779	-TTTTTTTG--GG-GGGGGGGGG-GG--G----G-GGGGGGGGG-GGGGGGGGGGG	0.001843
	3275	AAAA-A-AT--TT---TTTTT-TT-TTT--TTT---TTTTTTTTTTTTTTTTTTT	0.001741
	3686	GGGG-G-GC--CC---CCCCC-CC-CCC--CCC---CCCCCCCCCCCCCCCCCCCC	0.001248
D	626	GAGGGGGGGGGGGGGGGGGG-GGGGGGG-GGGGGGGGGGGGGGGGGGGGGGGG	−0.000006
	3725	-GCC----------	−0.000020
	6468	GAGGGGGGG--GG-GGAAAAGGG-GG--G----G-GGGGGGGGG-GGGGGGGGGG	−0.000037
	3104	TTTT-TCTT-T-T-T-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	−0.000073
	1863	TCCCCCCCC---CCCC-CCC-CCCCC----TC-CC-CT-C-CC-CCCCCCCCC-CC	−0.000093
E	2568	AGAA-CCTT-C-C-T-GGGGGGAGCCGGGGGAGGGGGGGGGGGGGAGGCGGGGAA	−0.01020
	3631	GGGG-A-C----------	−0.01066
	966	----CC--CGC-CGTCATAA-AACCCCCCCCCCCCCCCACCCCCCCGCCCCCCCTCC	−0.01326
	69	CCCCGGGGGGGCAAAACTTTT-CGGGGGA-GTGAGCGCGGGGCCGGGGGGGGGGGG	−0.01367
	3665	G----------CCC----------C-------	−0.01583
F	6825	AACCTCTTC--CC-GGTTTTCAC-CC--C----C-CCACCCCCC-CCCCCCCCCCC	−0.1134
	4710	TTCC--G-CCCCG-GCTCCCTTTCCCCCCCCCACC-CTGCCCCCCTC-CCCCCCTT	−0.1187
	6345	TCTCCGCAG--GT-CCCCCCCCC-CC--C----C-ACCTCCCCC-CCCCACCCCCC	−0.1189
	7872	TCCCCTACC--AGGGGTCCCCCGTCCCCC-CACC-TCCCCCCCCCCCC-CCCCCCC	−0.1214
	258	CGCTTTACCCCCATCCTCCCG-TCACCCT-CCCCCGCCCCCCCCCCCCCCCCCCCC	−0.1362

To provide a contrast to the sites in Group A of Table 2, the sites of Group C were selected because they appear to have experienced a change on the moderately long branch of Figure 5 that ends with the most recent common ancestor of Taxa #9-56. Among all sp-ESL values for this branch, the sites in Group C yield the highest sp-ESL values. Because this branch is not short, these sites have s-ESL values that are not far above 0.

Figure 6a plots the 109 sp-ESL values (y-axis) versus branch index (x-axis) for Site #5242. Only 22 of the 109 values are positive and only seven are greater than 1.0. The corresponding plot in Figure 6b is for Site #258. This site belongs to Group F of Table 2 and has the lowest s-ESL value in the data set. Figure 6 suggests that much of the difference in the sites with the highest and lowest s-ESL values can be explained by the large positive sp-ESL value for the change at Site #5242 on the shortest branch of the phylogeny. Rather than having any extreme negative sp-ESL values, Site #258 seems to have a low s-ESL value because of an absence of large or moderate sp-ESL values.

Figure 7 contrasts the distribution of sp-ESL values for the shortest branch on the tree with the distribution of sp-ESL values for the aforementioned moderately long branch that ends with the most recent common ancestor of Taxa #9-56. The shortest branch yields a small number of especially large positive sp-ESL values whereas the long branch yields a moderately large number of moderately large positive sp-ESL values. The contrast in sp-ESL distribution between these two branches occurs because a substitution on a very short branch is highly unusual but is a relatively big surprise when it does happen whereas substitutions on longer branches represent smaller surprises and happen somewhat often.

Empirical Data Analysis 3: Mitochondrial and Nuclear DNA of Mouse Lemur

Because the previous two empirical analyses involved highly diverged taxa, we also include an analysis of closely related taxa that were studied by Poelstra et al. (2021). Their mouse lemur RAD-seq data consists of both mitochondrial (mt-; 55 taxa) and nuclear (n-; 57 taxa) DNA from six species within the *Microcebus* genus. The PSL's of the mt-DNA and n-DNA are respectively 4048

