A. W. F. Edwards and the Origin of Bayesian Phylogenetics

ZIHENG YANG

In the early 1960s, Anthony Edwards and Luca Cavalli-Sforza made an effort to apply R. A. Fisher's maximum-likelihood method to estimate genealogical trees of human populations using gene-frequency data. They used the Yule branching process to describe the probabilities of the trees and branching times and the Brownian-motion process to model the drift of gene frequencies (after a suitable transformation) over time along the branches. They experienced considerable difficulties, including "singularities" in the likelihood surface, mainly because a distinction between parameters and random variables was not clearly made. In the process, they invented the distance (additive-tree) and parsimony (minimum-evolution) methods, both of which they viewed as heuristic approximations to maximum likelihood. The statistical nature of the inference problem was not clarified until Edwards's paper, [46], published in 1970, which pointed out that the trees should be estimated from their conditional distribution given the genetic data, rather than from the "likelihood function." In modern terminology, this is the Bayesian approach to phylogeny estimation: the Yule process specifies a prior on trees, while the conditional distribution of the trees given the data is the posterior. This article discusses the connections of the remarkable paper, [46], to modern Bayesian phylogenetics, and briefly comments on some modelling decisions Edwards made then that still concern us today in modern Bayesian phylogenetics. The reader I have in mind is familiar with modern phylogenetic methods but may not have read [46], which is published in a statistics journal.

The Model and the Statistical Problem of Phylogeny Estimation

The data considered by Edwards and Cavalli-Sforza in 1964 ([27]) and 1966 ([35]) consist of gene frequencies of common blood groups from different human populations. Edwards treated different human populations while

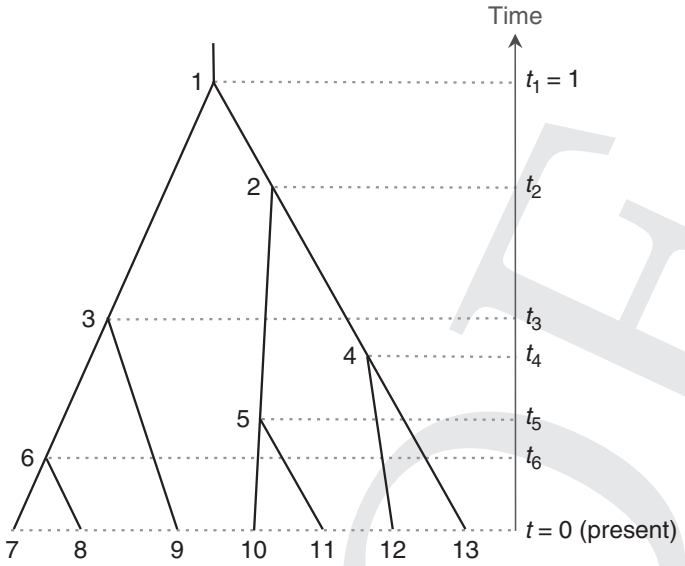

Figure 1. A phylogeny for seven ($n = 7$) species or populations used to illustrate the inference problem considered by Edwards in [46]. The tips (final particles) are numbered $n,\ n + 1, \ldots, 2n - 1$. The interior nodes are numbered $1, 2, \ldots, n - 1$. They represent the branching events and are ordered by time: $t_1 > t_2 > \ldots > t_{n-1}$. The time machine runs backwards, so that the present time is $t = 0$ while the age of the root (the origin at the first split) is fixed at $t_1 = 1$. The data are observed measurements in p characters from the n modern species: $\xi = \{\xi_{ik}\}$, where ξ_{ik} ($i = n, \ldots, 2\ n - 1$; $k = 1, \ldots, p$) is the measurement from species i in character k. Edwards considered the tree form (labelled history F), the times of non-root internal nodes, $t = \{t_1 = 1, t_2, \ldots, t_{n-1}\}$, as well as the ancestral character states $x = \{x_{ik}\}$, $i = 1,.., n - 1$; $k = 1, \ldots, p$, as quantities of interest.

