Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees

Ziheng Yang^{a,b,c,1} and Tianqi Zhu^c

^aDepartment of Genetics, University College London, London WC1E 6BT, United Kingdom; ^bRadcliffe Institute for Advanced Studies, Harvard University, Cambridge, MA 02138; and ^cKey Laboratory of Random Complex Structures and Data Science (RCSDS), National Center for Mathematics and Interdisciplinary Sciences (NCMIS), Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved January 2, 2018 (received for review July 17, 2017)

The Bayesian method is noted to produce spuriously high posterior probabilities for phylogenetic trees in analysis of large datasets, but the precise reasons for this overconfidence are unknown. In general, the performance of Bayesian selection of misspecified models is poorly understood, even though this is of great scientific interest since models are never true in real data analysis. Here we characterize the asymptotic behavior of Bayesian model selection and show that when the competing models are equally wrong, Bayesian model selection exhibits surprising and polarized behaviors in large datasets, supporting one model with full force while rejecting the others. If one model is slightly less wrong than the other, the less wrong model will eventually win when the amount of data increases, but the method may become overconfident before it becomes reliable. We suggest that this extreme behavior may be a major factor for the spuriously high posterior probabilities for evolutionary trees. The philosophical implications of our results to the application of Bayesian model selection to evaluate opposing scientific hypotheses are yet to be explored, as are the behaviors of non-Bayesian methods in similar situations.

Bayesian inference | fair-coin paradox | model selection | posterior probability | star-tree paradox

The Bayesian method was introduced into molecular phylogenetics in the 1990s (1–3) and has since become one of the most popular methods for statistical analysis in the field, in particular, for estimation of species phylogenies (4–7). It has been noted that the method often produces very high posterior probabilities for trees or clades (nodes in the tree). In the first-ever Bayesian phylogenetic calculation, a biologically reasonable tree for five species of great apes was produced from a dataset of 11 mitochondrial tRNA genes (739 sites), but the posterior probability for that tree, at 0.9999, was uncomfortably high (1). In the past two decades, the Bayesian method has been used to analyze thousands of datasets, with the computation made possible through Markov chain Monte Carlo (MCMC) (4, 5). It has become a common practice to report posterior clade probabilities only if they are <100% (because most estimates are 100%). In some cases the high posterior probabilities are decidedly spurious. For example, conflicting trees may be inferred from the same data under different evolutionary models. Different trees may be inferred depending on the species sampled in the dataset (8) or on whether protein sequences or the encoding DNA sequences are analyzed (9). In such cases, the different trees cannot all be correct, even if the true tree is unknown. The concern is not so much that the inferred species relationships may be wrong but that they are supported by extremely high posterior probabilities.

In the star-tree paradox, large datasets were simulated using the star tree and then analyzed to calculate the posterior probabilities for the three binary trees (Fig. 1). Most biologists would want the posterior probabilities for the binary trees to converge

to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ when the amount of data increases (10–12). Instead they fluctuate among datasets according to a statistical distribution, sometimes producing strong support for a binary tree even though the data do not contain any information either for or against any binary tree (13–15).

Bayesian model selection is known to be consistent (16). When the data size $n \rightarrow \infty$, the true model “dominates,” with its posterior probability approaching 1. If several models are equally right, the model with fewer parameters dominates. However, this theory applies only if the true model is included in the comparison. Given that a model is a simplified representation of the physical world, the more common situation in real data analysis should be the comparison of models that are all wrong. Not many theoretical results appear to exist concerning Bayesian comparison of misspecified models (17).

