Dwueng-Chwuan Jhwueng, Snehalata Huzurbazar, Brian C. O'Meara and Liang Liu*

Investigating the performance of AIC in selecting phylogenetic models

Abstract: The popular likelihood-based model selection criterion, Akaike's Information Criterion (AIC), is a breakthrough mathematical result derived from information theory. AIC is an approximation to Kullback-Leibler (KL) divergence with the derivation relying on the assumption that the likelihood function has finite second derivatives. However, for phylogenetic estimation, given that tree space is discrete with respect to tree topology, the assumption of a continuous likelihood function with finite second derivatives is violated. In this paper, we investigate the relationship between the expected log likelihood of a candidate model, and the expected KL divergence in the context of phylogenetic tree estimation. We find that given the tree topology, AIC is an unbiased estimator of the expected KL divergence. However, when the tree topology is unknown, AIC tends to underestimate the expected KL divergence for phylogenetic models. Simulation results suggest that the degree of underestimation varies across phylogenetic models so that even for large sample sizes, the bias of AIC can result in selecting a wrong model. As the choice of phylogenetic models is essential for statistical phylogenetic inference, it is important to improve the accuracy of model selection criteria in the context of phylogenetics.

Keywords: AIC; Kullback-Leibler divergence; model selection; phylogenetics.

DOI 10.1515/sagmb-2013-0048

1 Introduction

Probabilistic models are fundamental to statistical phylogenetic inference (Johnson and Omland, 2004; Sullivan and Joyce, 2005; Kelchner, 2009). A phylogenetic model assumes that the evolution of molecular sequences follows a substitution process along the branches of a phylogenetic tree. The random process of nucleotide substitutions over time is described probabilistically by a substitution model; over the years many such substitution models have been developed. The parameters in a phylogenetic model include the branch lengths and topology of the phylogenetic tree, as well as the parameters in the substitution model (Bos and Posada, 2005).

One of the major goals of phylogenetic model selection is to select a good substitution model for estimating phylogenetic trees from sequence data (Shapiro et al., 2006). Since statistical approaches for phylogenetic inference are based on particular models, model choice may significantly affect the resulting estimates of the phylogenetic parameters (Buckley and Cunningham, 2002; Posada and Buckley, 2004). Standard model selection criteria have been introduced for selecting phylogenetic models, but the biggest challenge in

***Corresponding author: Liang Liu,** Department of Statistics and Institute of Bioinformatics, University of Georgia, 101 Cedar Street, Athens, GA 30606 USA, Phone: +1-706-542-3309, Fax: +1-706-542-3391, e-mail: lliu@uga.edu
Dwueng-Chwuan Jhwueng: Department of Statistics, Feng-Chia University, Taichung, Taiwan 40724, R.O.C.
Snehalata Huzurbazar: Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709, USA; Department of Statistics, University of Wyoming, Laramie, WY 82071, USA; and Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
Brian C. O'Meara: Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA

phylogenetic model selection is that it involves a high dimensional discrete parameter, namely, the tree topology (Kelchner and Thomas, 2007). Likelihood ratio tests (LRT) have been used extensively for testing models in phylogenetics (Frati, 1997; Huelsenbeck and Crandall, 1997; Sullivan and Swofford, 1997; Posada and Crandall, 1998; Anisimova and Gascuel, 2006). However, criteria based on hypothesis testing are not suitable for selecting phylogenetic models, as they tend to favor complex models (Burham and Anderson, 2004), and also suffer from serious issues of implementation in this context (Cunningham et al., 1998; Pol, 2004). Likelihood-based model selection criteria, An Information Criterion (AIC) due to Akaike (1974), its modification or corrected AIC, namely, AICc derived by Hurvich and Tsai (1989), have been commonly used for assessing and comparing the fit of various phylogenetic models to sequence data (Holder et al., 2010; Posada and Crandall, 2001; Boettiger et al., 2012). In a Bayesian framework, phylogenetic models are compared via Bayes factors (Alfaro and Huelsenbeck, 2006). When the models have equal prior probabilities, Bayes factors are the ratio of the marginal likelihoods of the models. Bayesian Information Criterion (BIC) (Schwarz, 1978) was derived using Laplace's approximation for the log-transformed marginal likelihoods, and has been widely used for selecting phylogenetic models (Minin et al., 2003; Huelsenbeck et al., 2004; Evans and Sullivan, 2010; Wu et al., 2013). Although LRT, AIC and BIC rely on the initial phylogeny, it has been demonstrated by simulation that varying phylogenies has little impact on the performance of these model selection criteria (Posada and Crandall, 2001; Abdo et al., 2005; Luo et al., 2010). Moreover, previous studies have shown that although the choice of model can affect the maximum likelihood estimate of the phylogenetic tree, the differences are confined to the clades with low bootstrap support (Rippinger and Sullivan, 2008). In general, choice of the substitution model does not affect the main evolutionary inferences based on the estimated phylogeny (Rippinger and Sullivan, 2008).

The AIC and its variants were derived from information theory as an approximation to the expected Kullback-Leibler (KL) divergence which measures the distance of a candidate model from the true model. Specifically, subject to certain continuity conditions for the underlying model likelihood, the expected KL divergence can be approximated by $AIC=2\lambda-2L$ where λ is the number of parameters in the model and L is the log likelihood evaluated at the maximum likelihood estimates. This version of AIC was originally derived in the context of linear regression, but has been broadly applied to a wider range of models, including those in phylogenetics (Posada and Crandall, 2001; Posada and Buckley, 2004; Jermiin et al., 2008; Rippinger and Sullivan, 2008). In the latter context, the continuity restrictions underlying the approximation may not be met since parameters of phylogenetic models include tree structures which are not continuous when the phylogenetic trees are bifurcating trees. Hence, it is unclear whether AIC is an appropriate approximation to the expected KL divergence in the context of phylogenetic tree estimation, with the practical implication that model selection guided by AIC is potentially misleading.

Given the prevalent use of AIC in selecting phylogenetic models using sequence data, it is surprising that there is limited work on investigating its properties as an estimator of the expected KL divergence. In this paper, we investigate the asymptotic properties of AIC as an approximation to the expected KL divergence and then use simulations to evaluate the performance of AIC and AICc in selecting appropriate substitution models. We show that under the standard condition, AIC is an unbiased estimate of the expected KL divergence when the tree topology is given. However, when the tree topology is estimated from data, AIC underestimates the expected KL divergence. The simulation results suggest that the scales of bias of AIC and its variants (AICc) as the estimates of the expected KL divergence vary across substitution models. Thus the tree topology should be counted as model parameters when calculating AIC and AICc.

The paper is organized as follows: Section 2 defines AIC and KL in the context of phylogenetic models. In Section 3, we investigate the asymptotic properties of AIC as an estimate of the expected Kullback-Leibler divergence in the context of phylogenetic models. We show that AIC underestimates the expected KL divergence when the tree topology is estimated from data. Simulation results suggest that the scale of underestimation of AIC varies across phylogenetic models. Section 4 is devoted to a discussion on the consequences of using AIC without taking into account the tree topology, and potential solutions to correcting the bias of AIC using bootstrap techniques.

2 AIC and KL for phylogenetic models

As mentioned previously, AIC and its related criteria, AICc and BIC were not developed for complex models as those in phylogenetics but that has not precluded their current use with such models. In phylogenetics, we have molecular data, namely DNA sequences, which are assumed to have been generated from a substitution process along the lineages of a phylogenetic tree. The substitution process is modeled as a continuous time Markov process with a transition rate matrix Q describing the rates at which the nucleotides of one type change into other types. For a substitution model, the parameters, θ, include the rate parameters in Q and the equilibrium probabilities (π_A, π_C, π_G, π_T). The transition probability matrix $P(t)$ can be computed from rate matrix Q, i.e., $P(t)=e^{Qt}$, where the elements of $P(t)$, the transition probabilities, are exponential functions of the substitution model parameters θ. Thus most substitution models based on a continuous time Markov process have continuous second-order partial derivatives with respect to the model parameters, i.e., $\frac{\partial^2 P(t)}{\partial\theta\partial\theta^t}$ exists and is a continuous function of θ. In examining phylogenetic models, we consider only the substitution models for which the condition of existence of continuous second-order partial derivatives is satisfied. Given a phylogenetic tree, the set of such phylogenetic models is denoted by Ω, which includes most phylogenetic models available in a widely used phylogenetic model selection program jModelTest (Guindon and Gascuel, 2003; Darriba et al., 2012). Specifically, these models include JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), F81 (Felsenstein, 1981), HKY (Hayasaka et al., 1988), SYM (Zharkikh, 1994), GTR (Tavaré, 1986), and models that use a gamma distribution to handle variable rates among sites (Yang, 1994) as well as those considering invariant sites.

