Exact Bayesian evidence for phylogenetic trees

Matthew D. Schwartz*[1], Scott V. Edwards[2,3], and Paul O. Lewis[4]

[1]*Department of Physics, Harvard University*
[2]*Department of Organismic and Evolutionary Biology, Harvard University*
[3]*Museum of Comparative Zoology, Harvard University*
[4]*Department of Ecology and Evolutionary Biology, University of Connecticut*

This manuscript is preliminary. The first draft was written by Claude (Anthropic), and the text and computations are still under review by the authors.

Abstract

In Bayesian phylogenetics, evolutionary trees are inferred from DNA sequences as summaries of a posterior distribution over trees. Competing substitution models, and for a handful of species competing trees, are compared through their marginal likelihoods: the probability of the data under each, with branch lengths and other parameters integrated out. These marginal likelihoods are estimated by long Monte Carlo runs often with large uncontrolled errors. We show in many cases no estimate is needed: the marginal likelihood is an exact rational number, a ratio of explicit integers, computable by methods imported from the exact evaluation of Feynman integrals in particle physics. This holds for 4- and 5-taxon trees under the Jukes-Cantor, Kimura two-parameter, and any other time-reversible model. Including rate variation across sites, or the fractional powers of the likelihood that stepping-stone estimators use, the values are still exactly computable but are transcendental rather than rational. These methods can then be used to evaluate trees and compare substitution models. We also lay out how exact marginal likelihoods can be used to build larger trees, to combine many linked or unlinked genes, to flag genes corrupted by contamination or misassembly, to measure how a conclusion depends on the substitution model, and to adjudicate disagreements between data sets. This paper establishes the method, tests it on a number of known phylogenies as well as some unresolved controversies, and compares exact marginal likelihoods to a number of Monte Carlo tools. For example, two of the three resolutions of a hominid tRNA quartet are exactly tied, and the third, which the quartet nominally prefers although no site in the alignment supports it, ranks below both once a fifth species is added. We find that a widely used Monte Carlo tool, MrBayes, shows a small systematic offset, invisible when using it alone but apparent when compared to the exact result. For the sizes of data sets analyzed here the exact evaluation takes about as much computing time as the Monte Carlo runs, and that time grows as the fifth power of the alignment's pattern count. Applications discussed here include the placement of the extinct San Cristóbal tortoise and the chromosome-level discordance in the malaria-mosquito genome.

Keywords: [to be supplied]

*Work done with Anthropic, but results and views are not endorsed by Anthropic.

1 Introduction

Bayesian inference (Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001) is now a standard way to estimate phylogenies and to compare models of sequence evolution. It evaluates tree topologies by their posterior probabilities. Under a uniform prior on topologies, each such probability is proportional to the marginal likelihood of the topology, or evidence: the likelihood of the alignment averaged over the prior on branch lengths. The ratio of two marginal likelihoods is a Bayes factor (Kass and Raftery, 1995), used to compare trees as well as substitution models (Xie et al., 2011; Baele et al., 2012; Oaks et al., 2019).

For a topology T and an alignment $\mathbf{y}$ summarized by its site-pattern counts, the marginal likelihood is

$$Z(T \mid \mathbf{y}) \;=\; \int p(\mathbf{y} \mid T, t)\, \pi(t \mid T)\, \mathrm{d}t, \tag{1}$$

where t denotes the branch lengths of T, $p(\mathbf{y} \mid T, t)$ the likelihood computed by pruning (Felsenstein, 1981), and $\pi(t \mid T)$ the branch-length prior. The integral has one dimension per branch and has been regarded as analytically intractable for all but trivial cases (Xie et al., 2011; Oaks et al., 2019). In practice it is estimated from Markov chain Monte Carlo (MCMC) output, by the harmonic-mean estimator (Newton and Raftery, 1994), by thermodynamic integration (also called path sampling; Lartillot and Philippe, 2006; Friel and Pettitt, 2008), by stepping-stone sampling and its generalization (Xie et al., 2011; Fan et al., 2011; Baele et al., 2012), or by LoRaD (Wang et al., 2023). The harmonic mean needs only a sample from the posterior, but its variance can be infinite and it tends to overestimate the marginal likelihood, favoring parameter-rich models (Lartillot and Philippe, 2006; Xie et al., 2011; Baele et al., 2012). Thermodynamic integration and stepping-stone sampling instead sample a sequence of distributions between the prior and the posterior, with the likelihood raised to a power between zero and one, and are much more accurate at greater computational cost (Lartillot and Philippe, 2006; Xie et al., 2011). The Monte Carlo error of such an estimate is routinely reported. Thermodynamic integration has in addition a discretization bias, which depends on the number and placement of the powers and is not reduced by running longer chains at each of them (Xie et al., 2011; Friel and Pettitt, 2008). The LoRaD estimator (Wang et al., 2023) needs only a sample from the posterior, as the harmonic mean estimator does, but uses the draws in a high-density region of parameter space alone and is as accurate as stepping-stone sampling at a fraction of its cost; in the comparisons of Section 4.3 its estimates lie within error of the exact values.

The bias of an estimator, unlike its Monte Carlo error, cannot be read from its own output; it has to be checked against a value known by other means. The estimators have been tested against the analytic value of a conjugate normal model (Xie et al., 2011), against one another on real alignments (Wang et al., 2023), and against long reference MCMC runs (Fourment et al., 2020). In each such test, however, the reference value either is not a phylogenetic marginal likelihood or is itself an estimate. Direct evaluation has been possible only in special cases: Rannala and Yang (1996) obtained posterior probabilities for four hominoid species by numerical integration over node times, an approach limited to about five taxa (Yang and Rannala, 1997). Sinsheimer et al. (1996) tested four-taxon topologies under evolutionary parsimony, which involves no branch-length integral. The likelihood function itself has been treated analytically on the smallest trees (Chor et al., 2000; Chor and Snir, 2004), and analytic results for the posterior are known in the star-tree limit (Yang and Rannala, 2005; Lewis et al., 2005; Steel and Matsen, 2007; Susko, 2008). In none of these settings is the marginal likelihood of a resolved tree on a real alignment

known exactly.

Although few studies concern only four or five species, trees of that size arise in large numbers in current practice. Quartet-based methods assemble a tree from the topologies inferred for its subsets of four taxa (Ranwez and Gascuel, 2001). Under the multispecies coalescent, different loci may support different topologies (Pamilo and Nei, 1988; Degnan and Rosenberg, 2006), and summary methods for the species tree score candidate trees by the four-taxon topologies that the individual gene trees induce (Zhang et al., 2018). Genome scans for introgression build a tree for each short window of an alignment of a few species and follow the changes in preferred topology along the chromosomes (Fontaine et al., 2015).

Here, we show that for trees of four and five taxa, with independent exponential priors of rational rate on the branch lengths, $Z(T \mid \mathbf{y})$ is a rational number, a ratio of two integers, and we compute it exactly. This holds under the Jukes–Cantor model (Jukes and Cantor, 1969) and, with rate parameters and base frequencies fixed at rational values, under the Kimura two-parameter (Kimura, 1980) and general time-reversible models as well. Under the Jukes–Cantor model, pruning makes each site-pattern probability a polynomial in the quantities $x_e = e^{-\mu t_e}$, one for each branch e of length t_e, that is linear in each x_e separately; Kimura's models need two or three such exponentials on each branch, and the time-reversible case rests on a matrix identity instead. This multilinearity underlies Hadamard conjugation and phylogenetic invariants (Hendy and Penny, 1989; Székely et al., 1993; Sturmfels and Sullivant, 2005; Allman and Rhodes, 2006). The likelihood is then a product of such polynomials, and the prior contributes a power of each x_e. Since each x_e lies between zero and one, the integral runs over the unit interval in every variable and can be done term by term, giving a finite sum of fractions. At the length of a transfer-RNA gene, that sum has many millions of terms, and its value is a ratio of integers several hundred digits long. In the variables x_e the evidence has the form of a Feynman integral in parametric representation (Smirnov, 2012), known in algebraic statistics as an A-hypergeometric integral (Gel'fand et al., 1990; Sturmfels and Telen, 2021). Borinsky et al. (2023) wrote marginal likelihoods of toric statistical models in this form and evaluated them by Monte Carlo sampling. We instead evaluate the phylogenetic integral exactly, with modular-arithmetic methods adopted from work on Feynman integrals in particle physics (von Manteuffel and Schabinger, 2015).