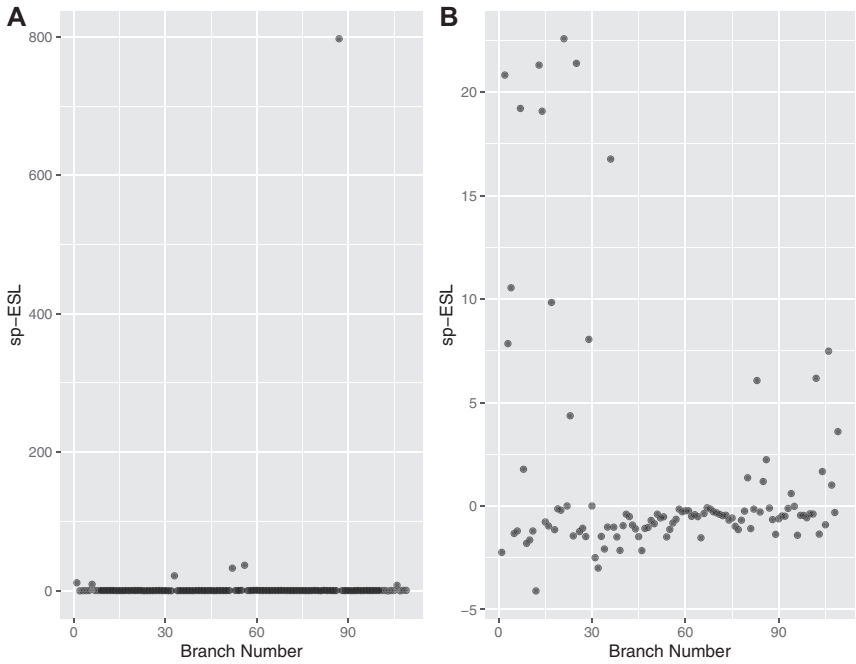

FIGURE 6. sp-ESL values for the sites with the highest and lowest s-ESL values of the ray-finned fish data. a) Individual sp-ESL values for the site with the highest s-ESL value (Site #5242). Values for 22 out of 109 branches are positive. The largest sp-ESL value corresponds to the shortest branch on the tree (i.e., the branch that ends with the most recent common ancestor of Taxa #47–56 in Fig. 5). b) Individual sp-ESL values for the site with the lowest s-ESL value (Site #258). The values are positive for 23 branches.

and 27,565 nucleotides. The mean proportions of non-gap characters per taxon are 0.4692 (±0.0271) for the mt-DNA and 0.8465 (±0.0209) for the n-DNA.

Poelstra et al. (2021) attributed the differences in topologies suggested by their concatenated mt-DNA and concatenated n-DNA to interspecific gene flow. Here, we avoid the important topic and potential consequences of concatenating RAD-seq data. We instead investigate the performance and applicability of our procedure for the concatenated mt-DNA and the concatenated n-DNA data. By adopting the GTRGAMMA model for the mt-DNA and the GTRCAT model for the n-DNA, we estimated the maximum likelihood phylogeny for the two data sets with RAxML software version 8.24 (Stamatakis 2014).

For the mt-DNA, only 46 of 107 branch length estimates are non-zero. Because our theory depends on branch length estimates having asymptotically normal distributions, its performance will be hampered when estimates are at or near their smallest possible value of zero (e.g., see Susko and Roger 2019). Therefore, we developed weak, moderate, and stringent constraints that can be used to identify branches whose length is reasonably far from zero (see Supplementary material [C] available on Dryad). The idea is to exclude branches that do not satisfy the constraints from G-factor and M-factor calculations. As explained in the Supplementary material available on Dryad, the moderate constraints seem to strike an appropriate balance between excluding branches that are not well approximated by a normal distribution and not excluding too many branches. All branches from the ray-finned fish DNA data, the eukaryotic protein data, and the mouse lemur n-DNA satisfy the moderate constraints. However, only 19 of the 107 branches of the mouse lemur mt-DNA satisfy the moderate constraints.

With only 19 branches being used for the analysis, the mt-DNA G-Factor and M-Factor estimates are 0.596 (±0.0421) and 0.955 (±0.0288). For the n-DNA, the G-Factor and M-Factor estimates are 0.810 (±0.00689) and 0.689 (±0.00993). These n-DNA estimates have substantially less uncertainty than the mt-DNA estimates due to the larger amount of sequence data and the fact that no branches are excluded by the constraints.

Applicability of ESL for Filtering or Bootstrap

For both the eukaryotic and ray-finned fish data, we performed the experiment of removing sites with negative s-ESL values, re-estimating the phylogeny, and then measuring bootstrap support of the maximum likelihood phylogeny. Some branches showed increased bootstrap support, but others showed decreases and there was no strong pattern of overall increase or decrease (data not shown). As illustrated in Figure 6, sites are typically associated with both positive and negative sp-ESL values. Furthermore, as shown in Figure 4, conventional filtering schemes may remove sites in which s-ESL's are positively large and individual sp-ESL's are distributed over both negative and positive ranges. The widespread distribution of sp-ESL values implies that data filtering is not a simple task and illustrates why