Specifically, the DNA sequence data is an $N\times M$ matrix in which N is the number of sequences and M is the sequence length (i.e., the number of nucleotides on each aligned sequence). Let D be this sequence matrix and D_{uv} be the v^{th} (v=1, …, M) nucleotide of sequence u (u=1, …, N); note that $D_{uv}\in\{A, C, G, T\}$. We denote a particular substitution model as m with parameters θ, as described above. We use $|\theta|$ to denote the length of θ, i.e., the number of parameters contained in θ. In addition, each phylogenetic tree consists of two types of parameters, tree topology τ and branch lengths b. Each column in the data matrix D is a site of the sequences. It is often assumed that the sites (or columns) evolve independently along the lineages of a phylogenetic tree. As each column is a vector of N nucleotides, there are 4^N possible nucleotide patterns. Let p_i denote the probability of observing the i^{th} nucleotide pattern, where p_i is the sum of the product of transition probabilities $P(t)$. Thus, $p_i=p_i(D|\phi)$ is a function of model parameters $\phi=(\theta, \tau, b)$, and has continuous second-order partial derivatives with respect to θ and b (see Appendix A1). Let $\xi=\{\xi_1, \xi_2, \ldots, \xi_{4^N}\}$ be the frequencies of 4^N patterns observed in data D. The probability density function of data D in terms of frequencies ξ of 4^N nucleotide patterns can be expressed as

$$f(D|\phi)=\prod_{i=1}^{4^N} p_i(D|\phi)^{\xi_i}. \tag{1}$$

It follows from (1) that the log likelihood function L is

$$L(D|\phi)=\sum_{i=1}^{4^N}\xi_i \ln p_i(D|\phi). \tag{2}$$

The log likelihood function for a fixed model $m\in\Omega$ is denoted by $L_m(D|\phi_m)$, or simply by L_m.

Kullback-Leibler (KL) divergence is a directed discrepancy measure for assessing the discrepancy of a candidate model m with parameters $\phi_m=(\theta_m, \tau_m, b_m)$ from the true model m^* with parameters $\phi_*=(\theta_*, \tau_*, b_*)$. KL divergence for a candidate model $m\in\Omega$ is defined as

$$KL(m, m^*)=\int \log\left(\frac{f_{m^*}(D|\phi_*)}{f_m(D|\phi_m)}\right) f_{m^*}(D|\phi_*)\,dD, \tag{3}$$

which is essentially the expectation of the ratio of the true and candidate models on the natural log scale, leading to the following difference,

$$KL(m, m^*)=E_{m^*}(L_{m^*}(D|\phi_*))-E_{m^*}(L_m(D|\phi_m)). \tag{4}$$

In (4), E_{m^*} represents the expectation with respect to the true model m^*. The Kullback-Leibler divergence is bounded below by 0, and if all candidate models have zero KL divergence, then the aim is to favor simpler models over more complicated counterparts.

The literature on AIC is vast, and various derivations are available. The overview here borrows heavily from Davison (2003) and Burham and Anderson (2004). Assuming that there exists a true model that generated the observed sequence data D. Applying the usual version of AIC, for a fixed phylogenetic model $m\in\Omega$ we simply use

$$AIC_m=2\lambda_m-2L_m(D|\hat{\phi}_m), \tag{5}$$

where $\hat{\phi}_m$ are the maximum likelihood estimates of $\phi_m=(\theta_m, \tau_m, b_m)$ under model m. The number λ_m of parameters in model m is equal to the sum of the number of parameters in the substitution model and the number of branches on the phylogenetic tree, i.e., $\lambda_m=|\theta_m|+|b_m|$. The topology of the tree is often not counted when calculating λ_m. Model selection starts with a set of candidate models $m\in\Omega$ and for each, AIC is calculated for the same data D, with final selection of the model which has the minimum AIC value. It should be noted that the maximum likelihood estimates of parameters (θ_m, τ_m, b_m) vary across distinct candidate models. Similarly, AICc is an extension for situations when the sample size, n, is small compared to the number of parameters (for instance, $\frac{n}{\lambda_m}<40$) and the correction gives $AICc=AIC+\frac{2\lambda_m(\lambda_m+1)}{n-\lambda_m-1}$. The Bayesian information criterion (BIC) is related to AIC via $BIC=AIC+\lambda_m(\ln n-2)$. The sample size n in *AICc* and *BIC* is defined as the sequence length (i.e., the number of base pairs on the alignments) (Posada, 2008).

Essentially, the performance of AIC depends on how well it approximates the expected Kullback-Leibler divergence. Although under certain conditions AIC is a good estimate of the expected KL divergence, these conditions have not been carefully checked in the context of phylogenetics. The model selection criteria (AIC and its variants) stated here were developed under continuity assumptions for the underlying model likelihood, assumptions which in general are not satisfied for phylogenetic models where two complicating issues arise. First is the issue of considering the tree topology as a parameter, and realizing that for a fixed scenario, a vast number of tree topologies are possible. For an unrooted phylogenetic tree with $j\geq 3$ taxa, there are $\frac{(2j-5)!}{2^{j-3}(j-3)!}$ possible tree topologies. With $j=10$, there are over two millions possible tree topologies. Hence searching for the best tree of a large number of taxa is computationally intensive. Furthermore, given that phylogenetic trees are bifurcating trees, this tree space is discrete, so that the underlying log likelihood function $L_m(D|\theta_m, \tau_m, b_m)$ as a function of topologies τ_m, fails to be continuous.

3 Approximation to expected KL divergence for phylogenetic models

Let $\hat{\phi}_m=(\hat{\theta}_m, \hat{\tau}_m, \hat{b}_m)$ be the maximum likelihood estimates (MLEs) of model parameters ϕ_m in the log likelihood function $L_m(D|\phi_m)$. The quantity in (4) is a conditional expectation given $\hat{\phi}_m$, i.e.,

$$KL(m, m^*|\hat{\phi}_m)=E_{m^*}(L_{m^*}(D|\phi_*))-E_{m^*}(L_m(D|\hat{\phi}_m)|\hat{\phi}_m). \tag{6}$$

To remove the dependence on the estimates $\hat{\phi}_m$, we take the expectation of the quantity in (6) with respect to the distribution of $\hat{\phi}_m$, giving the expected Kullback-Leibler divergence, $E_{\hat{\phi}_m}(KL(m, m^*|\hat{\phi}_m))$; AIC, AICc, and others are different estimators of this expectation. Specifically, in (6), when comparing various models, the first expectation is common to all. Thus, we drop the first term, but for simplicity, we continue to call the second term KL. Following convention rooted in likelihood theory, we express the expected Kullback-Leibler divergence as multiplied by two and concentrate on

$$E(KL)=-2E_{\hat{\phi}_m}(E_{m^*}(L_m(D|\hat{\phi}_m)|\hat{\phi}_m)). \tag{7}$$

The key to obtaining estimators of the expected KL divergence is a Taylor series expansion of $L_m(D|\hat{\phi}_m)$, and a necessary requirement of the expansion is the existence of the second derivative of $L_m(D|\phi_m)$ with respect to ϕ_m (Burham and Anderson, 2004). We now investigate the form of such approximations in the context of phylogenetic models.

3.1 AIC for phylogenetic models

We first consider a simple case where the tree topology τ_m is given, and simplify notation by dropping the parameter τ from the log likelihood function. The remaining parameters in ϕ are denoted by ρ, with $\hat{\rho}_m=(\hat{\theta}_m, \hat{b}_m)$ as the MLEs under model m. The expected KL divergence becomes

$$E(KL)=-2E_{\hat{\rho}_m}(E_{m^*}(L_m(D|\hat{\rho}_m)|\hat{\rho}_m))=-2E_{\hat{\rho}_m}\left(E_{m^*}\left(\sum_{i=1}^{4^N}\xi_i\ln\hat{p}_{mi}|\hat{\rho}_m\right)\right). \tag{8}$$

The second equality follows from (2), and $\hat{p}_{mi}=\hat{p}_{mi}(\hat{\rho}_m)$ is the probability of observing the i^{th} nucleotide pattern based on the maximum likelihood estimates $\hat{\rho}_m$. The inner expectation in (8) is

$$E_{m^*}\left(\sum_{i=1}^{4^N}\xi_i\ln\hat{p}_{mi}|\hat{\rho}_m\right)=\sum_{i=1}^{4^N}E_{m^*}(\xi_i)\ln\hat{p}_{mi}=\sum_{i=1}^{4^N}Mp_{*i}\ln\hat{p}_{mi}, \tag{9}$$

in which p_{*i} is the probability of observing the i^{th} nucleotide pattern under the true model m^* and M is the sequence length. Thus,

$$E(KL)=-2E_{m^*}\left(\sum_{i=1}^{4^N}Mp_{*i}\ln\hat{p}_{mi}\right). \tag{10}$$

The quantity in (10) suggests that we can select a model to minimize the expected KL information value.