I focus on different species here. The data-generating model consists of two components. A Yule branching process, with a constant per-lineage rate λ of splitting, is used to describe the probability distribution, $f(F, \boldsymbol{t} \mid \lambda, n)$, of the phylogeny ($F$) and the branching times: $\boldsymbol{t} = \{t_1 = 1, t_2, \ldots, t_{n-1}\}$ (Figure 1). The Yule process assigns uniform probabilities to the labelled histories. The term labelled history, due to Edwards ([46]), refers to a rooted tree topology, with internal nodes ordered by time. For example, the rooted tree topology ((*a, b*), (*c, d*)) corresponds to two distinct labelled histories depending on whether the age of the *a–b* ancestor is older or younger than the age of the *c–d* ancestor. Note that other models of cladogenesis, such as the coalescent process (Kingman 1982) and the constant-rate birth–death process (Kendall 1948), all generate labelled histories with equal probabilities. In using the Yule process to describe the process of species formation, Edwards fixed the first branching event (the root of the tree) at time $t_1 = 1$, and conditioned on the number of species at the present time to be n. He derived the joint density for the tree form (F) and branching times ($\boldsymbol{t}$) as

$$f(F,t\,|\,\lambda)=\frac{2^{n-1}\lambda^{n-1}\exp\left\{-\lambda\sum_{i=2}^{n-1}t_i\right\}}{(n-1)n!(1-\mathrm{e}^{-\lambda})^{n-2}} \tag{1}$$

The second component of the model is the Brownian-motion process, used to describe the evolution of the continuous characters over time. Note that such Brownian models are now widely used in phylogenetic and phylogeographic analyses of morphological measurements from different species (Lartillotand Poujol 2011, Solis-Lemus *et al.* 2015). The Brownian motion or random walk in one dimension gives the location of the particle time t later, given that it is at location x_0 at time 0, as a Gaussian variable, $x_t \sim N(x_0, t\sigma^2)$, where the parameter σ^2 controls how fast the particle drifts and represents the evolutionary rate for the continuous character. Gene frequencies at multiple blood-group loci are treated as a p-dimensional Brownian motion. The different dimensions (different variables) are moving independently, with the same σ^2. However, the measurements for the same character observed in the modern species (which are the tips of the tree) are correlated because they may have shared some branches. For example, $\mathrm{cov}(x_{7k}, x_{9k}) = (t_1 - t_3)\sigma^2$, where $(t_1 - t_3)$ is the time shared by the two paths from the root to the tips 7 and 9 (Figure 1). Thus, given σ^2, the tree form (labelled history F) and branching times ($\boldsymbol{t}$), and the state at the root (x_{1k}), each character k observed in a modern species is normally distributed with mean x_{1k} and variance $t_1\sigma^2$. The data or the measurements of the k characters among all modern species have the probability density

$$f(\xi\,|\,F,\boldsymbol{t},\sigma^2\boldsymbol{x}_1)=\frac{\mathbf{1}}{(2\Pi)^{np/2}|\boldsymbol{T}|^{p/2}}\exp\left\{\sum_{k=1}^{p}-\frac{1}{2}(\xi_k-x_{1k}1)'\boldsymbol{T}^{-1}(\xi_k-x_{1k}1)\right\}, \tag{2}$$

where $\mathbf{1}$ is a $p \times 1$ vector with all elements to be 1, and $\boldsymbol{T}$ is the variance–covariance matrix given by the tree. This is an n-variate normal density, and is nowadays known as the phylogenetic likelihood. This is equation (8) of Edwards (p. 160).

The true parameters in the model are the Yule branching rate λ, the Brownian parameter σ^2, and the initial state $\boldsymbol{x}_1$. Given those, it should be possible to simulate the process. The other unknowns, including the tree F, the branching times $\boldsymbol{t}$, and the character states at the interior nodes x_{ik}, $2 \le i \le n-1$, $1 \le k \le p$, are random variables. If one simulates the process using the true parameters, those random variables will have different realized values among simulated replicates. In [46], Edwards pointed out that the true parameters should be estimated by maximum likelihood, with the likelihood calculated by integrating over the random variables, that is, by summing over the trees (F) and integrating over the branching times ($\boldsymbol{t}$) as well as the ancestral states. The likelihood function is thus

$$L(\lambda, \sigma^2, \boldsymbol{x}_1) = f(\xi \mid \lambda, \sigma^2, \boldsymbol{x}_1) = \sum\nolimits_F \int_{\mathbf{t}} f(F, \boldsymbol{t} \mid \lambda) f(\xi \mid F, \boldsymbol{t}, \sigma^2, \boldsymbol{x}_1) \mathrm{d}\boldsymbol{t} \quad (3)$$