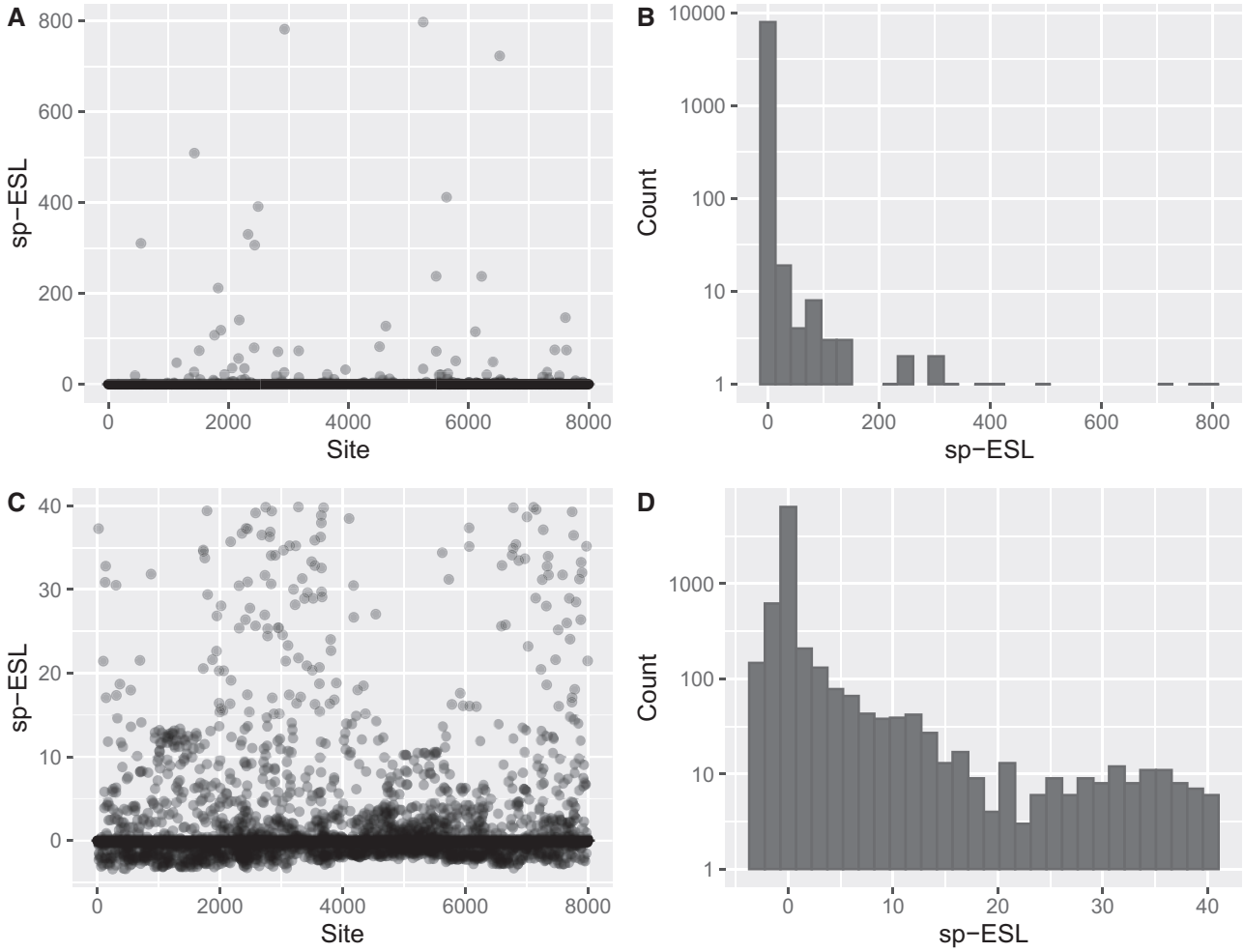

FIGURE 7. Plots and histograms of sp-ESL values from the ray-finned fish data. a) sp-ESL values from the shortest branch of the phylogeny (i.e., the branch that ends in the most recent common ancestor of Taxa #47–56). b) The log-scale histogram that corresponds to (a). c) sp-ESL values of a moderately long branch on the ray-finned fish phylogeny that ends with the most recent common ancestor of Taxa #9–56. d) The log-scale histogram that corresponds to (c).

removing columns may reduce information (Dessimoz and Gil 2010; Tan et al. 2015).

DISCUSSION

It is conventional in phylogenetics to report the number of columns in a sequence alignment (i.e., the PSL), but this statistic has little communicative value in isolation. Two alignments with the same number of columns are unlikely to be equally informative if one has no gaps and the other has abundant gaps. This work describes a procedure to measure the impact of gaps on phylogenetic information and to test model adequacy. It introduces the ESL measure that translates the phylogenetic information of a gapped data set to an intuitive summary representing how many ungapped columns would be in an equally informative alignment. An advantage of the ESL is its careful accounting of phylogenetic correlations among sequences. This work also introduces the s-ESL, p-ESL, and sp-ESL measures that can be employed to quantify the impacts of individual parameters and data components on phylogenetic analyses.

When a molecular phylogenetics study is performed, we suggest that both the ESL and the PSL be reported. If the ESL is substantially less than the PSL, investigators should consider potential sources of the disparity and how they might impact phylogenetic inferences. We discuss these issues in more detail below.

One possible cause of the PSL greatly exceeding the ESL is insertion and deletion events that occurred during the evolutionary history of the sequence data being analyzed. A large disparity between the PSL and ESL does not necessarily mean that alignment uncertainty is problematic for a phylogenetic analysis, but it is consistent with this possibility. When such a disparity occurs, more than a usual amount of attention to alignment uncertainty may be warranted. It might be possible to apply the sitewise s-ESL measure to the detection of alignment error or to have s-ESL

clarify whether filtering of alignment columns removes noise or hampers meaningful signal (e.g., Talavera and Castresana 2007; Dress et al. 2008; Capella-Gutierrez et al. 2009; Dessimoz and Gil 2010; Tan et al. 2015). If not for its computationally demanding nature, our preferred option would be to treat alignment uncertainty with a probabilistic model of insertion and deletion (e.g., Redelings and Suchard 2005).

An alternative but not mutually exclusive cause of PSL greatly exceeding ESL is that alignment gaps may represent sequence data that have not been collected for some taxa. We note that such uncollected sequence data are not necessarily missing "at random" because uncollected sequence data may be more diverged or may otherwise collectively differ from sequence data that are collected and that are therefore represented in a sequence alignment. Ascertainment bias warrants careful attention for phylogenetic inference (e.g., Felsenstein 1992; Leaché et al. 2015; Tamuri and Goldman 2017) and also for downstream analyses such as divergence time estimation.