Lemma 3.1 *Given the tree topology τ, AIC is an asymptotically unbiased estimator of the expected KL divergence for phylogenetic models under the standard condition (see Remark).*

Proof To derive the expected KL divergence, we first approximate the summation $\sum_{i=1}^{4^N}Mp_{*i}\ln\hat{p}_{mi}$ in equation (10), in which the probability $\hat{p}_{mi}$ is a function of parameters ρ_m. Let $\tilde{\rho}_m$ be the values of ρ_m that maximize $L_m(\rho_m)=\sum_{i=1}^{4^N}Mp_{*i}\ln p_{mi}$. It follows that

$$\frac{\partial L_m(\rho_m)}{\partial\rho_m}|_{\rho_m=\tilde{\rho}_m}=L_m'(\tilde{\rho}_m)=0. \tag{11}$$

When the sequence length $M\to\infty$, the frequencies of 4^N patterns under the true model m^* converge to Mp_{*i} for $i=1, \ldots, 4^N$. Thus $\hat{\rho}_m$ (the maximum likelihood estimates of parameters ρ_m for the finite data) converges to $\tilde{\rho}_m$ almost surely, as $M\to\infty$. Because the second-order partial derivatives of $L_m(\rho_m)$ with respect to parameters ρ_m exist and are continuous (see Appendix A1), we take the second order Taylor series approximation of $L_m(\hat{\rho}_m)=\sum_{i=1}^{k}Mp_{*i}\ln\hat{p}_{mi}$ around $\tilde{\rho}_m$ to obtain

$$L_m(\hat{\rho}_m) \approx L_m(\tilde{\rho}_m) + (\hat{\rho}_m - \tilde{\rho}_m) L'_m(\tilde{\rho}_m) + \frac{1}{2}(\hat{\rho}_m - \tilde{\rho}_m) L''_m(\tilde{\rho}_m)(\hat{\rho}_m - \tilde{\rho}_m)^t. \quad (12)$$

By (11), the first derivative $L'_m(\tilde{\rho}_m)$ is equal to 0. The quadratic term converges to a $-\chi^2$ random variable with $\lambda_m = |\rho_m|$ degrees of freedom. Thus the expected KL can be approximated by

$$E(KL) = -2E_{m^*}(L_m(\hat{\rho}_m)) \approx \lambda_m - 2E_{m^*}(L_m(\tilde{\rho}_m)). \quad (13)$$

We next approximate $L_m(\tilde{\rho}_m) = \sum_{i=1}^{k} Mp_{*i} \ln \tilde{p}_{mi}$, in which probability $\tilde{p}_{mi}$ is a function of parameters $\tilde{\rho}_m$. Because the expectation of the frequency of site pattern ξ_i in data D is Mp_{*i}, we have

$$L_m(\tilde{\rho}_m) = E\left(\sum_{i=1}^{k} \xi_i \ln \tilde{p}_{mi}\right). \quad (14)$$

Moreover, we take the second order Taylor approximation of $L_m(D|\tilde{\rho}_m) = \sum_{i=1}^{k} \xi_i \ln \tilde{p}_{mi}$ around the maximum likelihood estimates $\hat{\rho}_m$ of parameters ρ_m for data D,

$$L_m(D|\tilde{\rho}_m) \approx L_m(D|\hat{\rho}_m) + (\tilde{\rho}_m - \hat{\rho}_m) L'_m(\hat{\rho}_m) + \frac{1}{2}(\tilde{\rho}_m - \hat{\rho}_m) L''_m(\hat{\rho}_m)(\tilde{\rho}_m - \hat{\rho}_m)^t. \quad (15)$$

At its maximum, $L'_m(\hat{\rho}_m) = 0$ and the linear term in (15) vanishes. When the sample size (i.e., sequence length) M is large, we have $L''_m(\hat{\rho}_m) \approx L''_m(\tilde{\rho}_m)$, and the quadratic term converges to a $-\chi^2$ random variable with $\lambda_m = |\rho_m|$ degrees of freedom, i.e.,

$$-2E_{m^*}(L_m(D|\tilde{\rho}_m)) \approx \lambda_m - 2E_{m^*}(L_m(D|\hat{\rho}_m)). \quad (16)$$

Combining (13) and (16), we have

$$E(KL) \approx 2\lambda_m - 2E_{m^*}(L_m(D|\hat{\rho}_m)) = E(AIC_m). \quad (17) \blacksquare$$

Remark Approximation to $E(KL)$ is often obtained through a Taylor series expansion of the log-likelihood function L_m around the true parameter values $\tilde{\rho}_m$

$$L_m(\hat{\rho}_m) \approx L_m(\tilde{\rho}_m) + (\hat{\rho}_m - \tilde{\rho}_m) L'_m(\tilde{\rho}_m) + \frac{1}{2}(\hat{\rho}_m - \tilde{\rho}_m) L''_m(\tilde{\rho}_m)(\hat{\rho}_m - \tilde{\rho}_m)^t.$$

Self and Liang (1987) have shown that when the true parameter values lie on the boundary of the parameter space, the quadratic term is approximated by a mixture of χ^2 distributions. This result indicates that when some branch lengths in the phylogenetic tree are 0, the quadratic term may not have a χ^2 distribution with degrees of freedom equal to the number of parameters in the model.

Given a phylogenetic tree with k branches, let x_i be the number of substitutions on branch i. Under the substitution model, x_i is a Poisson random variable with parameter b_i, for i=1, 2, ⋯, k, in which $b_i \geq 0$ is the length of branch i (i.e., the expected number of substitutions on branch i). Assuming that substitution rates on different branches are independent, the joint distribution for $x=(x_1, x_2, \cdots, x_{k)}$ is $f(x|b) = \prod_{i=1}^{k} \frac{e^{-b_i} b_i^{x_i}}{x_i!}$ where $b=(b_1, b_2, \cdots, b_k) \in \mathbb{R}_+^k$. The first order partial derivative of the log-likelihood function, $\log L(b|x) = \sum_{i=1}^{k} (-b_i + x_i \log b_i - \log x_i!)$, with respect to b_i is given by $\frac{\partial L}{\partial b_i} = -1 + \frac{x_i}{b_i}$, for i=1, 2, ⋯, k. The maximum likelihood estimator of b_i is $\hat{b}_i = x_i$, $1 \leq i \leq k$.

The second order partial derivative (the Hessian matrix) of log L is a k by k diagonal matrix with $-\frac{x_i}{b_i^2}$ on the diagonal $\left(\text{i.e.,} \frac{\partial^2 \log L}{\partial b^2} = -diag\left[\frac{x_1}{b_1^2}, \frac{x_2}{b_2^2}, \cdots, \frac{x_k}{b_k^2}\right]\right)$. The Fisher information matrix is

$I(b)=-E\left[\frac{\partial^2}{\partial b^2}\log L(x;b)|b\right]=diag\left[\frac{1}{b_1},\frac{1}{b_2},\cdots,\frac{1}{b_k}\right]$. When no substitutions occur on some branches (i.e., $x_i=0$ for some i), the MLEs of the corresponding branch lengths $b_s=(b_{i1}, b_{i2}, \cdots, b_{is})$ are 0, which is the boundary of the parameter space of branch lengths b. In this case, the Fisher information matrix $I(b_s)=-E\left[\frac{\partial^2}{\partial b_s^2}\log L(x_s;b_s)|b_s\right]$ for branches b_s is a zero matrix. The quadratic term in (11) has an asymptotic χ^2 distribution with $(|\rho_m|-|b_s|)$ degrees of freedom. Thus, the distribution of the quadratic term depends on $|b_s|$, the number of branches on which no substitutions occur. Suppose we denote the random variable $|b_s|$ by y. Considering all possible values of $|b_s|$, the distribution of the quadratic term in (12) is a mixture of χ^2 distributions with various degrees of freedom, i.e., $\sum_{i=0}^{|b|}\{P(y=i)\times\chi^2_{|\rho_m|-i}\}$, in which $|\rho_m-i|$ represents degrees of freedom of the χ^2 distribution.