We compute exact values for twelve quartets of transfer-RNA genes, among them the tRNA-Gln gene of mouse, rat, chicken and *Xenopus*, and for several other published alignments, the longest of nearly four hundred sites. They serve first as references for stepping-stone sampling in MrBayes (Ronquist et al., 2012) and RevBayes (Höhna et al., 2016) and for the LoRaD estimator, so that the bias of each estimator can be told apart from its Monte Carlo error. In the MrBayes estimates they reveal a small offset, common to the three topologies of an alignment, that persists when more stones or longer chains are used. Exact values can also confirm ties, and resolve margins far below the Monte Carlo error of a sampling run. For the mitochondrial tRNA-Phe gene of human, chimpanzee, gorilla and orangutan, two of the three topologies have identical marginal likelihoods, with the same numerator and denominator, and the third is narrowly preferred although no site supports it. When gibbon is added and all fifteen five-taxon topologies are evaluated, that third topology ranks below a pair that remains exactly tied, one member of which is the accepted hominoid tree. Finally, we consider how exact single-locus values combine over many loci, with applications to the placement of the extinct San Cristóbal giant tortoise (Jensen et al., 2022) and to topological discordance across the genomes of the *Anopheles gambiae* complex (Fontaine et al., 2015).

Such values are exact only for the stated model and prior, and are within reach only for

small trees and for alignments with few variable sites. The cost grows as the fifth power of the number of **mixed sites**, the variable sites other than those whose pattern supports the topology under evaluation. For quartets of transfer-RNA length the three marginal likelihoods take from seconds to about seventy minutes on a multicore workstation, depending on the number of mixed sites, and considerably more processor time than a stepping-stone run. The term-by-term evaluation, and with it the guarantee of a rational value, is lost once rate variation across sites, or a substitution parameter shared across branches with a continuous prior, is integrated out. We treat the first only in the smallest cases (three taxa, one or two sites) and the second not at all. For trees of realistic size and the models in routine use the marginal likelihood must therefore still be estimated, and the exact values for small trees can then serve as references against which the estimators are tested.

2 The marginal likelihood of a small tree

For a small tree the candidate topologies can be enumerated, and Bayesian inference of the topology reduces to computing the marginal likelihood, or evidence, of each. This section contains no new results. We first recall the Monte Carlo estimators of this quantity and what is known of their errors. We then write the evidence of a four-taxon tree under the Jukes–Cantor model as the integral of a polynomial over the unit box, which is the form we evaluate exactly.

2.1 The evidence and its estimators

As in Eq. (1), $\mathbf{y}$ denotes an alignment summarized by its site-pattern counts, and the evidence $Z(T \mid \mathbf{y})$ for topology T is the likelihood averaged over the prior on branch lengths and any free substitution-model parameters. With prior probabilities $\pi(T)$ on the topologies, the posterior probability of a topology is

$$P(T \mid \mathbf{y}) \;=\; \frac{\pi(T)\, Z(T \mid \mathbf{y})}{\sum_{T'} \pi(T')\, Z(T' \mid \mathbf{y})}. \tag{2}$$

Four taxa have three unrooted topologies, and under the uniform prior used throughout this paper ($\pi(T) = \frac{1}{3}$ at four taxa) the posterior is the normalized vector of the three marginal likelihoods. Ratios of marginal likelihoods are Bayes factors (Kass and Raftery, 1995), reported here as natural logarithms (log units).

The harmonic-mean estimator (Newton and Raftery, 1994) needs only the likelihoods of an MCMC sample from the posterior, but its variance can be infinite and it tends to overestimate Z, favoring parameter-rich models (Lartillot and Philippe, 2006; Xie et al., 2011; Baele et al., 2012). The estimators now in standard use are based on the power posterior, proportional to $p(\mathbf{y} \mid T, t)^{\beta}\pi(t)$ for $0 \le \beta \le 1$, whose normalizing constant Z_β runs from $Z_0 = 1$ to $Z_1 = Z$ along a schedule $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$:

$$Z_\beta(T \mid \mathbf{y}) \;=\; \int p(\mathbf{y} \mid T, t)^{\beta}\, \pi(t)\, \mathrm{d}t, \qquad Z \;=\; \prod_{k=1}^{K} \frac{Z_{\beta_k}}{Z_{\beta_{k-1}}}. \tag{3}$$

Thermodynamic integration, or path sampling (Lartillot and Philippe, 2006; Friel and Pettitt, 2008), estimates $\log Z = \int_0^1 \mathbb{E}_\beta[\log p(\mathbf{y} \mid T, t)]\, \mathrm{d}\beta$ by quadrature over the schedule, with the expectation at each β_k replaced by the average of the sampled log-likelihoods from a chain run

at that β_k. Stepping-stone sampling (Xie et al., 2011) estimates each ratio in the product by importance sampling from the chain at β_{k-1}, with the β_k (the stones) crowded toward the prior, where the ratios change fastest. Generalized stepping-stone sampling (Fan et al., 2011) starts the path from a reference distribution fitted to the posterior instead of the prior. The LoRaD method (Wang et al., 2023) uses the posterior sample alone, restricted to a region of highest posterior density to keep its variance finite. Stepping-stone sampling is implemented in MrBayes (Ronquist et al., 2012) and, together with path sampling, in RevBayes (Höhna et al., 2016) and BEAST (Suchard et al., 2018; Baele et al., 2012); LoRaD is available as the R package lorad (Milkey et al., 2023).

On a conjugate normal model whose marginal likelihood is known analytically (Xie et al., 2011), stepping-stone sampling recovered the value within Monte Carlo error, thermodynamic integration fell slightly below it, and the harmonic mean was almost one log unit too high. The shortfall of thermodynamic integration is a discretization bias: it depends on the number and placement of the β_k and is not reduced by running longer chains at each β_k (Xie et al., 2011; Friel and Pettitt, 2008). On phylogenetic problems the reference value has instead been the output of another estimator (Wang et al., 2023) or of a very long chain, as in the comparison of nineteen methods by Fourment et al. (2020). In none of these tests was the reference an exactly known marginal likelihood of a tree on real sequence data. The Monte Carlo error of a stepping-stone or path-sampling estimate is routinely reported; its bias, however, has not been measured against such a value. In Section 4 we compute the exact marginal likelihoods of the 68-site tRNA-Gln quartet and use them to measure the bias of stepping-stone estimates from MrBayes, RevBayes, and LoRaD estimates.

2.2 The Jukes–Cantor quartet integral

Consider the unrooted quartet $T = 12|34$ with pendant edges $e_1, \dots, e_4$, internal edge e_5, and branch lengths t_e in expected substitutions per site. Under the Jukes–Cantor (JC69) model (Jukes and Cantor, 1969) the transition probabilities along an edge e depend on t_e only through the edge variable $x_e = e^{-\mu t_e}$, with $\mu = \frac{4}{3}$. The variable x_e is the probability that a site is not randomized along e ($x_e = 1$ for a branch of length zero, $x_e = 0$ for an infinitely long one). Pruning (Felsenstein, 1981) then gives the probability of site pattern $s = (s_1, \dots, s_4)$ as a sum over the states a, b at the two internal nodes,

$$\begin{aligned} p_s(x) \;&=\; \frac{1}{4} \sum_{a,b} P_{as_1}(x_1)\, P_{as_2}(x_2)\, P_{ab}(x_5)\, P_{bs_3}(x_3)\, P_{bs_4}(x_4), \\ P_{ii}(x) &= \tfrac{1}{4} + \tfrac{3}{4}\, x, \quad P_{ij}(x) = \tfrac{1}{4} - \tfrac{1}{4}\, x \;\; (i \neq j). \end{aligned} \tag{4}$$

Each term of this sum contains one factor per edge, linear in that edge's x_e. Hence p_s is a multilinear polynomial, of degree at most one in each of the five edge variables, with integer coefficients after multiplication by 256. Because the model treats the four nucleotides alike, the $4^4 = 256$ patterns fall into fifteen classes, the set partitions of $\{1, 2, 3, 4\}$ from $xxxx$ to $xyzw$. We write $p_k(x)$, $k = 1, \dots, 15$, for the class polynomials and m_k for the class sizes (Table A1, Appendix A); the fifteen p_k share thirteen monomials and satisfy $\sum_k m_k\, p_k(x) \equiv 1$.