Here, the notation is heuristic, and the sum is over all possible tree forms and the integral is over the $(n-2)$ branching times within each tree. The computation was deemed impossible and the model was not analyzed. Nevertheless, Edwards pointed out that random variables, F and $\boldsymbol{t}$, should be estimated from their conditional distribution given the data, with the true parameters replaced by their maximum-likelihood estimations.

The Singularity on the "Likelihood" Surface

Early attempts by Edwards and Cavalli-Sforza ([27] and [37]) treated the branching times ($\boldsymbol{t}$) and ancestral states ($\boldsymbol{x}$) as parameters. The "likelihood" function was defined as the product of the multivariate normal densities for character changes along the branches. For any branch $i \to j$, with branch length $t_i - t_j$, the density is

$$[2\pi(t_j - t_i)\sigma^2]^{-\frac{p}{2}} \times \exp\left\{-\frac{1}{2(t_j - t_i)\sigma^2} \sum\nolimits_k (x_{jk} - x_{ik})^2\right\}, \quad (4)$$

given by the Brownian-motion model. As pointed out by Edwards in [46], this "likelihood" increases without bound if $x_{ik} = x_{jk}$ for all k and if $t_i \to t_j$: there are lines of singularity in the likelihood surface when there is no change along a branch for any of the p characters and when the branch disappears. Here, the mistake was to treat the ancestral states and branching times as parameters, when they are in fact random variables with statistical distributions under the model.

Some Remarks

While the Yule process of pure birth is implemented in some Bayesian programs such as BEAST (Drummond and Rambaut 2007), it is more common to use the birth–death process, which is more general by allowing species extinctions and species sampling (Rannala and Yang 1996, Stadler 2010). Often, the distribution of times is obtained by conditioning on the number of lineages at the present time and on the age of the root, as in [46], but variations exist. For example, Thompson (1975) suggests fixing $\sigma^2 = 1$, instead of fixing $t_1 = 1$, and she treats n as data (as n may be informative about the birth rate λ), rather than conditioning on n tips at the present time. Gomberg, in an unfinished report on '"Bayesian"' postdiction in an evolution process' (Gomberg 1966, unpublished work), prefers not conditioning on

the present time, although the details are not so clear. The differences among those variants are not well understood.

The characters states at the interior nodes of the tree are known as ancestral states. On a species phylogeny, they represent the states of the characters in the extinct common ancestors of modern species. Edwards, in [46], and also Thompson (1975, p. 119) treated the states at the root ($\boldsymbol{x}_1$) as parameters. This has the problem of introducing many parameters in the model, such that the number of parameters increases without bound when the number of characters increases. For discrete characters such as the nucleotides (T, C, A, and G) in DNA sequences, it is customary to use continuous-time Markov chains to describe transitions between character states and to assume that the process has been stationary; for exceptions, see Yang and Roberts (1995). Then the root states have a distribution given by the stationary distribution of the Markov chain. For continuous characters, the Brownian motion does not have a stationary distribution. Felsenstein (1973) discussed an algorithm to estimate the ancestral states, eliminating $\boldsymbol{x}_1$. Statistical justifications for this procedure were discussed by Thompson (1975, p. 119).

The Yule process component of the model was dropped when Felsenstein (1973a) revisited the problem of phylogeny reconstruction using continuous characters. Thus, the tree (F) and times ($\boldsymbol{t}$) do not have statistical distributions anymore and become true parameters; these are estimated by maximizing the likelihood function, which averages over the ancestral states. Similarly, Thompson (1975, p. 60), in extending the method used by Edwards in [46], dropped the Yule process. These are the early maximum-likelihood implementations of the Brownian-motion model for continuous characters.