The boundary problem may also affect BIC. As the branch length b_i in the phylogenetic tree is nonnegative, the MLE $\hat{b}_i$ of b_i has asymptotically a truncated normal distribution, i.e., $\hat{b}_i\geq 0$. The size of truncation depends on the value of b_i. When the branch length b_i is close to 0, almost half of the asymptotic normal distribution is truncated, and thus the distribution of MLE $\hat{b}_i$ is no longer asymptotically normal. Since BIC relies on a normal approximation, it needs a correction for short branches, because the normal approximation does not hold for the MLE $\hat{b}_i$ when b_i is close to 0. As long as all internal branches receive more than five substitutions ($b_i>5$), the χ^2 approximation is valid; we call this the standard condition. ■

Thus, under the standard condition, AIC is an unbiased estimator of the expected KL when the tree topology is fixed. It shows that the use of AIC or similar procedures on a given tree is expected to give an unbiased estimate of the expected KL divergence of a phylogenetic model. However, this only holds with a fixed, true tree, as shown in the next lemma. If the tree topology is estimated from data D, it should be counted as a model parameter when calculating K (the number of parameters) in AIC. However, it is difficult to count the exact number of parameters for the tree topology. In practice, the tree topology is often omitted when calculating K in AIC, so that the resulting AIC without taking into account the effect of tree topology is a biased estimator of $E(KL)$; we show this in the following lemma.

Lemma 3.2 *If the tree topology τ is estimated from data, $E(AIC)<E(KL)$, even under the standard condition. Hence AIC underestimates $E(KL)$.*

Proof If the tree topology is estimated from data D, we need to take the expectation with respect to the MLE of the tree topology when calculating $E(KL)$, i.e.,

$$E(KL)=-2E_{\hat{\tau}_m,\hat{\rho}_m}(E_{m^*}(L_m(D|\hat{\rho}_m)|\hat{\rho}_m,\hat{\tau}_m)). \tag{18}$$

The MLE $\hat{\tau}_m$ of the tree topology is a discrete random variable with a probability distribution $P(\hat{\tau}_{mj})=w_{mj}$, where $0\leq w_{mj}\leq 1$, $\sum_{j=1}^{s}w_{mj}=1$, and $s=\frac{(2j-5)!}{2^{j-3}(j-3)!}$. The expected KL divergence is expressed as

$$E(KL)=-2\sum_{j=1}^{s}w_{mj}E_{\hat{\rho}_m}\left(\sum_{i=1}^{4^N}Mp_{*i}\ln\hat{p}_{mij}|\hat{\tau}_{mj},\hat{\rho}_{mj}\right). \tag{19}$$

In equation (19), the probability $\hat{p}_{mij}$ is a function of $\hat{\tau}_{mj}$ and $\hat{\rho}_{mj}$. It follows from Lemma 3.1 that given the tree topology $\hat{\tau}_{mj}$, the conditional expectation in equation (19) can be approximated by

$$-2E_{\hat{\rho}_m}\left(\sum_{i=1}^{4^N}Mp_{*i}\ln\hat{p}_{mij}|\hat{\tau}_{mj},\hat{\rho}_{mj}\right)\approx\lambda_m-2\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mij}, \tag{20}$$

in which $\tilde{p}_{mij}$ is a function of parameters $\tilde{\rho}_{mj}$ that maximize $\sum_{i=1}^{4^N}Mp_{*i}\ln p_{mij}$, given the tree topology $\hat{\tau}_{mj}$. Thus, the expected KL divergence becomes

$$E(KL)=\lambda_m-2\sum_{j=1}^{s}\left(w_{mj}\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mij}\right). \tag{21}$$

Let $\tilde{\tau}_m$ and $\tilde{\rho}_m$ be the tree topology and the values of parameters ρ_m that maximize $L_m(\rho_m,\tau)=\sum_{i=1}^{4^N}Mp_{*i}\ln p_{mi}$ for model m, in which probability p_{mi} is a function of τ_m and ρ_m. Let $\tilde{p}_{mi}$ be the probability of pattern i calculated from $\tilde{\tau}_m$ and $\tilde{\rho}_m$. Thus, $\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mij}<\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mi}$, if $\hat{\tau}_{mj}\neq\tilde{\tau}_m$. Here we assume that the tree topology is identifiable and thus $\tilde{\tau}_m$ is unique. It follows that

$$E(KL)>\lambda_m-2\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mi}. \tag{22}$$

With a fixed topology $\tilde{\tau}$, summation $\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mi}$ in equation (22) can be approximated by (see Lemma 3.1)

$$-2\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mi}\approx\lambda_m-2E\left(\sum_{i=1}^{4^N}\xi_i\ln\check{p}_{mi}\right), \tag{23}$$

in which $\check{p}_{mi}$ is the probability of pattern i based on parameters $\check{\rho}_{mj}$ that maximize $\sum_{i=1}^{4^N}\xi_i\ln p_{mi}$, given topology $\tilde{\tau}_m$. Moreover, let $\hat{p}_{mi}$ be the estimates of p_{mi} that maximize $\sum_{i=1}^{4^N}\xi_i\ln p_{mi}$ over all possible tree topologies. Because $\sum_{i=1}^{4^N}\xi_i\ln\check{p}_{mi}<\sum_{i=1}^{4^N}\xi_i\ln\hat{p}_{mi}$, we have

$$-2\sum_{i=1}^{4^N}Mp_{*i}\ln\tilde{p}_{mi}>\lambda_m-2E\left(\sum_{i=1}^{4^N}\xi_i\ln\hat{p}_{mi}\right). \tag{24}$$

Note that $\sum_{i=1}^{4^N}\xi_i\ln\hat{p}_{mi}$ is the log likelihood. Combining (22) with (24), we have shown that $E(KL)>E(AIC)$ ■.

Although we have shown that AIC underestimates the expected KL divergence, it is difficult to quantitatively assess the degree of underestimation under different substitution models. Next, we use simulations to shed some light on this, and conclude that the scale of underestimation varies among substitution models.

3.2 Assessing the performance of AIC in selecting phylogenetic models

Our assessment of the performance of AIC and the other criteria consisted of two studies. In the first study, we use simulations to evaluate the amount of underestimation, assuming that for a set of phylogenetic trees, the topology is unknown when counting the model parameters in AIC. If the degree of underestimation remains the same across phylogenetic models, it will not affect the model selected by AIC without a correction for underestimation. Otherwise, ignorance of underestimation will introduce extra errors when using AIC to select phylogenetic models. In the second study, we evaluate the performance of AIC in selecting the correct phylogenetic model, and we also investigate the relationship between model selection and tree estimation.

For the first study, simulations were conducted with the following experimental design. We assumed two true substitution models, HKY and GTR, and for each, we generated 10,000 data sets from an eight taxon tree. For each data set, we then calculated AIC (and AICc) and compared their expected versions with the expected KL divergence for each of 24 substitution models. This study gives us an indication of how the bias of AIC and AICc varies across different phylogenetic models.

Specifically, we used the phylogenetic program Seq-Gen (Rambaut and Grassly, 1997) to simulate the 10,000 data sets, and the 8-taxon tree has characteristics given by setting of three different branch lengths: 0.002, 0.005, 0.01 with each data set consisting of DNA sequences of 1000 base pairs; thus the expected

Table 1 Bias of AIC and AICc in estimating E(KL) for branch length 2.

Model	K	HKY			GTR		
		E(KL)	Bias (AIC)	Bias (AICc)	E(KL)	Bias (AIC)	Bias (AICc)
JC	14	−5663.28	10.93	10.72	−5728.55	8.73	8.52
JC+G	15	−5663.54	9.92	9.68	−5728.18	7.33	7.09
JC+I	15	−5663.41	9.55	9.31	−5728.25	7.29	7.05
JC+I+G	16	−5663.56	8.21	7.93	−5728.18	6.05	5.77
K80	15	−5365.13	5.61	5.37	−5721.55	8.78	8.54
K80+G	16	−5364.98	4.64	4.36	−5721.29	7.57	7.29
K80+I	16	−5365.22	4.89	4.61	−5721.28	7.38	7.1
K80+I+G	17	−5365.15	3.79	3.48	−5721.24	6.21	5.9
F81	17	−5590.25	10.26	9.95	−5531.75	7.5	7.19
F81+G	18	−5590.32	9.14	8.79	−5531.42	6.39	6.04
F81+I	18	−5590.27	8.93	8.58	−5531.45	6.33	5.98
F81+I+G	19	−5590.36	7.7	7.31	−5531.43	5.28	4.89
HKY	18	−5324.19	5.19	4.84	−5528.31	7.47	7.12
HKY+G	19	−5324.09	4.27	3.88	−5528.05	6.45	6.06
HKY+I	19	−5324.28	4.46	4.07	−5528.02	6.31	5.92
HKY+I+G	20	−5324.21	3.36	2.93	−5528.04	5.34	4.91
SYM	19	−5356.52	5.64	5.25	−5578.48	7.48	7.09
SYM+G	20	−5356.43	4.77	4.34	−5578.29	7.05	6.62
SYM+I	20	−5356.62	4.94	4.51	−5578.19	6.57	6.14
SYM+I+G	21	−5356.6	3.94	3.47	−5578.42	6.19	5.72
GTR	22	−5326.24	5.25	4.73	−5461.52	6.15	5.63
GTR+G	23	−5326.16	4.33	3.76	−5461.35	5.14	4.57
GTR+I	23	−5326.34	4.53	3.96	−5461.54	5.32	4.75
GTR+I+G	24	−5326.31	3.46	2.84	−5461.51	4.26	3.64