For an alignment of N independent sites with class counts $\mathbf{y} = (y_1, \dots, y_{15})$, $\sum_k y_k = N$, the likelihood is $\prod_k p_k(x)^{y_k}$, a polynomial of degree at most N in each x_e. We place on the five branch lengths independent exponential priors $\pi(t_e) = \lambda e^{-\lambda t_e}$ with rate λ (mean $1/\lambda$), the unconstrained exponential prior of MrBayes and RevBayes. In the edge variable this prior is a

power of x_e,

$$\lambda e^{-\lambda t_e}\,\mathrm{d}t_e \;\longrightarrow\; \alpha\, x_e^{\alpha-1}\,\mathrm{d}x_e \quad \text{on } [0,1], \qquad \alpha = \frac{\lambda}{\mu} = \frac{3\lambda}{4}, \tag{5}$$

and the evidence becomes an integral over the unit box,

$$Z(T;\mathbf{y},\alpha) \;=\; \alpha^5 \int_{[0,1]^5} \prod_{k=1}^{15} p_k(x)^{y_k} \prod_{e=1}^{5} x_e^{\alpha-1}\,\mathrm{d}x. \tag{6}$$

The factor α^5, the normalizing constant of the prior, is common to the three topologies and cancels from the Bayes factors and the posterior probabilities. We take $\alpha = 1$ throughout, that is $\lambda = \frac{4}{3}$, which in the parameterization of MrBayes and RevBayes corresponds to an exponential branch-length prior with rate parameter $\frac{4}{3}$ (mean 0.75 expected substitutions per site); each x_e is then uniform on $[0,1]$. Other values of α are used only in Sections 4.2 and 5.2, to examine how two hominoid tRNA-Phe margins depend on the prior mean. The marginal likelihoods of 13|24 and 14|23 are the same integral with the taxon labels permuted in $\mathbf{y}$.

3 Exact evaluation

Under exponential branch priors of rational rate, or gamma priors of integer shape and rational rate, the quartet evidence is a rational number. The proof is constructive and gives the algorithm by which we compute it; for a twenty-site example we print the fraction in full. On real alignments the cost is governed by the number of mixed sites, those neither constant nor of the class supporting the topology under evaluation ($xxyy$ for 12|34); whether a given value in this paper could be computed exactly depends on that number. Rate variation across sites and tempered likelihoods, under which the evidence is presumably no longer a rational number, are treated last (Table 1).

3.1 Rationality of the evidence

The following lemma uses only the fact that the quartet integrand is a polynomial in the five edge variables multiplied by the prior density (we write $N = \sum_k y_k$ for the number of sites).

Lemma 1. *For every count vector $\mathbf{y} \in \mathbb{Z}_{\geq 0}^{15}$ and every rational $\alpha = p/q > 0$ in lowest terms, the evidence $Z(T;\mathbf{y},\alpha)$ of Eq. (6) is a rational number, and the integral $\alpha^{-5} Z(T;\mathbf{y},\alpha)$ has denominator dividing*

$$D_N(\alpha) \;=\; 256^N \prod_{e=1}^{5} \prod_{j=0}^{N} (jq + p). \tag{7}$$

Proof. Each pattern polynomial p_k is multilinear in $x_1, \ldots, x_5$ with coefficients in $256^{-1}\mathbb{Z}$, so the product $\prod_k p_k^{y_k}$ is a polynomial of degree at most N in each x_e; write it as $256^{-N} \sum_a c_a\, x_1^{a_1} \cdots x_5^{a_5}$, where the sum runs over the exponent vectors $a = (a_1, \ldots, a_5) \in \{0, 1, \ldots, N\}^5$ of its monomials and the coefficients c_a are integers. Integrating term by term against the prior,

$$Z(T;\mathbf{y},\alpha) \;=\; \frac{\alpha^5}{256^N} \sum_a c_a \prod_{e=1}^{5} \int_0^1 x_e^{a_e+\alpha-1}\,\mathrm{d}x_e \;=\; \frac{\alpha^5}{256^N} \sum_a c_a \prod_{e=1}^{5} \frac{1}{a_e + \alpha}\,, \tag{8}$$

a finite sum of rational numbers whose denominators all divide $D_N(\alpha)$, since $1/(a_e+\alpha) = q/(a_e q+p)$. □

The term-by-term argument above also covers the other group-based models, the Kimura two- and three-parameter models (Kimura, 1980), whose pattern probabilities are multilinear in at most three exponentials per edge. At fixed rational rates (for the two-parameter model, a fixed rational transition/transversion ratio κ) the evidence is again rational. More generally, the proof needs only a prior that is independent across branches and has rational moments in the edge variables. The exact Kimura-type values used in Section 6.1 rely on this more general form. There each of the five branches carries its own transition and transversion rates, giving two edge variables per branch, ten in all, rather than one ratio κ for the tree. Writing v_e for the length of branch e in expected transversions per site and κ_e for its ratio of transition to transversion rate, the two variables are $x_e = e^{-2v_e}$, which coincides with the Jukes–Cantor variable when $\kappa_e = 1$, and $z_e = e^{-(1+\kappa_e)v_e} = x_e^{(1+\kappa_e)/2}$. In these variables the probabilities of no change, of a transition and of each transversion along the branch are $\frac{1}{4}+\frac{1}{4}x_e+\frac{1}{2}z_e$, $\frac{1}{4}+\frac{1}{4}x_e-\frac{1}{2}z_e$ and $\frac{1}{4}-\frac{1}{4}x_e$. The prior is uniform on the region of the unit square corresponding to nonnegative rates,

$$\pi(x_e, z_e) \;=\; \tfrac{3}{2} \qquad \text{on} \quad 0 < z_e \le \sqrt{x_e} \le 1, \qquad e = 1, \ldots, 5, \tag{9}$$

independently across branches; the region has area $\frac{2}{3}$, and the Jukes–Cantor model is its diagonal $z_e = x_e$. A single κ shared by all branches and integrated over a continuous prior is not covered, however: the value is then an integral of a rational function of κ against the prior, which need not be rational.

The general time-reversible (GTR) model, whose transition matrix $P(t) = e^{Qt}$ is usually computed from generically irrational eigenvalues of Q, needs a different argument.

Lemma 2 (GTR)**.** *Consider any time-reversible model with rate matrix Q, built from exchangeabilities and stationary frequencies, and independent exponential branch-length priors of rate λ. The evidence $Z(T;\mathbf{y})$ is a rational function of the exchangeabilities, the stationary frequencies and λ with rational coefficients, and in particular a rational number at rational parameter values.*

The proof (Appendix A) rests on one identity. Because the likelihood is linear in each branch's N-site transition matrix $P_e(t_e)^{\otimes N}$ and the priors are independent across branches, the prior expectation may be taken one branch at a time. The N sites on a branch evolve jointly under $Q^{\oplus N} = \sum_{i=1}^{N} I \otimes \cdots \otimes Q \otimes \cdots \otimes I$ on 4^N states, and

$$\int_0^\infty P(t)^{\otimes N}\, \lambda e^{-\lambda t}\, \mathrm{d}t \;=\; \lambda\big(\lambda I - Q^{\oplus N}\big)^{-1} \tag{10}$$

is a resolvent identity, the matrix form of the scalar formula $\int_0^\infty e^{qt}\lambda e^{-\lambda t}\,\mathrm{d}t = \lambda/(\lambda - q)$ for a rate $q < \lambda$. The right side, and the subsequent pruning (Felsenstein, 1981) against the stationary frequencies, involve only sums and products of the entries of Q and one matrix inversion, so that no eigenvalue of Q enters the result. The identity is a proof device rather than a practical algorithm (the matrix has 4^N rows). We have compared exact values obtained from each lemma with conventional spectral computations only on synthetic alignments of three to five sites for three and five taxa, including one case at each taxon number under a time-reversible rate matrix with three irrational eigenvalues (Supplement S3).

Both lemmas extend to any number of taxa. On an unrooted binary tree with n leaves the Jukes–Cantor pattern probabilities are multilinear in the $2n-3$ edge variables, with coefficients

in $4^{-n}\mathbb{Z}$, and Lemma 1 then holds with 256^N replaced by 4^{nN} and five edges by $2n-3$; the proof of Lemma 2 uses only that the pruning sum is multilinear in the per-edge transition matrices, which holds on any tree. A gamma prior on t_e with shape k and rate λ may also replace the exponential (its density in x_e is given in Appendix A). The term-by-term integral is again elementary,

$$\int_0^1 x^{a+\alpha-1}(-\log x)^{k-1}\,\mathrm{d}x = \frac{\Gamma(k)}{(a+\alpha)^k}, \qquad Z \;=\; \alpha^{k|E|}\sum_a \frac{c'_a}{\prod_e (a_e+\alpha)^k}, \quad c'_a \in \mathbb{Q}, \tag{11}$$

with $|E|$ the number of edges and $\alpha = \lambda/\mu$. For integer shape the evidence is a fraction, and for fractional shape (e.g. $k = 1/2$) an algebraic number.

3.2 Computing the exact value

The proof of Lemma 1 is already an algorithm (Fig. 1; Appendix A states it formally, with the fifteen polynomials). We scale the polynomials to integer coefficients, $P_k = 256\,p_k$, and store each as a small five-dimensional array indexed by the exponents of the edge variables. Multiplying the likelihood out one site at a time is then a sequence of N five-dimensional convolutions, ending in the coefficient array c_a of the proof, with at most $(N+1)^5$ entries. Finally we replace each monomial by $\prod_e (a_e+\alpha)^{-1}$ and sum the terms.