Origin of Bayesian Phylogenetics

Felsenstein (2004, p. 291) has included a discussion of early applications of Bayesian or Bayesian-like ideas to phylogenetics. Perhaps the most relevant is the calculation of posterior probabilities for trees by Kishino and Hasegawa (1989); see also Smouse and Li (1987), who calculated the likelihood by optimizing the branch lengths for each tree, while a fully Bayesian approach should average over the branch lengths or branching times. The modern approach to Bayesian inference of molecular phylogenies was introduced by three groups, working independently in the 1990s: Bob Mau and Michael Newton in Wisconsin (Mau and Newton 1997), a research student at Ohio State University, Shuying Li (Li *et al*. 2000), and Bruce Rannala and me in Berkeley (Rannala, and Yang 1996, Yang and Rannala 1997). In all those works, the branching times are integrated out through a prior to calculate the posterior probabilities of trees. The first two groups are statisticians, applying Bayesian Markov chain Monte Carlo (MCMC) algorithms to phylogeny

estimation. Note that the 1990s was the time when Bayesian MCMC algorithms were introduced into various branches of sciences, even though they were developed a few decades earlier (Metropolis 1953, Hastings 1970).

In our case, the motivation must be owed entirely to Edwards's paper, [46]. After finishing my PhD in Beijing in 1992, I went to Cambridge to work with Adrian Friday and Nick Goldman. We occasionally saw Edwards, but I believe he was working on his book on Venn diagrams, rather than phylogenetics. Adrian, Nick, and I were developing Markov models of sequence evolution for use in the maximum-likelihood method, and we had much discussion about whether the tree should be treated as a discrete parameter or a statistical model. I have provided detailed arguments elsewhere that the distinction is not a semantic one (see, e.g., Yang 1997, 2014, pp. 159–163). For example, it is not so clear how to decide whether a log-likelihood difference between two trees is due to chance. We have the rule of thumb (Edwards 1972, p. 202) that an improvement of 2 log-likelihood units (or 1.92 if we use the asymptotic χ^2 distribution) is good enough for including one additional parameter, but we lack such a "calibration" when two trees are compared. The use of bootstrap to evaluate the significance of trees has met with difficulties in interpretation (Zharkikh and Li 1992, Felsenstein and Kishino 1993, Hillis and Bull 1993). I was also concerned that the maximum-likelihood tree topology does not have the large-sample efficiency of the maximum-likelihood estimate of a conventional parameter (see, e.g., Yang 1997). Those concerns motivated my work with Bruce, when both of us were postdocs in Monty Slatkin's phylogenetics laboratory in Berkeley. We were curious to see what the alternative statistical methodology, the Bayesian, might offer, given the difficulties with the maximum likelihood method. We decided to try Edwards's prescription, but with two changes. First, we worked with DNA sequence data, using a continuous-time Markov chain model, instead of the Brownian-motion model, for continuous characters, with summation over the ancestral states achieved using Felsenstein's pruning algorithm (Felsenstein 1973b, 1981). Second, we used the birth–death process (instead of the Yule) to specify a prior on the trees and times, but this was an easy replacement. We used numerical integration to integrate over the times, so that the method is applicable to small trees only. This effort led to the work published by Rannala and Yang (1996).

This is one of the first Bayesian molecular phylogenetic analyses, and the results are interesting. We applied our program to two datasets of four or five ape species (human, common chimpanzee, pygmy chimpanzee, gorilla, and orangutan). We estimated the birth rate λ, death rate μ, and the transition/transversion rate ratio parameter κ by maximum likelihood, and used these parameters to calculate the posterior probabilities for trees, as stipulated by Edwards in [46]. The maximum a-posteriori (MAP) trees are reasonable in both datasets, grouping

the humans with the chimpanzees, but the posterior probability in one dataset, at 0.9999, is uncomfortably high. This dataset, of 11 mitochondrial tRNA genes (739 base pairs) and published by Horai *et al.* (1992), is fairly small, and the human–chimpanzee–gorilla relationship was a hard phylogenetic problem at the time. Spuriously high posterior probabilities for trees continue to trouble us today, especially as datasets for phylogenetic analysis are getting increasingly large (Lewis *et al*. 2005, Yang and Rannala 2005, Yang 2007). The problem appears to have to do with the asymptotic behaviour of Bayesian model selection when applied to opposing and nearly equally wrong models.