The sequence data were simulated from tree (((S1:0.5, S2:0.002):0.002, (S3:0.5, S4:0.002):0.002):0.002, ((S1:0.5, S2:0.002):0.002, (S3:0.5, S4:0.002):0.002):0.002) under the HKY and GTR models. The expected AIC and AICc of 24 substitution models were calculated from the simulated data. *Bias*(*AIC*)=*E*(*KL*)−*E*(*AIC*).

number of substitutions (b_i) is 0.002×1000=2, 0.005×1000=5, 0.01×1000=10, respectively. Recall that $b_i>5$ is the standard condition for validity of the χ^2 approximation. Specifically, we have results in Tables 1 and 2 from the two examples from non-standard conditions, (((S1:0.5, S2:0.002):0.002, (S3:0.5, S4:0.002):0.002):0.002, ((S5:0.5, S6:0.002):0.002, (S7:0.5, S8:0.002):0.002):0.002) and same with 0.005 replacing 0.002 and 0.5. In Table 3 we have results from an example from a standard condition scenario where 0.002 and 0.5 are replaced by 0.01. The parameters in the HKY model include base frequencies (π_A=0.2, π_C=0.3, π_G=0.2, π_T=0.3) and rate parameters (r_{AC}=1, r_{AG}=4.78, r_{AT}=1, r_{CG}=1, r_{CT}=4.78, r_{GT}=1), while the GTR model had base frequencies (π_A=0.32, π_C=0.3, π_G=0.11, π_T=0.27) and the rate parameters (r_{AC}=2.7, r_{AG}=1.3, r_{AT}=4.5, r_{CG}=5.7, r_{CT}=3.6, r_{GT}=1). The simulated data were then used to calculate E(KL) and E(AIC). We considered 24 candidate models (JC, JC+G, JC+I, JC+I+G, F81, F81+G, F81+I, F81+I+G, K80, K80+G, K80+I, K80+I+G, HKY, HKY+G, HKY+I, HKY+I+G, SYM, SYM+G, SYM+I, SYM+I+G, GTR, GTR+G, GTR+I, GTR+I+G), and values of AIC and AICc were calculated for each candidate model in jModelTest. It follows from equation (10) that the expected KL for a candidate model is

$$E(KL)=-2\sum_{i=1}^{4^N} Mp_{*i}E_{\hat{\rho}_m}(\ln\hat{p}_{mi}). \tag{25}$$

Because true model m^* is known, p_{*i} in (25) can be calculated under m^* and we need only to estimate the term $E_{\hat{\rho}_m}(\ln\hat{p}_i)$. By the law of large numbers, the expectation $E_{\hat{\rho}_m}(\ln\hat{p}_i)$ can be estimated by $\frac{1}{c}\sum_{l=1}^{c}(\ln\hat{p}_i^l)$, where l=1, …, c indexes the simulated "true" data sets. It suggests that the expected KL divergence can be estimated by

Table 2 Bias of AIC and AICc in estimating E(KL) for branch length 5.

Model	K	HKY			GTR		
		E(KL)	Bias (AIC)	Bias (AICc)	E(KL)	Bias (AIC)	Bias (AICc)
JC	14	−1902.33	3.12	2.69	−1902.39	2.58	2.15
JC+G	15	−1902.43	2.41	1.92	−1902.50	1.91	1.42
JC+I	15	−1902.37	2.17	1.68	−1902.41	1.63	1.14
JC+I+G	16	−1902.42	1.26	0.70	−1902.48	0.74	0.18
K80	15	−1880.98	2.82	2.33	−1902.56	2.65	2.16
K80+G	16	−1881.03	2.10	1.54	−1902.68	1.98	1.43
K80+I	16	−1881.07	2.13	1.58	−1902.57	1.69	1.14
K80+I+G	17	−1881.00	1.06	0.44	−1902.65	0.81	0.19
F81	17	−1880.21	3.07	2.45	−1832.37	2.59	1.97
F81+G	18	−1880.31	2.38	1.69	−1832.48	1.93	1.24
F81+I	18	−1880.24	2.13	1.43	−1832.37	1.64	0.94
F81+I+G	19	−1880.29	1.22	0.44	−1832.44	0.75	−0.02
HKY	18	−1861.03	2.81	2.11	−1832.70	2.65	1.96
HKY+G	19	−1861.07	2.08	1.30	−1832.82	2.00	1.22
HKY+I	19	−1861.11	2.12	1.34	−1832.70	1.70	0.92
HKY+I+G	20	−1861.04	1.05	0.19	−1832.77	0.82	−0.04
SYM	19	−1882.98	3.84	3.07	−1892.26	3.71	2.93
SYM+G	20	−1883.04	3.13	2.27	−1892.31	2.99	2.13
SYM+I	20	−1883.03	3.12	2.26	−1892.26	2.94	2.08
SYM+I+G	21	−1882.99	2.08	1.13	−1892.25	1.93	0.98
GTR	22	−1863.93	3.82	2.78	−1827.84	3.87	2.83
GTR+G	23	−1863.99	3.10	1.97	−1827.89	3.13	2.00
GTR+I	23	−1863.98	3.09	1.96	−1827.82	3.06	1.93
GTR+I+G	24	−1863.95	2.05	0.82	−1827.82	2.06	0.83

The sequence data were simulated from tree (((S1:0.005, S2:0.005):0.005, (S3:0.005, S4:0.005):0.005):0.005, ((S5:0.005, S6:0.005):0.005, (S7:0.005, S8:0.005):0.005):0.005) under the HKY and GTR models. The expected AIC and AICc of 24 substitution models were calculated from the simulated data. $Bias(AIC)=E(KL)-E(AIC)$.

$$E(KL)\approx -2\sum_{i=1}^{4^N}\left(\frac{1}{c}Mp_{*i}\sum_{l=1}^{c}(\ln\hat{p}_i^l)\right). \tag{26}$$

Specifically, we fitted the simulated data to the candidate models m=1, …, 24 and calculated the probability $\hat{p}_i^l$ of observing nucleotide pattern i based on the ML estimates of the parameters obtained for the simulated dataset l. Finally, the expected KL divergence was estimated by (26) (see Figure A2 in Appendix A2).

For the non-standard condition examples, the results indicate that both AIC and AICc underestimate the expected KL divergence (Tables 1 and 2), as expected by Lemma 3.2. The scale of underestimation varies among substitution models and by b_i (Tables 1 and 2). Overall, the underestimation for b_i=5 is smaller than that for b_i=2. For both branch lengths, the bias of AIC and AICc for the models with gamma parameter G or invariant proportion parameter I tends to be smaller than that for the models without parameters G and I (Table 1 and 2). In addition, the bias tends to increase as the target model moves away from the true model (in terms of the number of parameters in the model). In general, the over-simplified models have a larger bias than the over-parameterized models. Using b_i=2 as an example (Table 1), when the true model is HKY, the bias of AIC for the true model is 5.19, while the bias for GTR is 5.25. In contrast, the bias for JC is 10.93, which is significantly larger than the bias for GTR. Regardless of the type of the target model, AICc appears to have a smaller bias than AIC. Because AICc is AIC with a correction for finite sample sizes, the difference between the biases of AIC and AICc is the amount of bias of AIC due to finite sample sizes. After the correction for finite sample sizes, the amount of bias (bias of AICc in Table 1) is still substantial, i.e., the amount of bias due to the estimated tree topology is significantly larger than the amount of bias due to finite sample sizes. It indicates

Table 3 Bias of AIC and AICc in estimating E(KL) for branch length 10.