In practice we clear denominators with D_N, which makes $D_N\,\alpha^{-5}Z$ an integer, and because the c_a run to hundreds of digits we do all arithmetic modulo a prime just below 2^{31}, so that every intermediate quantity fits in a machine word. Repeating the computation for several primes and applying the Chinese remainder theorem gives the unique integer that lies below the product of the primes and has the computed remainders; once the product exceeds D_N, this integer is $D_N\,\alpha^{-5}Z$ itself. Thus the number of primes needed is known beforehand, and the whole procedure is standard computer algebra (von zur Gathen and Gerhard, 2013; von Manteuffel and Schabinger, 2015), with no floating-point operation at any stage. Because the coefficient array is an N-fold convolution of fifteen fixed $2\times2\times2\times2\times2$ arrays, a five-dimensional discrete Fourier transform over the exponent lattice, taken modulo the working prime, turns it into a pointwise product that can be summed without storing the array (Supplement S3). This transform acts on the multi-site coefficient array and is not the Hadamard conjugation of Hendy and Penny (1989), which acts on the pattern frequencies of a single site; the two share only the group-based structure of the model (Székely et al., 1993).

A twenty-site example. Consider the synthetic alignment on $T = 12|34$ with pattern counts $xxxx\colon 10$, $xxyy\colon 4$, $xyxy\colon 2$, $xyyx\colon 2$, $xxxy\colon 1$ and $xyzw\colon 1$ (all other classes zero, $N = 20$), at $\alpha = 1$. The integrand is $256^{-20}P_{xxxx}^{10}P_{xxyy}^{4}P_{xyxy}^{2}P_{xyyx}^{2}P_{xxxy}P_{xyzw}$, of degree at most twenty in each variable, and therefore expands into at most $21^5 = 4{,}084{,}101$ monomials, with a common denominator dividing $D = 256^{20}\,\mathrm{lcm}(1,\dots,21)^5 \approx 9.99\times10^{89}$. The procedure above returns

$$Z(12|34;\mathbf{y}) = \frac{8390002712695928886592363104613077773344802548241}{975764089122942807922559239182988593336005689311686648380848617182632879055070822400000} \tag{12}$$

in lowest terms (the denominator is exactly $D/2^{10}$), with $Z \approx 8.6\times10^{-39}$ and $\log Z(12|34;\mathbf{y}) = -87.6492$; the computation took 14 s on a workstation. Exchanging taxa 1 and 2 leaves $\mathbf{y}$ unchanged and interchanges the topologies 13|24 and 14|23. Those two marginal likelihoods are

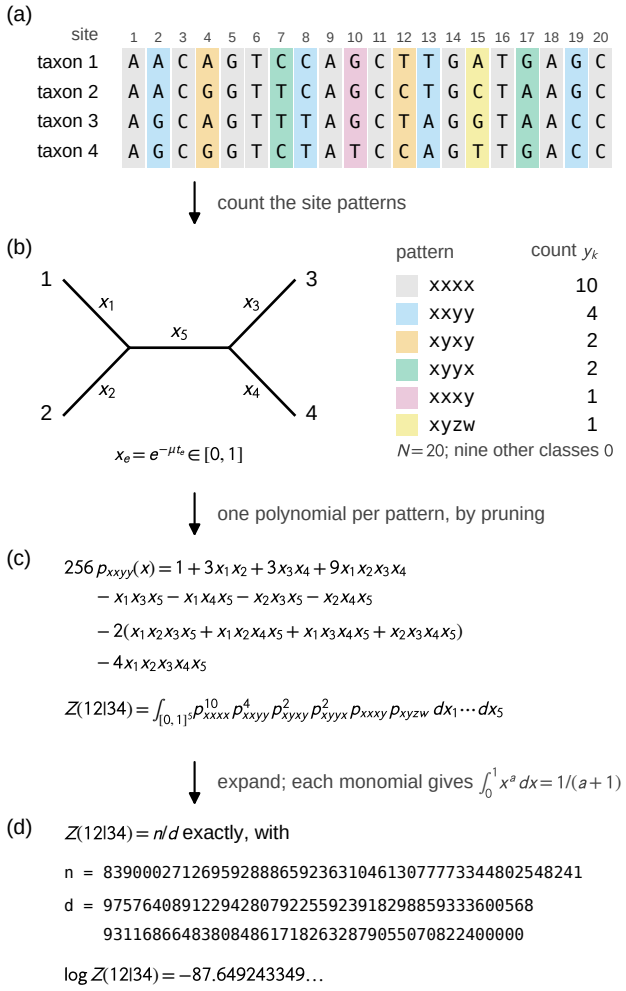

Figure 1: The marginal likelihood of a quartet is a finite sum of fractions, shown on the twenty-site example of Eq. (12). (a) A four-taxon alignment with the pattern counts $\mathbf{y}$ of that example, columns tinted by Jukes–Cantor pattern class (nucleotides are arbitrary representatives of each class). (b) The topology 12|34 with edge variables $x_e = e^{-\mu t_e}$ and the class counts y_k. (c) Pruning turns each class into a polynomial of degree at most one in each x_e with coefficients that are integers divided by 256 (shown for $xxyy$); the likelihood is the product of these polynomials, each raised to its count. (d) Under the exponential prior with $\alpha = 1$ each monomial integrates to a product of reciprocals of integers, giving the fraction of Eq. (12).

consequently equal as fractions, and the Bayes factor against either is an exact ratio of integers,

$$\mathrm{BF} = \frac{Z(12|34;\mathbf{y})}{Z(13|24;\mathbf{y})} = \frac{8390002712695928886592363104613077773344802548241}{3460407108832207189367527631015414554765886136 38} = 24.2457\ldots, \tag{13}$$

or $\log \mathrm{BF} = 3.1882$.

The mixed-site reduction. In real alignments most columns are constant and, for the topology 12|34, most of the rest fall in the class $xxyy$ that supports it. Sites of these two classes add little to the cost, because at such a site the two leaves of each cherry (sister pair) have the same nucleotide and therefore either both agree or both disagree with their common ancestor. In the pruning sum such a site contributes only products in which edges 1 and 2 (and likewise 3 and 4) carry the same one of the two Jukes–Cantor factors, $(1 + 3x_e)/4$ or $(1 - x_e)/4$. The product over these sites is therefore indexed by three integers, however many such sites there are. Only the remaining *mixed* sites, m in number, whose pattern is neither constant nor the evaluated topology's own $xxyy$ class, need the five-index expansion. The evidence becomes a single exact sum over products of entries of a three-index array, a five-index array of side $m+1$ and rational per-edge integrals, with cost governed by $(N - m + 1)(m + 1)^5$ in place of $(N + 1)^5$ once m exceeds about ten (Section 3.3). Most values on real alignments below were obtained this way.

Recurrences. The marginal likelihoods of alignments that differ by a few sites of one pattern are linearly related, much as Pascal's rule relates neighboring binomial coefficients. Along one count y_j the relation is $\sum_{i=0}^{r} c_i(y_j)\, Z(y_j + i) = 0$, with polynomial coefficients fixed by the tree and the model (Zeilberger, 1990). For the Jukes–Cantor quartet along the count of constant sites we found such a recurrence, of order $r = 14$, by modular linear algebra on exactly computed values, and verified it on values not used to find it (Supplement S2); it is not derived symbolically. We do not use it here, because its coefficients depend also on the other fourteen counts and, for a general alignment, determining them costs more than direct evaluation.

3.3 Cost and scaling

The cost of exact evaluation depends on two properties of the alignment: its length N after columns with gaps or ambiguity codes are removed, and the mixed-site count m. The "pattern count" of the Abstract, whose fifth power sets the cost, is m, and in real alignments it is typically much the smaller of the two. For example, the hominid tRNA-Phe quartet has $N = 69$ and $m = 10$, and the tRNA-Gln quartet $N = 68$ and $m = 25$. Of the longer alignments of Section 4, the lysozyme quartet has $N = 390$ but $m = 39$, whereas the two 900-site mitochondrial alignments have $m = 201$ and 357.

We timed exact evaluation and stepping-stone sampling on the tRNA-Gln alignment at the settings of Section 4.3. Each MrBayes stepping-stone analysis of one topology (two independent single-chain runs) took 8–9 s on one processor core, or 24–27 s for the three topologies. The three exact values took 75.5 s of wall time on 72 worker processes of a shared multicore workstation. The computation comprised 216 modular evaluations at a mean of 22.8 s, about 4.9×10^3 core-seconds, half of them a duplicate evaluation of each value on a disjoint set of primes. Exact evaluation thus takes roughly three times the sampler's wall time on seventy-two times as many cores, or about two hundred times its processor time (one hundred without the duplicate evaluation). The tRNA-Phe alignment ($m = 10$) took 7.2 s on the same workers, and

the largest of the twelve tRNA quartets ($m = 59$) about seventy minutes on 32 workers (Fig. 2). At these sizes, then, exact evaluation takes elapsed time of the same order as the sampling runs and considerably more processor time.