Here we study the asymptotic behavior of Bayesian model selection in a general setting where multiple misspecified models are compared. We are interested in how the posterior probabilities for models behave when the data size increases. Do the dynamics depend on whether there are any free parameters in the models? If one model is less wrong than another (in a certain sense appropriately defined), will the less wrong model always win? We present the proofs and mathematical analyses in *General Theory for Equally Wrong Models with No Free Parameters* ($d = 0$) and *General Theory for Equally Right or Equally Wrong Models with Free Parameters* ($d > 0$). In the main text, we summarize our results and illustrate them using three canonical simple problems. Our analysis suggests that the problem exposed by the

Significance

The Bayesian method is widely used to estimate species phylogenies using molecular sequence data. While it has long been noted to produce spuriously high posterior probabilities for trees or clades, the precise reasons for this overconfidence are unknown. Here we characterize the behavior of Bayesian model selection when the compared models are misspecified and demonstrate that when the models are nearly equally wrong, the method exhibits unpleasant polarized behaviors, supporting one model with high confidence while rejecting others. This provides an explanation for the empirical observation of spuriously high posterior probabilities in molecular phylogenetics.

Author contributions: Z.Y. designed research; Z.Y. and T.Z. performed research; Z.Y. and T.Z. analyzed data; and Z.Y. and T.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

¹To whom correspondence should be addressed. Email: z.yang@ucl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1712673115/-DCSupplemental.

whether they are both right or both wrong or by whether the compared models have unknown parameters. For example, cases B_1 (two right models) and B_3 (two equally wrong models) in Fig. 2 show the same volatile behavior, while cases C_1 (no free parameters) and C_2 (with free parameters) show the same polarized behavior.

Problem 1. Fair-Coin Paradox (Equally Wrong Models with No Free Parameter). Consider a coin-tossing experiment in which the coin is fair with the probability of heads $p = \frac{1}{2}$. We use the data of x heads in n tosses to compare two models: $H_1: p = 0.4$ (tail bias) and $H_2: p = 0.6$ (head bias). The two models are equally wrong. We assign a uniform prior for the two models ($\frac{1}{2}$ each) and calculate the posterior model probability $P_1 = \mathbb{P}(H_1|x)$. This is a type-3 problem (Fig. 2, C_1).

As the models involve no free parameters, the likelihood (L) and marginal likelihood (M) are the same, given by the binomial probability for data x . The posterior odds are the likelihood-ratio

$$\frac{P_1}{1 - P_1} = \frac{M_1}{M_2} = \frac{0.4^x \cdot 0.6^{n-x}}{0.6^x \cdot 0.4^{n-x}} = \left(\frac{0.4}{0.6}\right)^{2x-n}. \quad [2]$$

When n is large, P_1 tends to be extreme (close to 0 or 1). Indeed, $\alpha < P_1 < 1 - \alpha$ if and only if $|2x - n| < B = \frac{\log\{\alpha/(1-\alpha)\}}{\log\{0.4/0.6\}}$. If n is large, $2x - n$ is $\sim \mathcal{N}(0, n)$, so that

$$\mathbb{P}\{|2x - n| < B\} \approx 1 - 2\Phi\left(-\frac{B}{\sqrt{n}}\right) \approx \frac{2B}{\sqrt{2\pi n}}, \quad [3]$$

where Φ is the cumulative distribution function (CDF) for $\mathcal{N}(0, 1)$. If $\alpha = 1\%$, we have $B = 11.33296$, so that only 11 data outcomes will give P_1 in the range $(0.01, 0.99)$, with $x - \frac{n}{2}$ being $-5, -4, \dots, 5$. For $n = 10^3, 10^4, 10^5, 10^6$, we have $\mathbb{P}\{0.01 < P_1 < 0.99\} = 0.280, 0.090, 0.0286$, and 0.0090 using the normal approximation of Eq. 3 or $0.272, 0.0876, 0.0277$, and 0.0088 exactly by the binomial distribution. Thus, in large datasets, moderate posterior probabilities will be rare, and either H_1 or H_2 will be favored with posterior > 0.99 . When $n \rightarrow \infty$, P_1 has a degenerate two-point distribution, taking the values 0 and 1, each half of the times. This is the type-3 polarized behavior. Note that there is no information either for or against either model in the data. Fig. 3 A, *i* shows the distribution of P_1 for $n = 10^3$.