Our approach relies on the conventional phylogenetic treatment of gaps as data that are missing at random. This conventional treatment is justified if the substitution process is independent of the insertion–deletion process and if ascertainment bias of uncollected data can be neglected. With these provisos, the conventional treatment of gaps will not cause bias in phylogenetic estimation.

Our model adequacy test could be modified to examine a null hypothesis of independence between insertion–deletion and substitution. There is ample justification for examining this hypothesis. For example, amino acid replacement and insertion–deletion are correlated through protein structure. In fact, early methods utilized gaps in multiple sequence alignments to predict coils in protein secondary structure and they further leveraged patterns of amino acid variability within alignment columns to discriminate between coils, α-helices, and *M*-strands (e.g., Benner and Gerloff 1991; Thorne et al. 1991).

The assumption of independence between nucleotide substitution and insertion–deletion can be biologically unrealistic because of both natural selection and mutation. The reason why natural selection might violate the independence assumption is straightforward. Both point mutations and insertion–deletion mutations are likely to be selected against in genomic regions that are functionally constrained. The result is that there can be a positive correlation among genomic regions between the rate at which point mutations and insertion–deletion mutations fix (e.g., see Sjödin et al. 2010). In addition, insertion–deletion and nucleotide substitution might be correlated due to mutation. For example, Tian et al. (2008) suggest that segregating insertion-deletion polymorphism might be mutagenic in heterozygous individuals.

Our work also has relevance to model selection. To compare competing models and select the best one, conventional options include the AIC (Akaike 1974), the BIC (Schwarz 1978), and the likelihood ratio test. These options cannot determine if the "best model" is significantly different from an "unknown data-generating mechanism." Procedures to test model adequacy already exist (e.g., Goldman 1993; Duchêne et al. 2018; Chen et al. 2019), but these existing procedures have a substantially different basis than ours. An attractive feature of our M-Factor approach is that it can investigate model adequacy of individual model parameters.

Due to its low power, we do not suggest that our model adequacy test is superior to conventional model comparison options but we believe that it can supplement them in order to illuminate goodness-of-fit and potentially to help pinpoint parameters associated with model deficiencies. Although the model adequacy test has low power, it strongly rejected all models that were considered for the eukaryotic and ray-finned fish data sets. These results are consistent with the view that widely-used models of sequence evolution are deeply flawed. Duchêne et al. (2018) have emphasized the importance for phylogenetic inference of carefully assessing model adequacy.

Future directions: A limitation of our procedure is that the G-Factors and M-Factors do not account for topological uncertainty of the phylogeny. The impact of this limitation is likely to greatly vary among data sets. We hope to address and assess this limitation in the future.

Here, we focused on the parameter-wise M-Factors that correspond to individual branches of the phylogenetic tree. In future work, we hope to characterize how parameter-wise M-Factors can be leveraged to improve models of nucleotide substitution, especially codon-based models of nucleotide substitution (see also Seo and Kishino 2008, 2009). The ability to interrogate the effect of individual parameters on model fit can potentially guide the development of more useful probabilistic models of sequence change.

Appendix

[A] A criterion for matrix approximation

In our derivations, we repeatedly approximate some matrix $\mathbf{A}$ with a proportion of a different matrix $\mathbf{B}$. We represent this sort of approximation and the corresponding exact relationship as

$$\begin{aligned} \mathbf{A} &\approx \alpha\mathbf{B} \\ \mathbf{A} &= \alpha\mathbf{B}+\mathbf{R}(\alpha), \end{aligned} \tag{A.1}$$

where $\mathbf{R}(\alpha)$ is the residual matrix for given α. To get an optimal α, we use the Frobenius norm ($\|\mathbf{R}(\cdot)\|_F$; Golub and Van Loan 2013) of the residual matrix which is minimized at α,

$$\alpha := \underset{t}{\operatorname{argmin}} \left\{ \|\mathbf{R}(t)\|_F := \sqrt{\sum_{i,j}\left(A_{ij}-tB_{ij}\right)^2} \right\}, \tag{A.2}$$

where A_{ij} and B_{ij} are the elements of the ith row and jth column of matrices **A** and **B**, respectively. This will be referred to as the Minimum Frobenius Norm (MFN) criterion. The optimal α can be obtained with $\frac{d}{dt}\|\mathbf{R}(t)\|_F = 0$ and

$$\alpha = \frac{\sum_{i,j} A_{ij}B_{ij}}{\sum_{i,j} B_{ij}^2}.$$

Note that α is identifiable only when both **A** and **B** are identifiable.

[B] The approximation of Fisher information matrices Assume the data were generated by the true but unknown distribution $g(\cdot)$ and will be analyzed with model $f(\cdot|\boldsymbol{\theta})$. Subject to regularity conditions, the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$ follows an asymptotically multivariate normal distribution when n is large (White 1982),

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \doteqdot N\left(\mathbf{0}, \frac{1}{n}\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf}\mathbf{I}_{gf}^{-1}\right), \tag{A.3}$$

where

$$\mathbf{I}_{gf} := \mathrm{E}_g\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(X|\boldsymbol{\theta}_*)\right]$$

$$\mathbf{J}_{gf} := \mathrm{E}_g\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X|\boldsymbol{\theta}_*)\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X|\boldsymbol{\theta}_*)\right\}^T\right],$$

and where $\mathrm{E}_g[\cdot]$ implies the expectation is performed with respect to the true distribution $g(\cdot)$. If the true and adopted models are identical (i.e., $g(x) = f(x|\boldsymbol{\theta})$), $\mathbf{I}_{gf} = \mathbf{J}_{gf}$ and the variance of equation (A.3) is reduced to $\{\mathbf{I}_{gf}^{-1}/n\}$.