The costs of exact evaluation and of sampling scale in different variables: the exact cost is

$$t_{\text{exact}} \;\propto\; N\left[(N-m+1)(m+1)^5 + c_1(m+1)^6 + c_2(N-m+1)^4\right], \tag{14}$$

with constants c_1, c_2 of order one and lower-order terms given in Supplement S3. The first term is the contraction of the five-index array of side $m+1$ against the three-index array of side $N-m+1$, the second the construction of the five-index array, the third that of the three-index array. The overall factor N is the number of primes, which grows linearly with N through the denominator bound (in integer arithmetic it is the length of the integers instead). At the mixed-site counts of the tRNA quartets ($m = 25$ to 59) the first two terms dominate and the cost grows as about the fifth power of m; below $m \approx 10$ at tRNA length the third term dominates and the cost is nearly independent of m (Supplement S3). At $m = 25$ the five-index array holds $26^5 \approx 1.2 \times 10^7$ entries per prime. A sampler's cost per generation, in contrast, is nearly independent of N once site patterns are compressed (a quartet has at most 256 distinct patterns), and so, very nearly, is its Monte Carlo error at a fixed number of generations. Away from the prior end of the path the variance of the log-likelihood under the power posterior at β is close to $d/(2\beta^2)$ for a tree with d free branch lengths, whatever the alignment length. Only as $\beta \to 0$, where the power posterior approaches the prior, does it grow with N, and there as N^2 (Supplement S4.3 verifies both on alignments of 68 to 276 sites). The stepping-stone error at a fixed budget therefore depends on N only through the few stones nearest the prior: with the counts of the tRNA-Phe quartet multiplied by two and by four, the standard deviation of the MrBayes estimate across independent runs stayed between 0.006 and 0.007 log units at 50 stones of 400,000 generations, and its bias within error of zero (Supplement S4.3). The error falls as the inverse square root of the budget (Section 4.3). Thermodynamic integration adds a discretization bias that falls as K^{-2} in the number K of β values and does not decrease with more sampling at each value (Lartillot and Philippe, 2006; Xie et al., 2011). With N the bias grows less than proportionally (by a factor of about 1.8 per doubling of the tRNA-Phe counts at $K = 50$), through the stones nearest the prior, the contribution of the rest of the path being independent of N (Supplement S4.3). The tRNA-Phe and tRNA-Gln quartets, of nearly equal cost to a sampler, thus differ about seventyfold in exact cost by the fifth-power factor in Eq. (14). The wall times reported above for these two quartets differ by less than this, because job start-up dominates the per-prime times of the smaller computation.

Under the fifth-power scaling in the mixed-site count, the two 900-site alignments would cost more than 10^4 times the tRNA-Gln computation, and no exact value at that length exists. For those two alignments the exact cost would be dominated by the array product, which is batched integer matrix multiplication modulo a prime, a workload suited to graphics processors. Thus nearly all exact values on real data in this paper are Jukes–Cantor marginal likelihoods of quartets with at most 59 mixed sites (up to 390 sites in all) and of one quintet. The only others are Kimura-type values, with the branch-wise rates and prior described above, of six windows of at most 30 sites. Under the fixed-rate Kimura and GTR models we have computed exact values only for the small synthetic alignments already mentioned. All remaining values are floating-point quadrature or sampling estimates, marked as such in the tables (the two 900-site alignments, for example).

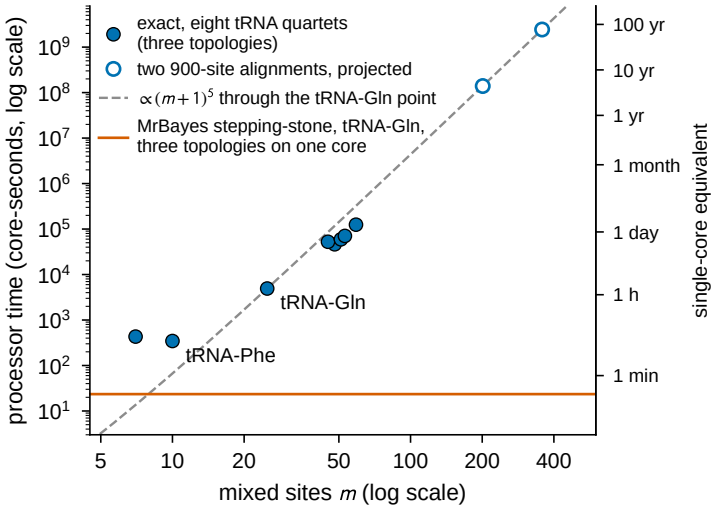

Figure 2: Exact evaluation is feasible at the mixed-site counts of the tRNA quartets and computationally prohibitive at several hundred mixed sites. Filled circles: processor time for the marginal likelihoods of all three topologies, each computed twice, of the eight timed tRNA quartets of Table 2 ($N = 63$–75) against the mixed-site count m; right axis, the same time on one core. Dashed line: $\propto (m+1)^5$ through the tRNA-Gln point; the two smallest cases sit above it because job start-up dominates their sub-second per-prime times. Open circles: the two 900-site alignments of Table 3 ($m = 201$, 357) placed on the line, not computed. Horizontal line: MrBayes stepping-stone, tRNA-Gln quartet, three topologies on one core (50 stones of 50,000 generations, two runs each); sampler cost is nearly independent of m.

3.4 Limits of exactness

Table 1 lists the number class of the evidence (rational, algebraic, or a constant defined by an exponential integral or by a linear differential equation) under each modeling ingredient. For two ingredients in routine use, rate variation across sites and tempered likelihoods, the term-by-term argument no longer applies in general. The evidence is then an exponential-integral constant under rate variation and, under tempering, a constant determined by a linear differential equation; both are expected to be irrational but can still be computed as precisely as required.

Rate variation across sites. Under the gamma model of among-site rate variation (Yang, 1993), with the rate integrated out continuously, each site carries a rate r with a mean-one gamma distribution of shape k. Integrating r out of a monomial $\prod_{e\in S} x_e^r = e^{-r\sum_{e\in S}\mu t_e}$ gives $(1 + \frac{1}{k}\sum_{e\in S}\mu t_e)^{-k}$, which couples the branches; the integral over branch lengths no longer factorizes edge by edge, and the term-by-term argument fails. For example, on the three-taxon star tree with a single constant site, shape $k = 1$ and the exponential prior used throughout ($\alpha = 1$),

$$Z_{+\Gamma} \;=\; \frac{10 - 6G}{64}, \qquad G = \int_0^\infty \frac{e^{-s}}{1+s}\,\mathrm{d}s = 0.5963\ldots, \tag{15}$$

where G is the Euler–Gompertz constant. With two or more sites a second constant of the same family enters, for which no expression as a rational combination of G and familiar constants is known, and the number of such constants appears to grow with the alignment length

(Supplement S1). We have identified these constants only for one and two sites. Under the discrete approximation with a fixed number of rate categories and a fixed shape (Yang, 1994), each pattern probability is instead the average, over the category rates r_c, of the same polynomial evaluated at the powers $x_e^{r_c}$. It remains multilinear in these powers, and the integrand remains a polynomial in them. The term-by-term evaluation then survives as a finite sum in closed form, which is rational only if the category rates are rational, and the mean or median category rates of the discrete gamma are not. The lattice of exponents generated by the r_c grows quickly with the number of mixed sites, and we expect this evaluation to be impractical beyond a few of them; a shape parameter integrated over its prior is not covered. Because most analyses of real data include gamma-distributed rates across sites (JC + Γ or GTR + Γ), rate variation remains a practical limitation of the method as it stands. Under the continuous model the term-by-term argument fails, and under the discrete model the value has a closed form that we have not attempted to evaluate on real alignments.

Tempered likelihoods. A stepping-stone or thermodynamic-integration analysis samples from densities proportional to L^β times the prior at $0 < \beta < 1$, with $L = \prod_k p_k^{y_k}$ the likelihood, and estimates ratios of the normalizing constants Z_β of Eq. (3) (Xie et al., 2011). When every product βy_k is an integer, Z_β is again rational; otherwise the integrand contains fractional powers of polynomials, to which the term-by-term argument does not apply. At fixed β, however, Z_β can be reached by introducing an auxiliary parameter that interpolates between an integrand with a known rational integral and the tempered one. As a function of that parameter the integral satisfies a linear differential equation with polynomial coefficients, which can be found by computer algebra from the leading terms of the integral's power series in that parameter. Solved from the exact initial values, the equation then gives Z_β to any stated precision. Consequently, on a small tree, and for schedules on which every βy_k has a small denominator, each ratio that a stepping-stone analysis estimates can be computed without sampling and used to test the estimators.