Fig. 3 A, *ii* shows the comparison of $H_1: p = 0.42$ against $H_2: p = 0.6$ when the truth is $p = 0.5$. Here H_1 is less wrong and will eventually dominate. However, in large and finite datasets, the more wrong model H_2 can often receive high support. For example, for $n = 10^3$, nonextreme posterior probabilities in the range $0.01 < P_1 < 0.99$ occur for only 13 data outcomes, with x being 504–516, and in 14.8% of datasets, x is greater than those values so that $P_2 > 0.99$. Indeed over the whole range $36 \leq n \leq 11,611$, the more wrong model H_2 is strongly favored too often, with $\mathbb{P}(P_2 > 0.99) > 0.01$. The method becomes overconfident before it becomes reliable. It may be noted that such strong support for the more wrong model occurs only when the two models are opposing each other. It does not occur if both models are wrong in the same direction: In the comparison of $H_1: p = 0.4$ and $H_2: p = 0.42$ when the truth is $p = 0.5$, the less wrong model H_2 dominates in the posterior.

Problem 2. Fair-Balance Paradox (Equally Right Models or Equally Wrong and Indistinct Models). The true model is $\mathcal{N}(0, 1)$, and we compare two models $H_1: \mathcal{N}(\mu, 1/\tau)$, $\mu < 0$ and $H_2: \mathcal{N}(\mu, 1/\tau)$, $\mu > 0$, with τ given. The data may represent measurement errors observed on a fair balance while the models claim that the balance has an unknown negative or positive bias. The best-fitting parameter value (the MLE when the data size $n \rightarrow \infty$) is $\mu^* = 0$

Fig. 3. The distribution of posterior model probability $P_1 = \mathbb{P}\{H_1|x\}$ in three inference problems. (A) Problem 1 (fair-coin paradox) is for a coin-tossing experiment, where the true model is $p = 0.5$ (a fair coin), and the compared models are (A, *i*) $H_1: p = 0.4$ and $H_2: p = 0.6$ so that the two models are equally wrong and (A, *ii*) $H_1: p = 0.42$ and $H_2: p = 0.6$ so that H_1 is less wrong than H_2 . The data size (the number of coin tosses) is 10^3 . (B) Problem 2 (fair-balance paradox) is for a normal-distribution example in which the true model is $\mathcal{N}(0, 1)$, and the two compared models are $H_1: \mathcal{N}(\mu, 1/\tau)$, $\mu < 0$ and $H_2: \mathcal{N}(\mu, 1/\tau)$, $\mu > 0$, with variance $1/\tau$ given. The two models are equally right when $\tau = 1$ and equally wrong but indistinct when $\tau = 1/9$ or 9 . The data size is $n = 10^3$. The plots for $n = 100$ or ∞ are nearly the same. (C) Problem 3 (fair-balance paradox) is for a normal-distribution example in which the true model is $\mathcal{N}(0, 1)$, and the two compared models are $H_1: \mathcal{N}(\mu, 1/\tau_1)$ and $H_2: \mathcal{N}(\mu, 1/\tau_2)$, with (C, *i*) $\tau_1 = 0.25$ and $\tau_2 = 2.58666$, so that the two models are equally wrong, and (C, *ii*) $\tau_1 = 0.3$ and $\tau_2 = 2.58666$, so that H_1 is less wrong than H_2 . The prior is $\mu \sim \mathcal{N}(0, 1/\xi)$ under each model, with $\xi = 1$. The data size is $n = 100$. All densities are estimated by simulating 10^5 samples for P_1 .

in each model, when the two models become identical (indistinct). Thus, the two models are equally right if $\tau = 1$ (Fig. 2, B_2), and are equally wrong if $\tau = 1/9$ or 9 (Fig. 2, B_4).