Phylogeny Estimation and Statistical Inference

I suppose my characterization of Edwards's paper [46] as the first effort to apply Bayesian statistics (rather than maximum-likelihood methods) to phylogeny estimation might not be looked upon by Edwards himself as reasonable. Indeed in his discussion, (p. 104) he has this to say about the Bayesian approach:

> . . . and a detailed study of the present problem has, if anything, strengthened my conviction that a Bayesian approach in this instance would be a gross over-simplification. I am not prepared to give the true parameters prior probability distributions because I can see no model which would justify them. We may also note that the adoption of a Bayesian approach would not automatically resolve the dilemma of how to summarize a posterior distribution in a great many variables in terms of a few descriptive parameters, for maximizing the posterior probability would lead to the singularities . . .

A few words of clarification may thus be called for, related to the changing usage of the term "Bayesian" and the philosophy of statistical inference. The word "Bayesian" was apparently coined by R. A. Fisher in 1950 (Edwards, 2004, Fienberg 2006) to refer, derogatorily, to the method of *inverse probability*, the inference method that uses Bayes theorem to derive probability distributions for parameters. The word "inverse" refers to the fact that the probability is here defined backwards, from the data to the parameter or hypothesis, or from effects to causes. Given the probability of heads for a fair coin, $\theta = \frac{1}{2}$, say, the probability of $x = 2$ heads in $n = 4$ coin tosses is $\frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = \frac{3}{8}$. This has a simple frequency interpretation: if we do many experiments, each of which involves tossing the coin four times, then in $\frac{3}{8}$ of the experiments, we will see two heads. Now suppose we have observed $x = 2$ heads in $n = 4$ tosses of a coin, what is the probability distribution of θ? To someone who is not a Bayesian statistician, this question is not meaningful. The approach of assigning a prior on θ, based on subjective beliefs or without a physical model, is the inverse method, and Edwards's formulation is not such a method.

In the tree problem, the Yule (or birth–death) process is a biologically plausible even if simple model. The true parameters are the Yule branching rate λ and the

Brownian-drift parameter σ^2, which should be estimated by maximizing the likelihood function, while the tree is a random variable and its realized value should be estimated from the conditional probability, given the data. This is a standard likelihood method for estimating realized values of random variables in the model, nowadays known as empirical Bayes. Edwards would thus consider his method to be a likelihood method (although not a maximum-likelihood method) of phylogeny estimation.

Nevertheless, modern use of the term "Bayesian" does not have the derogatory tone, and the use of a physical/biological process is a common approach to specifying a prior. Of course, one may argue that the full or hierarchical Bayesian approach would assign priors on parameters λ and σ^2, rather than using their maximum-likelihood estimates, but in this article, I have not made an effort to distinguish between the empirical Bayes and the full (hierarchical) Bayes methods, or between a prior based on a biological model and a prior chosen for convenience.

In passing, it may be noted that the challenge of summarizing the posterior distribution of phylogenetic trees still exists today. However, the singularity in the posterior mentioned by Edwards does not exist because the posterior for a tree is calculated by integrating over the ancestral states and branching times.

Modern Times

The early studies of Rananla and Yang (Rannala and Yang 1996, Yang and Rannal 1997). Mau and Newton (1997), and Li *et al.* (2000) assumed the molecular clock (rate constancy over time), which is often violated in comparisons of distant species. Bayesian phylogenetics really took off with the development of the program MrBayes (Huelsenbeck and Ronquist 2001, Ronquist *et al*. 2012), which adapted branch-swapping algorithms such as nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR) (Swofford *et al*. 1996) into MCMC proposal algorithms to move between trees. The clock constraint was relaxed, enabling phylogenetic inference to be conducted under more realistic models. A more recent program, BEAST, infers rooted trees under the clock, and relaxed-clock models (Drummond and Rambaut 2007), while PhyloBayes implements sophisticated non-stationary models to deal with substitution heterogeneity among lineages, which may be important for deep phylogenies (Lartillot *et al*. 2009). Nowadays, those Bayesian programs are standard tools in molecular phylogenetics, together with fast likelihood programs such as RAxML (Stamatakis 2006) and PhyML (Guindon and Gascuel 2003).