Similar to equation (A.3), the MLE $\tilde{\boldsymbol{\theta}}$ for incomplete data follows an asymptotically multivariate normal distribution when the data are generated by $g(\cdot)$ but analyzed with $f(\cdot|\boldsymbol{\theta})$,

$$(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \doteqdot N\left(\mathbf{0}, \frac{1}{n}\tilde{\mathbf{I}}_{gf}^{-1}\tilde{\mathbf{J}}_{gf}\tilde{\mathbf{I}}_{gf}^{-1}\right), \tag{A.4}$$

where

$$\tilde{\mathbf{I}}_{gf} := \mathrm{E}_g\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(\tilde{X}|\boldsymbol{\theta}_*)\right]$$

$$\tilde{\mathbf{J}}_{gf} := \mathrm{E}_g\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\tilde{X}|\boldsymbol{\theta}_*)\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\tilde{X}|\boldsymbol{\theta}_*)\right\}^T\right],$$

and where the notation $\tilde{X}$, $\tilde{\mathbf{I}}_{gf}$ and $\tilde{\mathbf{J}}_{gf}$ imply that the data contains gaps. The inverse of a variance such as found in equation (A.4) is conventionally referred to as information. For equation (A.4), the inverse of the variance represents the amount of information contained in an incomplete data set with a PSL of n.

Our purpose is to develop an approximate relationship between the information in the observed n columns of incomplete (gapped) data and the information in n_e columns of complete (ungapped) data such that $n\tilde{\mathbf{I}}_{gf}\tilde{\mathbf{J}}_{gf}^{-1}\tilde{\mathbf{I}}_{gf} \approx n_e\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf}$ and so that we can estimate n_e. We define $k_1 := n_e/n$ and parallel equation (A.1) with the following approximate and exact relationships,

$$\begin{aligned}\tilde{\mathbf{I}}_{gf}\tilde{\mathbf{J}}_{gf}^{-1}\tilde{\mathbf{I}}_{gf} &\approx k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf},\\ \tilde{\mathbf{I}}_{gf}\tilde{\mathbf{J}}_{gf}^{-1}\tilde{\mathbf{I}}_{gf} &= k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf} + \mathbf{E}_1,\end{aligned} \tag{A.5}$$

where $\mathbf{E}_1$ is the residual matrix for k_1. With only an incomplete data set, the complete-data quantities $\mathbf{I}_{gf}$ and $\mathbf{J}_{gf}$ are unidentifiable. These quantities represent expectations with respect to the unknown $g(\cdot)$ and there is no way to identify them even via simulation. Because $\mathbf{I}_{gf}$ and $\mathbf{J}_{gf}$ are unidentifiable, k_1 is unidentifiable in equation (A.5). In contrast, $\tilde{\mathbf{I}}_{gf}$ and $\tilde{\mathbf{J}}_{gf}$ are identifiable because the observed incomplete data were generated with $g(\cdot)$.

Similar to equation (A.1), consider the following approximate and exact relationships

$$\begin{aligned}\tilde{\mathbf{I}}_{gf}^{-1}\tilde{\mathbf{J}}_{gf} &\approx k_2\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf}\\ \tilde{\mathbf{I}}_{gf}^{-1}\tilde{\mathbf{J}}_{gf} &= k_2\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf} + \mathbf{E}_2,\end{aligned} \tag{A.6}$$

where k_2 is still unidentifiable because of $\mathbf{I}_{gf}$ and $\mathbf{J}_{gf}$. Define $\mathbf{I}_{ff}$ and $\tilde{\mathbf{I}}_{ff}$ as

$$\begin{aligned}\mathbf{I}_{ff} &= \mathrm{E}_f\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(X|\tilde{\boldsymbol{\theta}})\right]\\ \tilde{\mathbf{I}}_{ff} &= \mathrm{E}_f\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(\tilde{X}|\tilde{\boldsymbol{\theta}})\right],\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ is the MLE for given incomplete data and is compatible with $\tilde{\boldsymbol{\theta}}$ of equation (A.4). An important feature of both $\mathbf{I}_{ff}$ and $\tilde{\mathbf{I}}_{ff}$ is that the expectation is performed with respect to the adopted model $f(\cdot|\tilde{\boldsymbol{\theta}})$ so that these quantities are identifiable via simulation. By generating extremely long complete and incomplete sequences with model $f(\cdot|\tilde{\boldsymbol{\theta}})$, $\mathbf{I}_{ff}$ and $\tilde{\mathbf{I}}_{ff}$ can be estimated.

Similar to equation (A.1), consider the following approximate and exact relationships

$$\begin{aligned}\mathbf{I}_{gf} &\approx k_3\mathbf{I}_{ff}\\ \mathbf{I}_{gf} &= k_3\mathbf{I}_{ff} + \mathbf{E}_3,\end{aligned} \tag{A.7}$$

where k_3 is unidentifiable because of $\mathbf{I}_{gf}$.