We assign a uniform prior on the two models ($\frac{1}{2}$ each), and $\mu \sim \mathcal{N}(0, 1/\xi)$ with ξ fixed, truncated to the appropriate range under each model. The data (x), an i.i.d. sample from $\mathcal{N}(0, 1)$, can be summarized as the sample mean $\bar{x}$. It can be shown that the posterior model probability $P_1 = \mathbb{P}\{H_1|x\}$ varies among datasets according to the density

$$f(P_1) = \frac{\sqrt{\tau + \xi/n}}{\tau} \cdot \exp\left\{\frac{[\Phi^{-1}(P_1)]^2}{2} \left[1 - \frac{1}{\tau} - \frac{\xi}{n\tau^2}\right]\right\}, \quad [4]$$

where Φ^{-1} is the inverse CDF for $\mathcal{N}(0, 1)$ (Analysis of Problem 2 (Two Equally Right Models or Equally Wrong but Indistinct Models)).

Fig. 3B shows the density of P_1 for different values of precision (τ), with $n = 10^3$. If $\tau = 1$, the two models are equally right, and $f(P_1) \rightarrow 1$ when $n \rightarrow \infty$ so that P_1 behaves like a $\mathcal{U}(0, 1)$ random number (11, 12). If $\tau < 1$, the assumed variance ($1/\tau$) is larger than the true variance, so that the distribution has a mode at $\frac{1}{2}$. If $\tau > 1$, the assumed variance is too small, and P_1 has a U-shaped distribution. If one overstates the precision of the experiment, one tends to overinterpret the data and generate extreme posterior model probabilities. In all three cases ($\tau < 1, = 1, > 1$), P_1 has a nondegenerate distribution.

Problem 3. Fair-Balance Paradox (Equally Wrong and Distinct Models). The true model is $\mathcal{N}(0, 1)$, and the two compared models are $H_1: \mathcal{N}(\mu, 1/\tau_1)$ and $H_2: \mathcal{N}(\mu, 1/\tau_2)$, with $\tau_1 < 1 < \tau_2$ given, while μ is a free parameter in each model. The best-fitting parameter value is $\mu^* = 0$ in each model, irrespective of the value of τ assumed. Both models are wrong because of the misspecified variance: H_1 is overdispersed while H_2 is underdispersed. They are equally wrong, in the sense that $D_1 = D_2$ in Eq. 1, if

$$\log \frac{\tau_1}{\tau_2} = \tau_1 - \tau_2 \quad [5]$$

method, but those studies examined case A only, so the real situation is worse than previously realized.

A practically important scenario is where all binary trees are wrong because of violation of the evolutionary model but the true tree is less wrong than the other trees. We present such a case in Table S2, in which the data are simulated under JC+ Γ (with $\alpha=1$) using a binary tree with a short internal branch ($t_0=0.002$) and then analyzed under JC. When the amount of data approaches infinity, the true tree will eventually win, but there exists a twilight zone in which high posterior probabilities for wrong trees occur too frequently; according to Table S2, this zone is wider than $10^3 < n < 10^5$. For example, at sequence length $n=10^4$ and at the 1% nominal level, the error rate of rejecting the true tree is 25.0% and the error rate of accepting a wrong tree is 16.6% (Table S2).

Discussion

High Posterior Probabilities for Phylogenetic Trees. This work has been motivated by the phylogeny problem and in particular by the empirical observation of spuriously high posterior probabilities for phylogenetic trees (9–14). We note that certain biological processes such as deep coalescence (24, 25), gene duplication followed by gene loss (26), and horizontal gene transfer (24, 26) may cause different genes or genomic regions to have different histories. However, as discussed in the Introduction, posterior probabilities for many trees or clades observed in real data analyses are decidedly spurious even if the true tree is unknown.

One explanation for the spuriously high posterior probabilities for phylogenetic trees is the failure of current evolutionary models to accommodate interdependence among sites in the sequence, leading to an exaggeration of the amount of information in the data. Interacting sites may carry much less information than independent sites. This explanation predicts the problem to be more serious in coding genes than in noncoding regions of the genome as noncoding sites may be evolving largely independently due to lack of functional constraints. However, empirical evidence points to the opposite, with noncoding regions having higher substitution rates and higher information content (if they are not saturated with substitutions), generating more extreme posteriors for trees.