A brief introduction to Bayesian phylogenetics is provided in Yang (2016). More extensive recent reviews include Zwickl and Holder (2004) and Yang (2014, chs. 8

and 9). An edited book by Chen *et al*. (2014) summarizes current research topics in the field. Phylogenetics may well be the largest application area of Bayesian statistics. It provides a rich testing ground for advanced Monte Carlo computational algorithms. Jerzy Neyman (1971) was certainly right to identify molecular phylogenetics as "a source of novel statistical problems."

References

Chen, M.-H., Kuo, L., and Lewis, P. 2014. *Bayesian Phylogenetics: Methods, Algorithms, and Applications*. London: Chapman & Hall/CRC.

Drummond, A. J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7, 214.

Edwards, A. W. F. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Edwards, A. W. F. 2004. Comment on Bellhouse, David R. 'The Reverend Thomas Bayes FRS: a biography to celebrate the tercentenary of his birth. *Statistical Science* 19, 34–37.

Felsenstein, J. 1973a. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* 25, 471–492.

Felsenstein, J. 1973b. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* 22, 240–249.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17, 368–376.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland, MA. Sinauer Associates.

Felsenstein, J. and Kishino, H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* 42, 193–200.

Fienberg, S. E. 2006. When did Bayesian inference become "Bayesian"? *Bayesian Analysis* 1, 1–40.

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52, 696–704.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their application. *Biometrika* 57, 97–109.

Hillis, D. M. and Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42, 182–192.

Horai, S., Satta Y., Hayasaka, K., *et al*. 1992. Man's place in Hominoidea revealed by mitochondrial DNA genealogy [Erratum *J Mol Evol* 1993; 37:89]. *Journal of Molecular Evolution* 35, 32–43.

Huelsenbeck, J. P. and Ronquist, F. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.

Kendall, D. G. 1948. On the generalized birth-and-death process. *Annals of Mathematical Statistics* 19, 1–15.

Kingman, J. F. C. 1982. The coalescent. *Stochastic Processes and their Applications* 13, 235–248.

Kishino, H. and Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution* 29, 170–179.

Lartillot, N. and Poujol, R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution* 28, 729–744.

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

Lewis, P. O., Holder, M. T., and Holsinger, K. E. 2005. Polytomies and Bayesian phylogenetic inference. *Systematic Biology* 54, 241–253.

Li, S., Pearl, D., and Doss, H. 2000. Phylogenetic tree reconstruction using Markov chain Monte Carlo. *Journal of the American Statistical Association* 95, 493–508.

Mau, B. and Newton, M. A. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* 6, 122–131.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–1092.

Neyman, J. 1971. In *Statistical Decision Theory and Related Topics*, eds. S. S. Gupta and J. Yackel, New York: Academic Press, 1–27.

Rannala, B. and Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution* 43, 304–311.

Ronquist, F. Teslenko, M., van der Mark P., *et al*. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61, 539–542.

Smouse, P. E. and Li, W.-H. 1987. Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution* 41, 1162–1176.

Solis-Lemus, C., Knowles, L. L. and Ane, C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution* 69, 492–507.

Stadler, T. 2010. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* 267, 396–404.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. 1996. In *Molecular Systematics*, eds. D. M. Hillis, C. Moritz, and B. K. Mable, Sunderland, MA: Sinauer Associates, 407–514.

Thompson, E. A. 1975. *Human Evolutionary Trees*. Cambridge: Cambridge University Press.

Yang, Z. 1997. How often do wrong models produce better phylogenies? *Molecular Biology and Evolution* **14**, 105–108.

Yang, Z. 2007. Fair-balance paradox, star-tree paradox and Bayesian phylogenetics. *Molecular Biology and Evolution* **24**, 1639–1655.

Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford: Oxford University Press.

Yang, Z. 2016. In *Encyclopedia of Evolutionary Biology*, ed. R. M. Kliman, New York: Elsevier, 137–140,

Yang, Z. and Roberts, D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution* **12**, 451–458.

Yang, Z. and Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Molecular Biology and Evolution* **14**, 717–724.

Yang, Z. and Rannala, B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology* **54**, 455–470.

Zharkikh, A. and Li, W.-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* **9**, 1119–1147.

Zwickl, D. J. and Holder, M. T. 2004. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Systematic Biology* **53**, 877–888.