Model	K	HKY			GTR		
		E(KL)	Bias (AIC)	Bias (AICc)	E(KL)	Bias (AIC)	Bias (AICc)
JC	14	−2274.69	−0.78	−1.21	−2303.23	0.45	0.03
JC+G	15	−2274.74	−1.54	−2.03	−2303.27	−0.30	−0.79
JC+I	15	−2275.05	−1.56	−2.04	−2303.59	−0.31	−0.80
JC+I+G	16	−2275.03	−2.58	−3.13	−2303.55	−1.35	−1.90
K80	15	−2239.29	−0.32	−0.81	−2303.02	0.50	0.01
K80+G	16	−2239.35	−1.02	−1.58	−2303.05	−0.25	−0.81
K80+I	16	−2239.50	−0.87	−1.43	−2303.37	−0.26	−0.82
K80+I+G	17	−2239.43	−1.94	−2.56	−2303.34	−1.29	−1.92
F81	17	−2237.02	−0.65	−1.27	−2226.03	0.53	−0.09
F81+G	18	−2237.07	−1.36	−2.06	−2226.08	−0.19	−0.88
F81+I	18	−2237.25	−1.19	−1.88	−2226.35	−0.23	−0.92
F81+I+G	19	−2237.17	−2.26	−3.04	−2226.32	−1.26	−2.03
HKY	18	−2300.13	−0.73	−1.43	−2226.10	0.55	−0.15
HKY+G	19	−2300.17	−1.52	−2.30	−2226.15	−0.17	−0.95
HKY+I	19	−2300.52	−1.51	−2.29	−2226.42	−0.21	−0.99
HKY+I+G	20	−2300.50	−2.54	−3.40	−2226.39	−1.24	−2.10
SYM	19	−2258.21	−0.57	−1.35	−2279.30	1.22	0.44
SYM+G	20	−2258.26	−1.29	−2.14	−2279.35	0.53	−0.33
SYM+I	20	−2258.43	−1.11	−1.97	−2279.48	0.66	−0.19
SYM+I+G	21	−2258.35	−2.19	−3.13	−2279.41	−0.40	−1.34
GTR	22	−2258.66	−0.29	−1.32	−2212.97	1.27	0.23
GTR+G	23	−2258.73	−0.98	−2.11	−2213.01	0.55	−0.58
GTR+I	23	−2258.88	−0.83	−1.96	−2213.14	0.67	−0.46
GTR+I+G	24	−2258.81	−1.89	−3.12	−2213.08	−0.38	−1.61

The sequence data were simulated from tree (((S1:0.01, S2:0.01):0.01, (S3:0.01, S4:0.01):0.01):0.01, ((S5:0.01, S6:0.01):0.01, (S7:0.01, S8:0.01):0.01):0.01) under the HKY and GTR models. The expected AIC and AICc of 24 substitution models were calculated from the simulated data. $Bias(AIC)=E(KL)-E(AIC)$.

that the correction for the estimated tree topology is more important than the correction for finite sample sizes in the context of selecting phylogenetic models.

For the example from the standard condition, b_i=10, the trees estimated from the simulated data are identical with the true tree. Thus the conditions in Lemma 3.1 are satisfied. As expected form Lemma 3.1, the biases of AIC are close to 0 for the models without gamma (G) and invariant site (I) parameters (Table 3), and they generally hover in the interval (–1,0) for the GTR model and a bit lower for HKY. The bias of AIC and AICc for the models with gamma parameter *G* or invariant proportion parameter *I* tends to be smaller than that for the models without parameters *G* and *I* (Table 3).

For our second simulation study, to investigate the performance of AIC and other criteria in selecting phylogenetic models, we simulated DNA sequences from two 4-taxon phylogenetic trees (Figure 1). The first tree (tree A in Figure 1) has a relatively long internal branch, while the second tree (tree B) has a short internal branch. We again simulated 10,000 data sets from trees A and B (Figure 1) under the JC, HKY and GTR models, respectively, using Seq-Gen (Rambaut and Grassly, 1997). The parameters in the HKY model include base frequencies (π_A=0.2, π_C=0.3, π_G=0.2, π_T=0.3) and rate parameters (r_{AC}=1, r_{AG}=4.78, r_{AT}=1, r_{CG}=1, r_{CT}=4.78, r_{GT}=1). For the GTR model, the base frequencies are (π_A=0.32, π_C=0.3, π_G=0.11, π_T=0.27) and the rate parameters are (r_{AC}=2.7, r_{AG}=1.3, r_{AT}=4.5, r_{CG}=5.7, r_{CT}=3.6, r_{GT}=1). The simulated data were then used to evaluate the performance of AIC, AICc, and BIC. The values of AIC, AICc, and BIC were calculated for each candidate model in jModelTest. In addition, we calculated the bias of AIC using simulation as in the first study, and then calculated the bias-corrected AIC (BiasC), i.e., $BiasC=AIC+bias$. Thus, BiasC represents an unbiased selection criterion for phylogenetic models. Finally, we calculated the proportion of datasets for which the true tree was

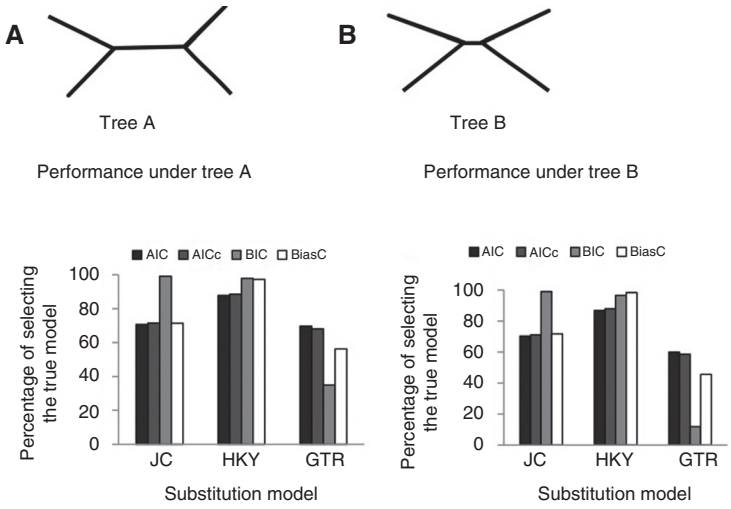

Figure 1 Performance of AIC, AICc, BIC, and BiasC in selecting phylogenetic models. The sequence data were simulated under the true model (JC, HKY, GTR), respectively from (A) tree A: (S1:0.01, S2:0.01, (S3:0.01, S4:0.01):0.01), which has a relatively long internal branch, or from (B) tree B: (S1:0.01, S2:0.01, (S3:0.01, S4:0.01):0.001), which has a short internal branch. The data simulated from trees A and B were used to estimate phylogenetic trees using the model selected by jModelTest.

successfully recovered, given that AIC selected the true or wrong model. This may illustrate whether model selection can affect the accuracy of phylogenetic tree estimation.

The performance of AIC, AICc, BIC, and BiasC depends on the true model and the true phylogenetic tree. When the true model is JC, BIC appears to perform best among the 4 model selection criteria (Figure 1A and B). Because BIC imposes more penalty on the number of parameters than AIC, AICc, and BiasC, it tends to select simpler models such as the JC model. Thus BIC performs much better than the other three criteria when the true model is JC. In contrast, when the true model is GTR, the percentage for BIC selecting the correct model drops to 35% for treeA and 12% for treeB. The poor performance of BIC for selecting complex phylogenetic models indicates that BIC over-penalizes the number of parameters and tends to be biased towards selecting simpler phylogenetic models. The performance of AIC and AICc is similar to that of BiasC when the true model is JC. In addition, AIC and AICc tend to select more complex models compared to BiasC. The bias of AIC and AICc, however, is not as serious as that of BIC.

In the second simulation, BiasC does not perform best among the four model selection criteria. This is because the other three criteria are based on the biased estimates of the expected KL divergence. When the true model happens to be the one favored by a biased criterion, it is more likely for the biased criterion to select the true model. In this case, the unbiased criterion BiasC may perform worse than the biased model selection criteria. As the performance of a selection criterion also depends on the variance of the log likelihood, an unbiased selection criterion, such as BiasC, may perform poorly when the variance of the log likelihood is large. For example, the chance of BiasC selecting the true model is less than 60% when the true model is GTR (Figure 1a–b). It suggests that improvement in accuracy of a selection criterion may be achieved by reducing the variance of the estimate of E(KL) on which the selection criterion is based. We will discuss this issue further in the following section.

4 Discussion and future directions

Our theoretical derivations and simulations highlight some deficiencies of AIC for selection of substitution models as well as phylogenetic models. When the tree topology is known, AIC is unbiased but the large variance results in its poor performance in practice. When the topology of a phylogenetic tree is unknown and estimated from sequence data, the most frequently used selection criteria, including AIC and AICc, are biased

estimates of the expected KL divergence. As such, we suggest that there need to be more appropriate model selection criteria tailored for use in phylogenetics.