Our results suggest that the problem may lie deeper and may be a consequence of the polarized nature of Bayesian model selection when all models under comparison are misspecified. As the assumptions about the process of sequence evolution are unrealistic, the likelihood model is wrong whatever the tree, although the true tree may be expected to be less wrong than the other trees. As the different trees constitute opposing models that are nearly equally wrong, the inference problem is one of type 3 (Fig. 2, C_4). Bayesian tree estimation may then be expected to produce extreme posterior probabilities in large datasets.

Bayesian Selection of Opposing Misspecified Models. We have provided a characterization of model selection problems according to the asymptotic behavior of the Bayesian method as the data size $n \rightarrow \infty$ [Fig. 2 and *General Theory for Equally Wrong Models with No Free Parameters* ($d=0$) and *General Theory for Equally Right or Equally Wrong Models with Free Parameters* ($d>0$)]. While all of the problems considered here involve comparison of two equally right or equally wrong models, three different asymptotic behaviors are identified, which we label as type 1, type 2, and type 3. The type-1 behavior is for the posterior model probability P_1 to converge to a sensible point value, such as $\frac{1}{2}$. We consider this to be a good balanced behavior, following phylogeneticists (10–12). The rationale is that one would like a sure answer given an infinite amount of data and the only reasonable sure answer

should be $\frac{1}{2}$ for each model, since the data contain no information for or against either model. This behavior occurs only when the two models are identical or overlapping, a situation that does not appear relevant to scientific inference. With type-2 behavior, P_1 fluctuates among datasets (each of infinite size) like a random number, so that strong support may be attached to a particular model in some datasets. Biologists were surprised at this erratic behavior (10–12), which we label as volatile. This occurs when the models are equally right or equally wrong but indistinct. In theory, type-2 behavior may not pose a serious problem, because the parameter posteriors under the models, if examined carefully, should make it clear that the competing models essentially gave the same interpretation of the data and should lead to the same scientific conclusion. In data simulated in ref. 12 or in Fig. 4A and A', the estimates of t_0 should be very close to 0, and all binary trees are similar to the same star tree. Nevertheless this escaped our attention at the time.

With type-3 behavior, P_1 is ~ 0 in half of the datasets and ~ 1 in the other half. We describe this behavior as polarized. This occurs when the two models are equally wrong and distinct. Type-3 problems may be the most relevant to practical data analysis given that all models are simplified representations of reality and are thus wrong. A variation to type-3 problems is when one model is only slightly less wrong than another (Fig. 3A, ii and C, ii and Table S2). While the less wrong model eventually wins in the limit of infinite data, Bayesian model selection is overconfident in large but finite datasets, supporting the more wrong model with high posterior too often.

Note that the question of how the posterior model probability should behave when large datasets are used to compare two equally wrong models is somewhat philosophical and may not have a simple answer. One position is to accept whatever behavior the Bayesian method exhibits. This may be legitimate given that Bayesian theory is the correct probability framework for summarizing evidence in the prior and likelihood. The polarized behavior in type-3 problems may then be seen as a consequence of “user error” (for not including the true model in the comparison), exacerbated by the large data size. In this regard we note that the posterior predictive distribution (27, 28) can be used to assess the general adequacy of any model or the compatibility between the prior and the likelihood, and indeed this has been widely used to assess the goodness of fit of models in phylogenetics (29, 30). Nevertheless, a number of sophisticated and parameter-rich models have been developed for Bayesian phylogenetic analysis, due to three decades of active research (31), and furthermore extreme sensitivity to the assumed model is not a desirable property of an inference method. Seven decades ago, Egon S. Pearson (ref. 32, p.142) wrote that “Hitherto the user has been accustomed to accept the function of probability theory laid down by the mathematicians; but it would be good if he could take a larger share in formulating himself what are the practical requirements that the theory should satisfy in application.” This stipulation may be relevant even today.