We carried out this computation for one small case (Supplement S4), the three-taxon star tree under Jukes–Cantor with counts $\mathbf{y}'$ given by xxx: 4, xxy: 2, xyz: 2 ($N = 8$), at $\alpha = 1$, on the uniform schedule $\beta_k = k/4$, $K = 4$. At $\beta = \frac{1}{2}$ and $\beta = 1$ the exponents $\beta y'_k$ are integers and

$$Z_{1/2} = 64^{-4}\,\frac{17872}{3375}, \qquad Z_1 = 64^{-8}\,\frac{50612054}{1157625}, \tag{16}$$

while $Z_0 = 1$ (the powers of 64 are the factor $4^{-nN\beta}$ of the likelihood at $n = 3$). For $\beta = \frac{1}{4}$ and $\frac{3}{4}$ the exponents are half-integers; there the differential equations, of orders 10 and 11, were found and solved, giving $64^2 Z_{1/4} = 2.1466466\ldots$ and $64^6 Z_{3/4} = 14.592196\ldots$, in agreement with numerical quadrature. The four ratios $Z_{\beta_{k+1}}/Z_{\beta_k}$ of this schedule are therefore free of sampling error and serve as the reference for the stepping-stone comparison of Section 4.3. The $\beta = \frac{3}{4}$ value took 27 minutes on one workstation, mostly in finding its differential equation; the rational values take seconds. The demonstration is limited to three taxa, eight sites and half-integer exponents. For exponents with denominator three or more the differential equation could not be found at the series depths tried, and the Beta-quantile schedules used in practice (Xie et al., 2011) are therefore outside the scope of this method at present.

Table 1: Number class of the marginal likelihood by modeling ingredient, with the cases evaluated in this paper and the method used for each. The constants in the fourth and fifth rows are computable to any stated precision from their defining expressions; their transcendence is expected but unproved (the irrationality of G itself is an open problem). Priors are independent across branches throughout; β is the exponent of the likelihood in a power posterior.

substitution model	branch prior	β	number class	cases evaluated here	method
JC69, K2P, K3P (group-based); rates fixed and rational, or a branch-wise prior with rational moments	exponential	1	rational	JC69: every quartet and quintet value in this paper ($m \leq 59$, $N \leq 390$); K2P with branch-wise rates: six windows, $N \leq 30$; K3P: none	term by term, Lemma 1
GTR, fixed rational rates and frequencies	exponential	1	rational (a rational function of rates and frequencies)	synthetic only: 3 taxa, $N = 3$; 5 taxa, $N = 4$	matrix inverse, Eq. (10)
group-based	gamma, shape k	1	rational for integer k; algebraic for rational k	synthetic: 3-taxon star, $k = 1/2$	term by term, Eq. (11)
JC69 + Γ (continuous gamma, shape k)	exponential	1	exponential-integral constant, e.g. $(10 - 6G)/64$	3 taxa, one constant site, $k = 1$; constants identified for one and two sites only	closed form, Eq. (15)
any of the first three rows	exponential	$\notin \mathbb{Z}$	rational whenever every βy_k is an integer (a sufficient condition); otherwise outside Lemma 1, the value of a solution of a Fuchsian differential equation, presumably irrational	3 taxa, $N = 8$, $\beta = \frac{1}{4}, \frac{3}{4}$	linear differential equation solved from exact initial values
GTR + Γ; substitution parameters (κ, exchangeabilities, frequencies) integrated over continuous priors			not treated (an integral of a rational function of the parameters; in general not rational)		

4 Exact values on real alignments

The four-taxon values in this section are computed under the model and priors of Section 2.2: JC69, a uniform prior over the three topologies, and independent exponential branch-length priors of rate 4/3 ($\alpha = 1$). Decimals are rounded to the last digit shown; for the benchmark alignments Table B1 gives the exact values to ten decimals and Supplement S10 their numerators and denominators.

4.1 Benchmark alignments

Table 2 lists twelve quartets of transfer-RNA genes. Three are single mitochondrial tRNA genes (tRNA-Phe, tRNA-Lys and tRNA-Leu(UUR)) from the reference mitochondrial genomes of human, chimpanzee, gorilla and orangutan. Each gene was taken at the coordinates of its annotated tRNA feature in the RefSeq record. The four sequences were aligned pairwise to a reference among them (the longest, or human when the lengths are equal) by the Needleman–Wunsch algorithm (match 1, mismatch -1, gap -2), and the pairwise alignments were merged on that reference. The other nine consist of rows of the Rfam RF00005 seed alignment (954 sequences by 118 columns) (Kalvari et al., 2021). One of the nine, our reference quartet, comprises the mitochondrial tRNA-Gln genes of mouse, rat, chicken and *Xenopus*; for the other eight we took, for each organism, the row whose seed identifier sorts first alphabetically.

Columns with a gap or ambiguity code in any sequence were removed, leaving $N = 63$ to 75 sites. Of these, $m = 7$ to 59 are mixed sites, that is, neither constant nor of the class $xxyy$ that supports the split 12|34. Each of the thirty-six marginal likelihoods is a ratio of integers of 131 to 331 digits.

The reference quartet prefers the pairing of mouse with rat by 14.5368 log units, and the two other orthologous Rfam rows, primates and vertebrates, recover the accepted groupings by 7.2575 and 4.8380 log units. In the remaining six Rfam rows, however, our alphabetical rule selected tRNAs that are not orthologous but different isoacceptors. In two of them the genes also come from different compartments: nuclear against chloroplast in the plant row, and human nuclear against yeast mitochondrial in the cross-domain row. In these six rows the preferred split reflects the similarity of the aligned sequences and is not evidence about the relationships of the four organisms. We retain them because a paralogous or contaminated locus would enter a multilocus analysis with a margin of this kind, indistinguishable from the margin of an orthologous locus.

Table 3 gives the three tRNA-Gln marginal likelihoods, together with those of four longer published alignments and of the twenty-site synthetic alignment of Eq. (12). The lysozyme quartet of Messier and Stewart (1997) (human, gibbon, colobus, marmoset; $N = 390$, $m = 39$), distributed with PAML, was evaluated exactly, and the hominoid pairing is preferred by 14.5647 and 16.3561 log units over the two alternatives. A 5S ribosomal RNA quartet from the Rfam RF00001 seed (mouse, chicken, *Xenopus*, *Drosophila*; $N = 116$), also evaluated exactly, has no site of the class supporting the amniote split. The accepted topology is disfavored by 1.3266 log units relative to the best alternative, a margin comparable to those by which single tRNA genes prefer wrong pairings (next subsection). The mitochondrial alignments of Brown et al. (1982) and Hayasaka et al. (1988), distributed with PAML and MrBayes, have $m = 201$ and 357, beyond the reach of the exact computation at present; for these two alignments the table gives floating-point quadrature values instead (the rows marked †).

locus	species, numbered 1 to 4	N	m	preferred split	log BF	orthologous
mt tRNA-Phe	human, chimpanzee, gorilla, orangutan	69	10	13|24; 12|34 = 14|23 exactly	0.0013	yes
mt tRNA-Lys	human, chimpanzee, gorilla, orangutan	70	11	12|34	0.9105	yes
mt tRNA-Leu(UUR)	human, chimpanzee, gorilla, orangutan	75	7	13|24	2.8180	yes
mt tRNA-Gln	mouse, rat, chicken, *Xenopus*	68	25	12|34	14.5368	yes
tRNA, primates	human, lemur, macaque, squirrel monkey	67	48	13|24	7.2575	yes
tRNA, vertebrates	chicken, mouse, rat, *Xenopus*	66	51	14|23	4.8380	yes
tRNA, bacteria	*B. subtilis*, *E. coli*, *M. pneumoniae*, *T. thermophilus*	70	45	12|34	1.1070	no (different isoacceptors)
tRNA, fungi	*A. nidulans*, *N. crassa*, *S. cerevisiae*, *S. pombe*	71	53	13|24	2.2423	no (different isoacceptors)
tRNA, insects	mosquito, silkmoth, fruit fly, locust	63	59	14|23	0.0557	no (different isoacceptors)
tRNA, plants	*A. thaliana*, *N. tabacum*, *T. aestivum*, *Z. mays*	70	50	13|24	6.4205	no (nuclear, chloroplast)
tRNA, cross-domain	*E. coli*, human, *M. vannielii*, *S. cerevisiae*	70	44	14|23	4.9463	no (nuclear, mitochondrial)
tRNA, archaea	*Haloarcula*, *Haloferax*, *M. vannielii*, *T. acidophilum*	72	48	13|24	3.0225	no (different isoacceptors)

Accessions: hominids RefSeq NC_012920, NC_001643, NC_001645, NC_001646; tRNA-Gln GenBank V00711, X14848, X52392, M10217; other rows, Rfam RF00005 seed identifiers as listed in Supplement S10.