Applying equations (A.6) and (A.7) to equation (A.5),

$$\begin{aligned}\tilde{\mathbf{I}}_{gf}\tilde{\mathbf{J}}_{gf}^{-1}\tilde{\mathbf{I}}_{gf} &= k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf} + \mathbf{E}_1\\ \Longleftrightarrow \tilde{\mathbf{I}}_{gf}\tilde{\mathbf{J}}_{gf}^{-1}\tilde{\mathbf{I}}_{gf} \times \tilde{\mathbf{I}}_{gf}^{-1}\tilde{\mathbf{J}}_{gf} &= \left\{k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf} + \mathbf{E}_1\right\}\\ &\quad \times \left\{k_2\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf} + \mathbf{E}_2\right\}\end{aligned}$$

$$\begin{aligned}\Longleftrightarrow \tilde{\mathbf{I}}_{gf} &= k_1k_2\mathbf{I}_{gf}+k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf}\mathbf{E}_2 \\ &\quad +k_2\mathbf{E}_1\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf}+\mathbf{E}_1\mathbf{E}_2 \qquad \text{(A.8)}\\ &= k_1k_2\left\{k_3\mathbf{I}_{ff}+\mathbf{E}_3\right\}+k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1} \\ &\quad \times\mathbf{I}_{gf}\mathbf{E}_2+k_2\mathbf{E}_1\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf}+\mathbf{E}_1\mathbf{E}_2 \\ &= k_1k_2k_3\mathbf{I}_{ff}+k_1k_2\mathbf{E}_3+k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1} \\ &\quad \times\mathbf{I}_{gf}\mathbf{E}_2+k_2\mathbf{E}_1\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf}+\mathbf{E}_1\mathbf{E}_2.\end{aligned}$$

Simplifying the last line of the above equation yields

$$\begin{aligned}\tilde{\mathbf{I}}_{gf} &\approx k_1k_2k_3\mathbf{I}_{ff} \\ \tilde{\mathbf{I}}_{gf} &= k_1k_2k_3\mathbf{I}_{ff}+\mathbf{E}, \qquad \text{(A.9)}\end{aligned}$$

where $\mathbf{E}=k_1k_2\mathbf{E}_3+k_1\mathbf{I}_{gf}\mathbf{J}_{gf}^{-1}\mathbf{I}_{gf}\mathbf{E}_2+k_2\mathbf{E}_1\mathbf{I}_{gf}^{-1}\mathbf{J}_{gf}+\mathbf{E}_1\mathbf{E}_2$.

Because both $\tilde{\mathbf{I}}_{gf}$ and $\mathbf{I}_{ff}$ are identifiable in equation (A.9), the product of $\{k_1k_2k_3\}$ can be estimated by applying the MFN criterion of equation (A.2). While the product $k_1k_2k_3$ can be inferred, the three factors in this product cannot be individually estimated with our approach.

Finally, consider the following approximate and exact relationships,

$$\begin{aligned}\tilde{\mathbf{I}}_{ff} &\approx k_4\mathbf{I}_{ff} \\ \tilde{\mathbf{I}}_{ff} &= k_4\mathbf{I}_{ff}+\mathbf{E}_4, \qquad \text{(A.10)}\end{aligned}$$

where both $\tilde{\mathbf{I}}_{ff}$ and $\mathbf{I}_{ff}$ are identifiable via simulation. Because random data can be generated and analyzed with $f(\cdot|\tilde{\boldsymbol{\theta}})$, k_4 is free from the model misspecification issue. Applying the MFN criterion to equation (A.10),

$$k_4 := \frac{\sum_{i,j} I_{ffij}\tilde{I}_{ffij}}{\sum_{i,j} I_{ffij}^2}, \qquad \text{(A.11)}$$

where I_{ffij} and $\tilde{I}_{ffij}$ are respectively the elements at the ith row and jth column of matrices $\mathbf{I}_{ff}$ and $\tilde{\mathbf{I}}_{ff}$.

Although equation (A.11) relies upon the model $f(\cdot|\tilde{\boldsymbol{\theta}})$ being used for both data generation and analysis, we found via simulation and empirical data analysis (see Results section) that estimates of k_4 in equation (A.10) are relatively robust to the choice of model used for analysis. That is, when we generate random sequences with $f(\cdot|\tilde{\boldsymbol{\theta}})$ and analyze them with an incorrect model $h(\cdot|\boldsymbol{\theta})$, we can consider the following relationship that is similar to equation (A.10),

$$\begin{aligned}\tilde{\mathbf{I}}_{fh} &\approx k_4'\mathbf{I}_{fh} \\ \tilde{\mathbf{I}}_{fh} &= k_4'\mathbf{I}_{fh}+\mathbf{E}_4'. \qquad \text{(A.12)}\end{aligned}$$

As described in the Results section, we found $k_4' \approx k_4$ even for an incorrect model $h(\cdot|\boldsymbol{\theta})$. We note that equation (A.8) has the same structure as equation (A.12). Therefore, we can expect $k_1k_2 \approx k_4' \approx k_4$ for a carefully selected model $f(\cdot|\tilde{\boldsymbol{\theta}})$ that is not too far from the truth.