Our simulation study suggests that AIC and AICc are biased estimators of the expected KL divergence for phylogenetic models, but the scale of the bias is intractable unless the true model is given. For real data analysis, where the true model m^* is unknown, we suggest the following in an attempt to develop a new selection criterion based on an unbiased estimate of the expected KL divergence. In equation (8), the expected KL divergence is expressed as the sum of the products of the true frequencies and the expected probabilities under the candidate model. It implies that the expected KL divergence could be estimated by a bootstrap technique, in which the true model m^* is replaced by the empirical distribution $\hat{F}_{m^*}$. The bootstrap version of information criterion for a general setting was introduced by Ishiguro et al. (1997). The expectation of log likelihood L_m based on the empirical distribution $\hat{F}_{m^*}$ is given by

$$E_{m^*}(L_m) \approx E_{\hat{m}^*}(L_m) = \int \left(\sum_{i=1}^{4^N} M\hat{p}_i \ln \hat{p}_i \right) d\hat{F}_{m^*}. \tag{27}$$

The integrand $\sum_{i=1}^{4^N} M\hat{p}_i \ln \hat{p}_i$ is the log likelihood L_m under a candidate model m. Equation (27) suggests that we could use bootstrap to produce a better estimate for the expected log likelihood. The idea is as follows: generate c bootstrap samples from the empirical distribution $\hat{F}_{m^*}$. For each bootstrap sample, we calculate the log likelihood score. The expected log likelihood is then estimated by

$$E_{\hat{m}^*}(L_m) \approx \frac{M}{c} \sum_{j=1}^{c} \sum_{i=1}^{k} \hat{p}_i^j \ln \hat{p}_i^j. \tag{28}$$

The performance of a selection criterion is determined by not only the bias of the estimate of E(KL), but also the variance of the estimate (i.e., the variance of log likelihood). In the simulation study, the error rates of BiasC with correction for the bias of AIC are still greater than 45% when the true model is GTR (Figure 1A and B). The large error rate of BiasC indicates that log likelihood has a large variance. Thus accuracy of a selection criterion for phylogenetic models can be improved by developing an estimate of E(KL) with a smaller variance.

In practice, we should remain cautious when using AIC (or AICc) as a criterion for selecting the best phylogenetic model, even though given the tree topology, AIC is an unbiased estimator of the expected KL divergence for phylogenetic, in this case, substitution models. As shown in the previous section, AIC underestimates the expected KL divergence if the tree topology is estimated from data but not counted as a model parameter in calculating AIC. As the scale of underestimation depends on the target model, the bias of AIC as a model selection criterion may result in selecting a wrong model even when the sample size is large. The theory developed in this paper has laid foundation for further studying the statistical properties of AIC and other model selection criteria based on the Kullback-Liebler divergence. While we sketched out a bootstrapping idea, further work is underway to assess its feasibility and performance.

Acknowledgments: We thank David Posada for the helpful discussion on phylogenetic model selection. We thank Diego Darriba for his generous help with implementing the phylogenetic model selection program jModelTest. Jhwueng's research was supported by the National Science Council Award #NSC-101-2118-M-035-001 Taiwan and Postdoctoral Fellowship at the National Institute for Mathematical and Biological Synthesis (NIMBioS), an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Awards #EF-0832858 and #DBI-1300426, with additional support from The University of Tennessee, Knoxville. We also thank NIMBioS Working Group-Gene Tree/Species Tree Reconciliation for providing the opportunity that the authors of this paper could meet to discuss this project. This research was partially supported by the National Science Foundation (DMS-1222745 to LL). Huzurbazar's research was supported by a grant to the University of Wyoming from the National Science Foundation under grant DMS-1100615. Huzurbazar's contribution was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix A

A1 The likelihood function of general substitution models

Given the DNA sequence data D of size $N\times M$ where each element of D, D_{uv} represents the v^{th} (v=1, …, M) nucleotide of sequence u (u=1, …, N). There are 4^N possible nucleotide patterns for each column in data D. Let p_i be the probability of observing the i^{th} nucleotide pattern. The probability p_i is the summation of the product of transition probabilities $\{P(t)=P_{ij}(t); i, j=A, C, G, T\}$ (see Figure A1). Thus, $p_i=p_i(\phi)$ is a function of model parameters $\phi=(\theta, \tau, b)$ where θ represents the parameters in the substitution model (typically the rate matrix Q, and the equilibrium vector $\pi=(\pi_A, \pi_C, \pi_G, \pi_T)$). When substitution rates are variable over sites, the heterogeneity of rates such as invariant parameter I, and the gamma rate parameter γ will be included in the substitution models. The tree topology τ and its branch lengths b are also treated as parameters for phylogenetic tree estimation. The rate matrix, Q, describes the rate at which bases of one type change into bases of another type. The rate matrix has the following structure

$$Q=\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & B & G & T \end{array} \\ \begin{pmatrix} -\mu_A & \mu_{CA} & \mu_{GA} & \mu_{TA} \\ \mu_{AC} & -\mu_C & \mu_{GC} & \mu_{TC} \\ \mu_{AG} & \mu_{CG} & -\mu_G & \mu_{TG} \\ \mu_{AT} & \mu_{CT} & \mu_{GT} & -\mu_T \end{pmatrix} \end{array}$$

where μ_{xy} represents the transition rate from base x to base y and the diagonals of the matrix are chosen so that the rows sum to zero: $\mu_x=-\Sigma_{\{y|y\neq x\}}\mu_{xy}$. The equilibrium row vector π must satisfy $\pi Q=0$. Let $P(t)$ be the transition probability matrix, in which $P_{xy}(t)$ is the probability of base x changing to y after a period of time t. Since this is a continuous time Markov Chain, the transition probability matrix satisfies a first order ordinary system differential equation $P'(t)=QP(t)$, to which the solution is $P(t|Q)=e^{Qt}$. The substitution models consid-

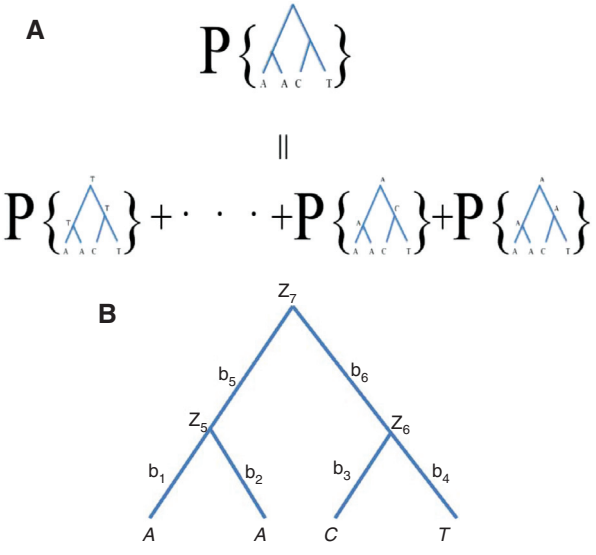

Figure A1 The probability density function $p(\cdot)$ of a column of nucleotides given the phylogenetic tree. (A) The column of nucleotides (A, A, C, T) are at the tips of the tree. The probability of (A, A, C, T) given the tree (on the top) is equal to the sum of the probabilities at the bottom, in which the nucleotides at the internal nodes of the tree are given. Each probability at the bottom is the product of $P_{ij}(t)$s, the probabilities for individual branches on the tree. (B) A four taxa rooted phylogenetic tree. The i^{th} site observed at tip has pattern *AACT*. The internal nodes $z_k\in\{A, C, G, T\}$, $5\leq k\leq 7$ are ancestral status; and $b_j\geq 0$, $1\leq j\leq 6$ are branches lengths.

ered in this paper are time reversible. Thus, the rate matrix Q can be diagonalized and has real eigenvalues. It is straightforward that all elements $P_{xy}(t)$ in the transition probability matrix $P(t)$ have a continuous second order partial derivative with respect to t and parameters in the rate matrix Q.