Table 2: Orthologous non-hominid quartets recover accepted groupings by 4.8 to 14.5 log units; the three hominid genes disagree with one another. Exact comparison of the three topologies for twelve tRNA quartets under JC69, exponential branch-length priors of rate 4/3 (mean 0.75; $\alpha = 1$), uniform topology prior. Split 12|34 joins species 1 with 2; log BF, exact log Bayes factor (log units) of the preferred split over the closer alternative; N, gap-free sites; m, mixed sites when 12|34 is evaluated. "Orthologous": whether the four sequences are the same gene; if not, the preferred split is not evidence about the organisms' relationships.

(a) $\log Z$ to ten significant digits; rows not marked † are exact

alignment	N	m	topology	$\log Z$	$\log \mathrm{BF}$
mt tRNA-Gln (mouse, rat, chicken, *Xenopus*)	68	25	12|34	−261.4030642	
			13|24	−275.9398267	14.5367625
			14|23	−276.9593357	15.5562715
lysozyme (Messier and Stewart, 1997) (human, gibbon, colobus, marmoset)	390	39	12|34	−829.7542320	
			13|24	−844.3189102	14.5646782
			14|23	−846.1103094	16.3560775
5S rRNA, Rfam RF00001 (mouse, chicken, *Xenopus*, *Drosophila*)	116	38	12|34	−345.7452740	
			13|24	−344.7711532	−0.9741208
			14|23	−344.4186439	−1.3266301
synthetic, Eq. (12)	20	6	12|34	−87.64924335	
			13|24 = 14|23	−90.83748304	3.18823969
mtDNA 895 bp (Brown et al., 1982) (human, chimpanzee, gorilla, orangutan)†	895	201	12|34	−2427.531701	
			13|24	−2439.667267	12.135566
			14|23	−2433.111176	5.579475
mtDNA 898 bp (Hayasaka et al., 1988) (*Homo*, *Macaca*, *Saimiri*, *Lemur*)†	890	357	12|34	−3365.932479	
			13|24	−3385.261108	19.328630
			14|23	−3397.798577	31.866098

(b) Estimates on the tRNA-Gln quartet, mean ± one standard deviation of four independent estimates

	exact	MrBayes SS	RevBayes SS	LoRaD
$\log Z(12\|34)$	−261.4031	-261.4239 ± 0.0266	-261.3815 ± 0.0388	-261.3966 ± 0.0074
$\log Z(13\|24)$	−275.9398	-275.9632 ± 0.0291	-275.9588 ± 0.0145	-275.9321 ± 0.0121
$\log Z(14\|23)$	−276.9593	-276.9884 ± 0.0154	-276.9502 ± 0.0235	-276.9588 ± 0.0017
$\log \mathrm{BF}$, 12|34 over 13|24	14.5368	14.5393	14.5773	14.5355
$\log \mathrm{BF}$, 12|34 over 14|23	15.5563	15.5646	15.5687	15.5621

Table 3: Marginal likelihoods of four-taxon alignments under JC69 (equal base frequencies, no rate variation); five branch lengths independently exponential with rate 4/3 in the MrBayes and RevBayes parameterization (mean 0.75 substitutions per site); uniform topology prior. (a) $\log Z$ to ten significant digits (ten decimals in Table B1); $\log \mathrm{BF} = \log Z(12|34) - \log Z(T)$, negative when the accepted split 12|34 is disfavored. †Floating-point Gauss–Legendre quadrature, digits stable under grid refinement; these two alignments exceed the present reach of the exact computation. (b) Stepping-stone (SS) estimates from MrBayes 3.2.7a and RevBayes (Höhna et al., 2016) at 50 stones, and LoRaD estimates computed with the R package lorad (Milkey et al., 2023) from MrBayes posterior samples, with the settings of Section 4.3.

4.2 Two exact ties

The three hominid rows of Table 2 are three genes from the same four non-recombining mitochondrial genomes, and they give three different results. Whereas tRNA-Lys prefers the accepted pairing of human with chimpanzee by 0.9105 log units, tRNA-Leu(UUR) prefers human with gorilla by 2.8180 log units, at odds with the established phylogeny. For tRNA-Phe ($N = 69$), with the species numbered as in the table,

$$Z(12|34) \;=\; Z(14|23) \qquad \text{as identical rational numbers.} \tag{17}$$

Both marginal likelihoods are the same ratio of a 217-digit to a 287-digit integer. Of the 69 columns, 59 are constant and 10 carry a change private to one species: five in orangutan, three in chimpanzee, and one each in human and gorilla. Since no column groups two of the apes against the other two, exchanging the labels of human and gorilla leaves every site-pattern count unchanged, and the same exchange carries the topology 12|34 into 14|23. Hence under the Jukes–Cantor model, with or without rate variation across sites, the two marginal likelihoods are one and the same integral whenever the branch-length priors are exchangeable. Under Kimura's model the tie would break, because the change private to human is a transversion and the one private to gorilla a transition. This symmetry was not used in the computation: the two values were evaluated separately, and the resulting numerators and denominators are identical.

The third topology, 13|24 (human with gorilla), is mapped to itself when the labels of human and gorilla are exchanged, and the symmetry therefore ties it to neither of the other two. Its evidence is the largest of the three, by 0.0013 log units, although no column supports that grouping (Fig. 3). Human and gorilla are the two sequences with the fewest private changes. The margin in favor of grouping them is not a difference in fit: the three likelihood functions attain the same maximum, on the star tree, where the internal branch has length zero and the three functions coincide (Supplement S12). They differ only at positive internal lengths, and there in opposite ways according to how the pendant branches are treated: with the pendant lengths at their best values the (human, gorilla) topology fits slightly worse than the other two, but with the pendant lengths integrated against their prior it is favored, by an amount that grows with the internal length (0.001 log units at 0.02, 0.03 at 0.1 substitutions per site). The evidence margin is this difference averaged over the posterior of the internal length, nine tenths of it from lengths below 0.09, and is therefore an interaction between the likelihood and the branch-length prior rather than an effect of either alone. Recomputed exactly at prior means of 0.1 and 0.01 substitutions per site (exponential rates 10 and 100), it keeps its sign and falls to 0.0010 and 0.0002 log units (Table B1), less than in proportion to the mean because the posterior of the internal length is set by the likelihood until the prior becomes the narrower of the two. Under gamma priors of shape two the margin rises to 0.02, and a prior that lengthens only the chimpanzee branch reverses it (Supplement S12). The tie of Eq. (17) persists under every prior exchangeable across branches, as the symmetry requires. Stepping-stone estimates on this alignment (Section 5.3) vary between runs by several hundredths of a log unit (Fig. 4b), and can neither confirm the tie nor distinguish a difference of a thousandth of a log unit from zero.

A second exact tie occurs in the alignment from which Krings et al. (1997) reported the first Neandertal DNA sequence, 379 bp of hypervariable region I of the mitochondrial control region. We took the alignment as printed in their Figure 2, which shows the human reference and 49 clones over positions 16,210 to 16,270 of that region plus one insertion column. Dropping the insertion column, which JC69 cannot model, leaves 61 consecutive sites. The quartet comprised

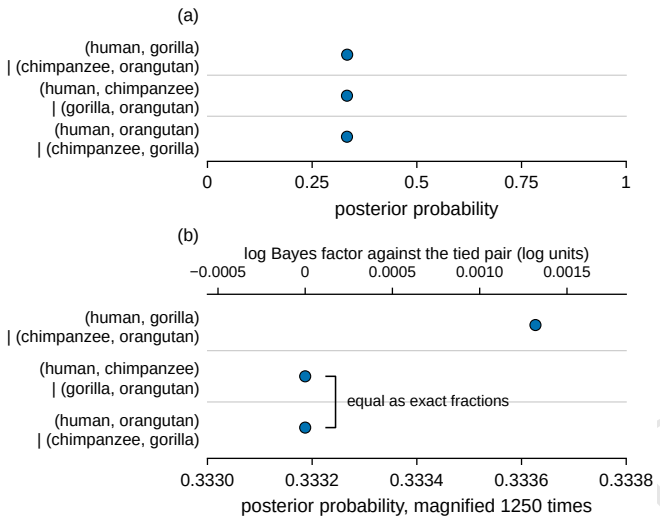

Figure 3: Two of the three posterior topology probabilities for the hominid tRNA-Phe quartet are equal as exact fractions, and the third exceeds them by less than the stepping-stone runs on this alignment can resolve. (a) Posterior probabilities under the model and priors of Table 2. (b) The same, magnified 1250 times; upper axis, log Bayes factor against the tied pair. The (human, gorilla) grouping, which no column supports, leads by 0.0013 log units. The three likelihood functions have the same maximum, on the star tree; the margin arises in the integration over branch lengths and shrinks when the prior mean is lowered (Section 4.2, Supplement S12). Stepping-stone estimates vary between runs by 0.017 to 0.067 log units (Fig. 4b).

the human reference, a modern-human contaminant clone identified by the authors, the Neandertal sequence and an independent Neandertal amplification clone. The split placing the two humans together is preferred by 25.1717 log units over each alternative. The two alternatives are tied exactly, by the same label-exchange symmetry as in the tRNA-Phe quartet: the two Neandertal sequences are identical over these 61 sites. A change at a single site shifts this margin appreciably: introducing an artificial substitution into the human clone (A to G at position 16,230) lowers it to 20.3435 log units. The exact comparison thus recovers the 1997 separation of Neandertal from modern human sequences as a four-taxon Bayes factor under one model and prior (full analysis in Supplement S6). Alignments with such label symmetries are perhaps the simplest on which no method can separate two topologies. A second such case is an internal branch of length zero, at which the three topologies prescribe the same pattern distribution; Supplement S5 locates that point, and every other degeneration of the quartet integrand in the internal-edge variable, as an explicit algebraic variety.