Applying the MFN criterion to equation (A.9) followed by replacing k_1k_2 with k_4 leads to

$$\begin{aligned}k_3 &\approx \frac{1}{k_4}\frac{\sum_{i,j} I_{ffij}\tilde{I}_{gfij}}{\sum_{i,j} I_{ffij}^2} \\ &= \frac{\sum_{i,j} I_{ffij}\tilde{I}_{gfij}}{\sum_{i,j} I_{ffij}\tilde{I}_{ffij}}. \qquad \text{(A.13)}\end{aligned}$$

The formula for k_4 in equation (A.11) and the formula for k_3 in equation (A.13) both involve off-diagonal elements. Assuming these off-diagonal elements can be ignored, we have

$$\begin{aligned}k_3 &\approx \frac{\sum_i I_{ffii}\tilde{I}_{gfii}}{\sum_i I_{ffii}\tilde{I}_{ffii}} =: M \\ k_4 &\approx \frac{\sum_i I_{ffii}\tilde{I}_{ffii}}{\sum_i I_{ffii}^2} =: G.\end{aligned}$$

We respectively refer to G and M as the overall G-Factor and the overall M-Factor, with G being the ESL/PSL ratio and M being the model misspecification factor. The ESL of the given incomplete sequence data can be obtained with $G\times$PSL. If we define

$$\begin{aligned}u_i &:= \frac{I_{ffii}^2}{\sum_i I_{ffii}^2} \\ v_i &:= \frac{I_{ffii}\tilde{I}_{ffii}}{\sum_i I_{ffii}\tilde{I}_{ffii}},\end{aligned}$$

G and M can be represented as the weighted average of the parameter-wise G_i's and M_i's as defined in equations (6) and (7),

$$\begin{aligned}G &= \sum_{i=1}^{d} u_iG_i \\ M &= \sum_{i=1}^{d} v_iM_i.\end{aligned}$$

[C] Fisher information implementation

In our implementation, two important simplifications are made regarding Fisher Information estimates. First, branch lengths are the only type of parameter considered. This is mainly motivated by a desire to avoid numerical instability complications. We expect that a consequence of this simplification is reduced power of our model adequacy test because parameters controlling character-state transitions are not considered. Second, our implementation adopts the analytic formula for diagonal elements of the Hessian matrix that correspond to branch length estimates (Yang 2000) but it ignores off-diagonal elements in the Fisher information matrix

because they could be burdensome with regard to computational time and memory. Furthermore, our bootstrap implementation saves computation and storage by resampling only sitewise second derivatives (i.e., sitewise diagonal elements of the Hessian matrix).

One way to assess the importance of off-diagonal elements is to measure their contribution to the Frobenius norm of equation (A.2). This is straightforward when the data-generation mechanism and adopted model match so that $\tilde{\mathbf{I}}_{ff} = \tilde{\mathbf{J}}_{ff}$ and so that off-diagonal elements of $\tilde{\mathbf{J}}_{ff}$ can be estimated by using the outer product of sitewise first derivatives (Porter 2002; Seo et al. 2004). The proportional contribution of the diagonal elements of the Frobenius norm is

$$\widehat{r}_d := \frac{\sqrt{\sum_i |\widehat{\tilde{J}}_{ffii}|^2}}{\sqrt{\sum_i \sum_j |\widehat{\tilde{J}}_{ffij}|^2}}.$$

We used the maximum likelihood estimates of the eukaryotic data set for the LG+Gamma model to simulate a data set with the observed size and gap patterns. We then analyzed this simulated data set with LG+Gamma and obtained $\widehat{r}_d = 0.962$. This high proportion suggests that the diagonal elements are summarizing most of the information.

In a separate experiment, we randomly selected three of the 500 data sets that were simulated according to Figure 1. These three data sets yielded $\widehat{r}_d$ values of 0.940, 0.944, and 0.945. Coupled with the simulation results of $\widehat{\overline{G_1}} = 0.500$ (Fig. 1), the high $\widehat{r}_d$ proportions imply ignoring the off-diagonals will have a minor impact on G-factor estimation.

Although the high $\widehat{r}_d$ values calculated in these experiments are all for the situation where $f(\cdot|\boldsymbol{\theta}) = g(\cdot)$, we expect that ignoring off-diagonal elements of $\tilde{\mathbf{I}}_{gf}$ will tend not to be problematic when $f(\cdot|\boldsymbol{\theta})$ is a reasonably good approximation for $g(\cdot)$.

Availability

Program code and example files are available at https://github.com/diploid2n/ESL.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.zs7h44j9f.

Funding

This work was supported by the Korea Polar Research Institute [PE21130 and PE21140 to T.-K.S.]; PRAIRIE [ANR-19-P3IA-0001 to O.G.]; N.S.F. [DEB-1754142] and N.I.H. [R01 GM118508] to J.L.T.