Two heuristic approaches have been suggested to remedy the high posterior model probabilities in the context of phylogenies. The first one is to assign nonzero probabilities to multifurcating trees (such as the star tree of Fig. 1) in the prior (11). This is equivalent to assigning some prior probability to the model $p=0.5$ in the fair-coin example of problem 1. While this resolves the star-tree paradox, it suffers from the conceptual difficulty that the multifurcating trees may not be plausible biologically. The second approach is to let the internal branch lengths in the binary trees become increasingly smaller in the prior when the data size increases (12, 14). This is non-Bayesian in that the prior depends on the size of the data. With both approaches, the posterior is extremely sensitive to the prior (9).

Non-Bayesian Methods. The phylogeny problem was described by Jerzy Neyman (ref. 33, p. 1) as “a source of novel statistical problems.” In the frequentist framework, the test of phylogeny, or test of nonnested models in general, offers challenging inference problems. Note that in many model selection problems, the model itself is not the focus of interest. For example, when an experiment is conducted to evaluate the effect of a new fertilizer, the sensitivity of the inference to the assumed normal distribution with homogeneous variance may be of concern, but the focus is not on the normal distribution itself. In phylogenetics, the phylogeny (which is a model) is of primary interest, far more important than the branch lengths (which are parameters in the model). The test of phylogeny is thus more akin to significance/hypothesis testing than to model selection. Model-selection criteria such as Akaike information criteria (34) or Bayesian information criteria (35) simply rank the trees by their likelihood (maximized over branch lengths) and will not be useful for attaching a measure of significance or confidence in the estimated tree. The phylogeny problem (or the problem of comparing nonnested models in general) falls outside the Fisher–Neyman–Pearson framework of hypothesis testing, which involves two nested models, one of which is true (36, 37).

In principle Cox’s likelihood-ratio test (38), which conducts multiple tests with each model used as the null, can be used to compare nonnested models. For type-3 problems (Fig. 2, C_1 – C_4), this test should lead to rejection of all models. Cox’s test has not been used widely in phylogenetics, apparently because of the existence of a great many possible trees and the heavy computation needed to generate the null distribution by simulation.

The most commonly used method for attaching a measure of confidence in the maximum-likelihood tree is the bootstrap

(39), which samples sites (alignment columns) to generate bootstrap pseudodatasets and calculates the bootstrap support value for a clade (a node on the species tree) as the proportion of the pseudodatasets in which that node is found in the inferred ML tree. This application of bootstrap for model comparison appears to have important differences from the conventional bootstrap for calculating the standard errors and confidence intervals for a parameter estimate (40); a straightforward interpretation of the bootstrap support values for trees remains elusive (31, 41–43). At any rate, the asymptotic behavior of bootstrap support values under the different scenarios of Fig. 2 merits further research. For the fair-coin example of problem 1 (Fig. 2, C_1), the bootstrap support converges to $U(0, 1)$, different from the posterior probability, although other cases are yet to be explored.

Materials and Methods

Star-Tree Simulations. For Fig. 4 A, A', B, and B', the true tree is T_0 of Fig. 1A. The data of counts of five site patterns (xxx, xxy, yxx, yxy, and xyz) were simulated by multinomial sampling (21) and analyzed using a C program, which calculates the 2D integrals in the marginal likelihood by Gaussian–Legendre quadrature with 128 points (14). For Fig. 4 C and C', the true tree is T_0 of Fig. 1B. Sequence alignments were simulated using EVOLVER and analyzed using MrBayes (4).

ACKNOWLEDGMENTS. We thank Philip Dawid and Wally Gilks for stimulating discussions and Jeff Thorne and an anonymous reviewer for constructive comments. Z.Y. was supported by a Biotechnological and Biological Sciences Research Council grant (BB/P006493/1) and in part by the Radcliffe Institute for Advanced Study at Harvard University. T.Z. was supported by Natural Science Foundation of China grants (31671370, 31301093, 11201224, and 11301294) and a grant from the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2015080).

- Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J Mol Evol* 43:304–311.
- Mau B, Newton M (1997) Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J Comput Graph Stat* 6:122–131.
- Li S, Pearl D, Doss H (2000) Phylogenetic tree reconstruction using Markov chain Monte Carlo. *J Am Stat Assoc* 95:493–508.
- Ronquist F, et al. (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61:539–542.
- Bouckaert R, et al. (2014) Beast 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10:e1003537.
- Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
- Chen MH, Kuo L, Lewis P (2014) *Bayesian Phylogenetics: Methods, Algorithms, and Applications* (Chapman and Hall/CRC, London).
- Bourlat SJ, et al. (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* 444:85–88.
- Yang Z (2008) Empirical evaluation of a prior for Bayesian phylogenetic inference. *Philos Trans R Soc Lond B* 363:4031–4039.
- Suzuki Y, Glazko G, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci USA* 99:16138–16143.
- Lewis P, Holder M, Holsinger K (2005) Polytomies and Bayesian phylogenetic inference. *Syst Biol* 54:241–253.
- Yang Z, Rannala B (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst Biol* 54:455–470.
- Steel MA, Matsen F (2007) The Bayesian “star paradox” persists for long finite sequences. *Mol Biol Evol* 24:1075–1079.
- Yang Z (2007) Fair-balance paradox, star-tree paradox and Bayesian phylogenetics. *Mol Biol Evol* 24:1639–1655.
- Susko E (2008) On the distributions of bootstrap support and posterior distributions for a star tree. *Syst Biol* 57:602–612.
- Dawid A (2011) Posterior model probabilities. *Philosophy of Statistics*, eds Bandyopadhyay PS, Forster M (Elsevier, New York), pp 607–630.
- Berk R (1966) Limiting behavior of posterior distributions when the model is incorrect. *Ann Math Stat* 37:51–58.
- White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50:1–25.
- Yang Z, Goldman N, Friday AE (1995) Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Syst Biol* 44:384–399.
- Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17:368–376.
- Yang Z (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst Biol* 43:329–342.
- Jukes T, Cantor C (1969) Evolution of protein molecules. *Mammalian Protein Metabolism*, ed Munro H (Academic, New York), pp 21–123.
- Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* 10:1396–1401.
- Maddison W (1997) Gene trees in species trees. *Syst Biol* 46:523–536.
- Xu B, Yang Z (2016) Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204:1353–1368.
- Nichols R (2001) Gene trees and species trees are not the same. *Trends Ecol Evol* 16:358–364.
- Roberts H (1965) Probabilistic prediction. *J Am Stat Assoc* 60:50–62.
- Box G (1980) Sampling and Bayes’ inference in scientific modelling and robustness. *J R Stat Soc A* 143:383–430.
- Sullivan J, Joyce P (2005) Model selection in phylogenetics. *Annu Rev Ecol Syst* 36:445–466.
- Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol* 23:1762–1775.
- Yang Z (2014) *Molecular Evolution: A Statistical Approach* (Oxford Univ Press, Oxford).
- Pearson E (1947) The choice of statistical tests illustrated on the interpretation of data classed in the 2x2 table. *Biometrika* 34:139–167.
- Neyman J (1971) Molecular studies of evolution: A source of novel statistical problems. *Statistical Decision Theory and Related Topics*, eds Gupta SS, Yackel J (Academic, New York), pp 1–27.
- Akaike H (1973) Information theory and an extension of the likelihood principle. *Proceedings of the Second International Symposium of Information Theory*, eds Petrov BN, Csaki F (Akademiai Kiado, Budapest), pp 267–281.
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
- Lehmann E (1997) *Testing Statistical Hypothesis* (Springer, New York), 2nd Ed.
- Goldman N, Anderson J, Rodrigo A (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol* 49:652–670.
- Cox D (1961) Tests of separate families of hypotheses. *Proc 4th Berkeley Symp Math Stat Prob* 1:105–123.
- Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
- Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap* (Chapman and Hall, London).
- Felsenstein J, Kishino H (1993) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst Biol* 42:193–200.
- Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci USA* 93:7085–7090, and correction (1996) 93:13429–13434.
- Susko E (2009) Bootstrap support is not first-order correct. *Syst Biol* 58:211–223.