As an example, we express probability p_i for a particular pattern $\{AACT\}$ observed at the tips of a 4-taxon rooted tree (Figure A1B). Let $z_k \in \{A, C, G, T\}$, $5 \le k \le 7$ be the ancestor status, let $b_j > 0, 1 \le j \le 6$ be the branch lengths. Let π be the equilibrium base frequencies, and we assume that the nucleotides at the root of the tree have reached the equilibrium frequencies π. The probability of observing pattern $AACT$ at the tips of the tree is a sum over all possible assignments of nucleotides to internal nodes, i.e.,

$$\begin{aligned} p_i &= Pr[\{AACT\}|\tau, b, Q] \\ &= \sum_{z_7}\sum_{z_6}\sum_{z_5}\{\pi_{z_7} P_{z_7 z_5}(b_5|Q)\cdot P_{z_7 z_6}(b_6|Q)\cdot P_{z_5 A}(b_1|Q) \\ &\quad \cdot P_{z_5 A}(b_2|Q)\cdot P_{z_6 C}(b_3|Q)\cdot P_{z_6 T}(b_4|Q)\}, \end{aligned} \tag{A1}$$

where $P_{xy}(b_j|Q)$ is the transition probability that nucleotide y is substituted for x over a branch length b_j. Equation (A1) has $4^3=64$ terms. In general the expression for k species will have 2^{2k-2} terms. As p_i is the sum of multiplications of equilibrium frequencies π and transition probabilities $P_{xy}(b)$, p_i has a continuous second-order partial derivative with respect to branch lengths b and parameters in the rate matrix Q and equilibrium frequency vector π.

Let $\xi = \{\xi_1, \xi_2, \ldots, \xi_{4^N}\}$ be the frequencies of 4^N patterns observed in data D. Assuming that the sites evolve independently along the lineages of a phylogenetic tree, the probability density function of data D in terms of the frequencies ξ of 4^N nucleotide patterns can be expressed as

$$f(D|\phi) = \prod_{i=1}^{4^N} (p_i)^{\xi_i}. \tag{A2}$$

It follows from (2) that the log likelihood function L is

$$L(\phi|D) = \sum_{i=1}^{4^N} \xi_i \ln p_i. \tag{A3}$$

The log likelihood function for a fixed model $m \in \Omega$ is denoted by $L_m(\phi_m|D)$, or simply by L_m. If the tree topology τ is given, p_i has continuous second-order partial derivatives with respect to parameters ϕ. It follows immediately that the likelihood function $L_m(\phi_m|D)$ has continuous second-order partial derivatives with respect to ϕ as L_m is the sum of the weighted logarithm of p_i shown in (A3).

A2 The flow chart for estimating the expected KL divergence

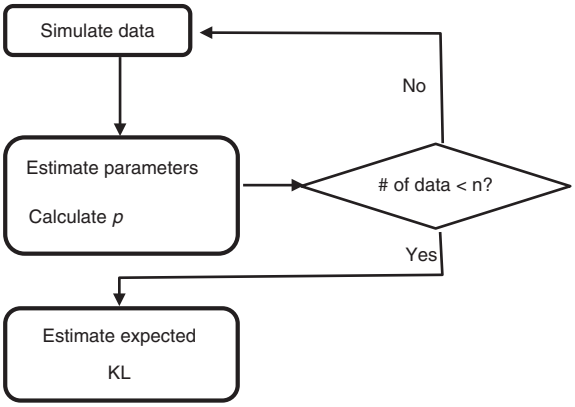

Figure A2 The flow chart for estimating the expected KL divergence.

References

Abdo, Z., V. Minin, P. Joyce and J. Sullivan (2005): "Accounting uncertainty in the tree topology has little effect on the decision theoretic approach on model selection in phylogenetic estimation." Mol. Biol. Evol., 22, 691–703.

Akaike, H. (1974): "A new look at the statistical model identification," IEEE Trans. Aut. Control, 19, 716–723.

Alfaro, M. and J. Huelsenbeck (2006): "Comparative performance of bayesian and aicbased measures of phylogenetic model uncertainty," Syst. Biol., 55, 89–96.

Anisimova, M. and O. Gascuel (2006): "Approximate likelihood-ratio test for branches: a fast, accurate and powerful alternative," Syst. Biol., 55, 539–552.

Boettiger, C., G. Coop and P. Ralph (2012): "Is your phylogeny informative? Measuring the power of comparative methods," Evolution, 66, 2240–2251.

Bos, D. and D. Posada (2005): "Using models of nucleotide evolution to build phylogenetic trees," Developmental and Comparative Immunolology, 29, 211–227.

Buckley, T. and C. Cunningham (2002): "The effects of nucleotide substituion model assumptions on estimates of nonparametric bootstrap support," Mol. Biol. Evol., 19, 394–405.

Burham, K. and D. Anderson (2004): Model selection and multimodel inference, Springer-Verlag: New York.

Cunningham, C., H. Zhu and D. Hillis (1998): "Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies," Evolution, 52, 978–987.

Darriba, D., G. Taboada, R. Doallo and D. Posada (2012): "Jmodeltest 2: more models, new heuristics and parallel computing," Nature Methods, 9, 772.

Davison, A. (2003): Statistical models, Cambridge University Press: New York.

Evans, J. and J. Sullivan (2010): "Approximation model probabilities in bic and dt approaches to model selection in phylogenetics," Mol. Biol. Evol., 28, 343–349.

Felsenstein, J. (1981): "Evolutionary trees from dna sequences: a maximum likelihood approach," J. Mol. Evol., 17, 368–376.

Frati, F. (1997): "Evolution of the mitochondrial coii gene in collembola," J. Mol. Evol., 44, 145–158.

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

Hayasaka, K., T. Gojobori and S. Horai (1988): "Molecular phylogeny and evolution of primate mitochondrial DNA," Mol. Biol. Evol., 5, 626–644.

Holder, M., P. Lewis and D. Swofford (2010): "The akaike information criterion will not choose the no common mechanism model," Syst. Biol., 59, 477–485.

Huelsenbeck, J. and K. Crandall (1997): "Phylogeny estimation and hypothesis testing using maximum likelihood," Annu. Rev. Ecol. Evol. Syst., 42, 247–264.

Huelsenbeck, J., B. Larget and M. Alfaro (2004): "Bayesian phylogenetic model selection using reversible jump markov chain monte carlo," Mol. Biol. Evol., 21, 1123–1133.

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

Ishiguro, M., Y. Sakamoto and G. Kitagawa (1997): "Bootstrapping log likelihood and eic, an extension of aic," Ann. I. Stat. Math., 49, 411–434.

Jermiin, L., V. Jayaswal, F. Ababneh and J. Robinson (2008): "Phylogenetic model evaluation," Methods Mol. Biol., 452, 31–64.

Johnson, J. and K. Omland (2004): "Model selection in ecology and evolution," Trends Ecol. Evol., 19, 101–108.

Jukes, T. and C. Cantor (1969): "Evolution of protein molecules," In: Munro, H.N. (Eds.), Mammalian protein metabolism. Academic Press: New York. 21–132.

Kelchner, S. (2009): "Phylogenetic models and model selection for noncoding Dna," Plant Syst. Evol., 282, 109–126.

Kelchner, S. and M. Thomas (2007): "Model use in phylogenetics: nine key questions," Trends Ecol. Evol., 282, 109–126.

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

Luo, A., H. Qiao, Y. Zhang, W. Shi, Y. Ho, W. Xu, A. Zhang and C. Zhu (2010): "Performance of crtiteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets," BMC Evol. Biol., 10, 242.

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

Pol, D. (2004): "Empirical problems of the hierarchical likelihood ratio test for model selection," Syst. Biol., 53, 949–962.

Posada, D. (2008): "Jmodeltest: phylogenetic model averaging," Mol. Biol. Evol., 25, 1253–1256.

Posada, D. and T. Buckley (2004): "Model selection and model averaging in phylogenetics: advantage of akaike information criterion and baysian approaches over likelihood ratio tests," Syst. Biol., 53, 793–808.

Posada, D. and K. Crandall (1998): "Modeltest: testing the model of DNA substitution," Bioinformatics, 14, 817–818.

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

Rambaut, A. and N. Grassly (1997): "Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic tree," Comput. Appl. Biosci., 13, 235–238.

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

Schwarz, G. (1978): "Estimating the dimension of a model," Ann. Stat., 6, 461–464.

Self, S. and K.-Y. Liang (1987): "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions," J. Am. Stat. Assoc., 82, 605–610.

Shapiro, B., A. Rambaut and A. Drummond (2006): "Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences," Mol. Biol. Evol., 23, 7–9.

Sullivan, J. and P. Joyce (2005): "Model selection in phylogenetics," Annu. Rev. Ecol. Evol. Syst., 36, 445–466.

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

Tavaré, S. (1986): "Some probabilistic and statistical problems in the analysis of dna sequences," Lect. Math. Life Sci. (American Mathematical Society), 17, 57–86.

Wu, C., M. Suchard and A. Drummond (2013): "Bayesian selection of nucleotide substitution models and their site assignments," Mol. Biol. Evol., 30, 669–688.

Yang, Z. (1994): "Maximum likelihood phylogenetic estimation from dna sequeces with variable rates over sites: approximate methods," J. Mol. Evol., 39, 306–314.

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