Acknowledgments

We thank Bryan Carstens, Simon Ho, and two anonymous reviewers for providing invaluable suggestions.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19:716–723.
Benner S.A., Gerloff D. 1991. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv. Enzyme Regul. 1991:121-181.
Bishop Y.M., Fienberg S.E., Holland P.W. 2007. Discrete multivariate analysis. New York: Springer. p. 475–484.
Bouchard-Côté A., Jordan M.I. 2013. Evolutionary inference via the Poisson Indel Process. Proc. Natl. Acad. Sci. USA 110(4):1160-1166.
Capella-Gutierrez S., Silla-Martinez J.M., Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.
Chen W., Kenney T., Bielawski J., Gu H. 2019. Testing adequacy for DNA substitution models. BMC Bioinformatics 20(1):1-6.
Dayhoff M.O., Schwartz R. M., Orcutt B.C. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure, vol. 5, Suppl. 3. Washington DC: National Biomedical Research Foundation. p. 345–352.
De Maio N. 2021. The cumulative indel model: fast and accurate statistical evolutionary alignment. Syst. Biol. 70(2):236–257.
Derelle R., Lang B.F. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol. Biol. Evol. 29(4):1277–1289.
Dessimoz C., Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11:R37.
Dress A.W., Flamm C., Fritzsch G., Grunewald S., Kruspe M., Prohaska S.J., Stadler P.F. 2008. Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol. Biol. 3:7.
Duchêne D.A., Duchêne S., Ho S.Y.W. 2018. Differences in performance among test statistics for assessing phylogenomic model adequacy. Genome Biol. Evol. 10(6):1375-1388.
Felsenstein J. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution 46(1):159-173.
Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.
Fleissner R., Metzler D., von Haeseler A. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst. Biol. 54:548–561.
Goldman N. 1993. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. 37:650–661.
Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B 265:1779–1786.
Goldman N., Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.
Golub G.H., Van Loan C.F. 2013. Matrix computations. 4th ed. Baltimore: John Hopkins University Press. p. 71.
Hall P., Wilson S.R. 1991. Two guidelines for bootstrap hypothesis testing. Biometrics 47:757–762.
Hein J., Wiuf C., Knudsen B., Moller M.B., Wibling G. 2000. Statistical alignment: computational properties, homology testing and goodness-of-fit. J. Mol. Biol. 302(1):265-279.
Holmes I. 2020. A model of indel evolution by finite-state, continuous-time machines. Genetics 216(4):1187-1204.
Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.
Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25(7):1307–1320.
Leaché A.D., Banbury B.L., Felsenstein J., De Oca A.N., Stamatakis A. 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Syst. Biol. 64(6):1032-1047.
Li C., Lu G., Orti G. 2008. Optimal data partitioning and a test for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst. Biol. 57:519–539.
Metzler D. 2003. Statistical alignment based on fragment insertion and deletion models. Bioinformatics. 19(4):490-499.
Piel W.H., Chan L., Dominus M.J., Ruan J., Vos R.A., Tannen V. 2009. TreeBASE v. 2: a database of phylogenetic knowledge. In: e-BioSphere.
Poelstra J.W., Salmona J., Tiley G.P., Schler D., Blanco B.M., Andriambeloson J.B., Bouchez O., Campbell C.R., Etter P.D., Hohenlohe P.A., Hunnicutt K.E., Iribar A., Johnson E.A., Kappeler P.M., Larsen P.A., Manzi S., Ralison J.M., Randrianambinina B., Rasoloarison R.M., Rasolofoson D.W., Stahlke A.R., Weisrock D.W., Williams R.C., Chikhi L., Louis E.L. Jr., Radespiel U., Yoder A.D. 2021. Cryptic patterns of speciation in cryptic primates: microendemic mouse lemurs and the multispecies coalescent. Syst. Biol. 70(2):203-218.
Porter J. 2002. Efficiency of covariance matrix estimators for maximum likelihood estimation. J. Bus. Econ. Stat. 20:431-440.
Posada D., Buckley T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793-808.
Posada D., Crandall, K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.
Robins J.M., Van Der Vaart A., Ventura V. 2000. Asymptotic distribution of P values in composite null models. J. Am. Stat. Assoc. 95(452):1143-1156.
Redelings B.D., Suchard M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54:401–418.
Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.
Seo T.-K., Kishino H., Thorne J.L. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. Mol. Biol. Evol. 21:1201–1213.
Seo T.-K., Kishino H. 2008. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst. Biol. 57:367-377.
Seo T.-K., Kishino H. 2009. Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. Syst. Biol. 58:199–210.
Seo T.-K., Thorne J.L. 2018. Information criteria for comparing partition schemes. Syst. Biol. 67:616–632.
Sjödin P., Bataillon T., Schierup M.H. 2010. Insertion and deletion processes in recent human history. PLoS One 5(1):e8650.
Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Susko E., Roger A.J. 2019. On the use of information criteria for model selection in phylogenetics. Mol. Biol. Evol. 37(2):549-562.
Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564-577.
Tamura K., Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.
Tamuri A., Goldman N. 2017. Avoiding ascertainment bias in the maximum likelihood inference of phylogenies based on truncated data. BioRxiv:186478. Available from: https://doi.org/10.1101/186478.
Tan G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., Dessimoz C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst. Biol. 64:778-791.
Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.
Thorne J.L., Kishino H., Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.
Thorne J.L., Kishino H., Felsenstein J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34:3–16.
Thornton J.M., Flores T.P., Jones D.T., Swindells M.B. 1991. Prediction of progress at last. Nature 354:105-106.
Tian D., Wang Q., Zhang P., Araki H., Yang S., Kreitman M., Nagylaki T., Hudson R., Bergelson J., Chen J.Q. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455(7209):105-108.
Vos R.A., Balhoff J.P., Caravas J.A., Holder M.T., Lapp H. Maddison W.P., Midford P.E., Priyam A., Sukumaran J., Xia X., Stoltzfus A. 2012. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. Syst. Biol. 61:675-689.
White H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50:1–25.
Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18:691–699.
Yang Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.
Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.
Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51:423–432.