4.3 Estimators against the exact values

The exact values permit a direct test of marginal-likelihood estimators (Lartillot and Philippe, 2006; Xie et al., 2011; Fourment et al., 2020; Wang et al., 2023) on a phylogenetic quantity. We carried out two such tests, and in each of them the sampler settings were chosen beforehand and not tuned against the exact values. Below, ± denotes one standard deviation across independent runs unless a standard error is stated.

The first test uses the three-taxon, eight-site alignment and the four equally spaced stones

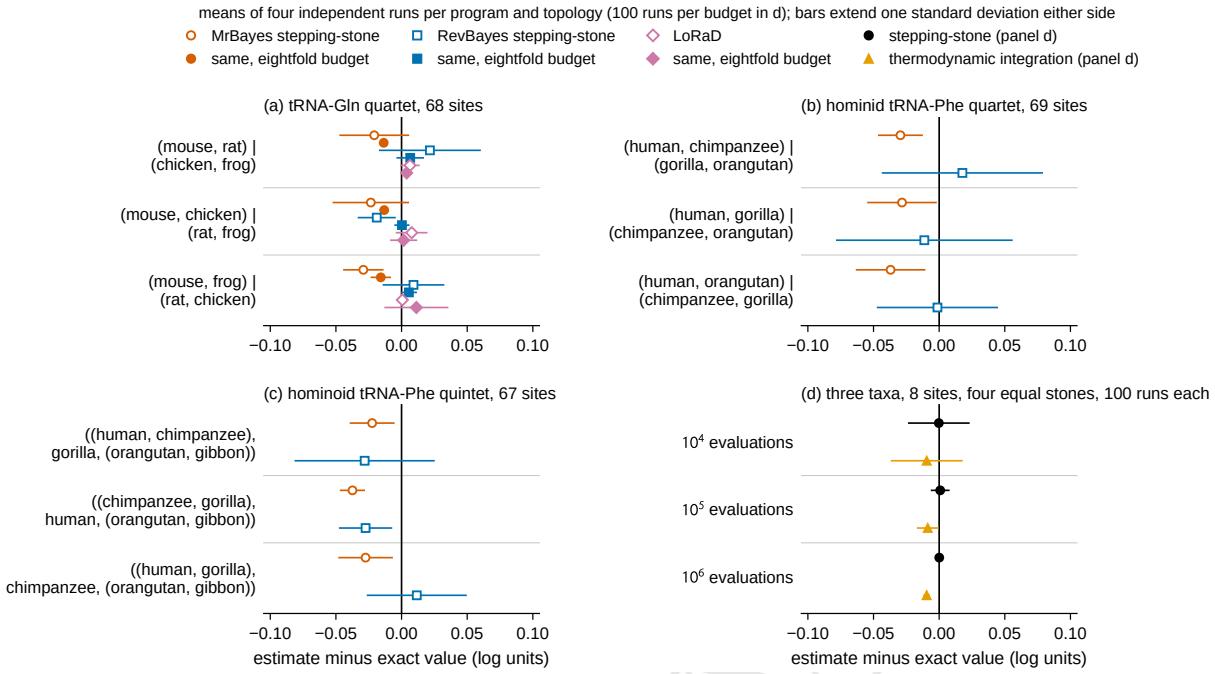

Figure 4: Deviations of MrBayes stepping-stone estimates from the exact values are slightly negative throughout, those of RevBayes and LoRaD within error of zero, and on a uniform four-interval schedule thermodynamic integration carries a discretization bias that stepping-stone sampling on the same schedule does not. Symbols, mean over runs minus the exact $\log Z$; bars, one standard deviation across runs (four runs in a–c, 100 in d). (a) tRNA-Gln quartet at two budgets matched within 2% between programs (LoRaD budgets were 0.78 of these). (b) Hominid tRNA-Phe quartet and (c) three leading topologies of the hominoid tRNA-Phe quintet, both at 50 stones with the settings of Section 5. (d) The three-taxon, eight-site case of Section 3.4 ($K = 4$).

of Section 3.4, a schedule on which each stepping-stone ratio is available as an exact value. Stepping-stone and thermodynamic-integration estimates were formed from the same power-posterior samples in $R = 100$ replicates, at total budgets of 10^4, 10^5 and 10^6 likelihood evaluations divided equally among the four stones. At $\beta = 0$ the draws came directly from the prior, and at $\beta > 0$ from a random-walk Metropolis chain with Gaussian proposals reflected into the unit cube. The first tenth of each chain, during which the step size was tuned, was discarded. For thermodynamic integration we applied the trapezoidal rule to $\mathrm{E}_\beta[\log L]$ at the five β values, with the expectation at $\beta = 1$ obtained by reweighting the $\beta = 3/4$ sample.

The stepping-stone mean lies on the exact $\log Z = -29.4932$ within Monte Carlo error at each budget (0.4 standard errors at the largest), and its spread falls as the inverse square root of the budget (Fig. 4d). Thermodynamic integration over the same four intervals, however, is low by 0.0095 log units, which at the largest budget is 39 standard errors. The size of this shortfall is set by the coarse uniform schedule: the trapezoidal rule misses the curvature of $\mathrm{E}_\beta[\log L]$ between so few stones, and sampling longer at each stone does not reduce the error (Lartillot and Philippe, 2006; Xie et al., 2011). Xie et al. (2011) found the same pattern (stepping-stone sampling accurate, thermodynamic integration slightly low) on their conjugate normal example,

whose true value is known in closed form; here the reference value at each stone is the normalizing constant of a phylogenetic power posterior, computed without sampling.

In the second test, stepping-stone sampling in MrBayes and RevBayes and the LoRaD estimator were applied to the tRNA-Gln quartet, with the topology fixed to each of the three trees in turn. In MrBayes 3.2.7a (Ronquist et al., 2012) stepping-stone sampling used $K = 50$ stones at the program's default spacing (evenly spaced quantiles of a Beta$(0.4, 1)$ distribution; Xie et al., 2011) and one unheated chain per run. After an initial burn-in of one stone's length, the chain ran for 50,000 generations at each stone, sampled every 500, with the first quarter discarded as burn-in. Two replicate analyses, each of two independent runs, gave four estimates per topology, one from each run. Eighteen further analyses varied the two settings that control stepping-stone error: chains eight times longer at 50 stones, and 100 and 200 stones at the baseline length per stone.

In RevBayes (Höhna et al., 2016) the same model and prior were specified and stepping-stone sampling was run with 50 stones, spacing parameter 0.4, a burn-in of 5,000 and then 10,000 generations per stone, in four replicate analyses per topology. Path-sampling estimates came from the same samples, and the analyses were repeated at 100 and 200 stones and at eightfold length. Measured in MCMC proposals, the computational budgets of the two programs agree within 2% at each setting.

LoRaD (Wang et al., 2023) (R package lorad 0.0.1.0, Milkey et al., 2023) was applied to fixed-topology posterior samples from MrBayes (2 million generations sampled every 100, 25% burn-in, four chains per topology). The coverage fraction (the share of the sample enclosed by the region of highest posterior density to which the estimator is restricted) was 0.5, at which the spread across chains was smallest; the training sample, used to determine that region, was taken from the start of each chain. Because the released package reports no Monte Carlo standard error, that spread serves in its place.

At the baseline settings (Table 3b and Fig. 4a) the three MrBayes means lie 0.02 to 0.03 log units below the exact values, and eleven of the twelve individual estimates also fall below them. Because this offset is common to the three topologies, it cancels in differences, and the two stepping-stone log Bayes factors are 14.5393 and 15.5646 against exact values of 14.5368 and 15.5563.

In the eighteen further analyses roughly 0.010 log units of the offset at 50 stones is removed either by more stones or by longer chains, as expected of the finite-sample bias of the stepping-stone estimator (Xie et al., 2011). The remainder, -0.016 log units with a standard error of 0.002 over the 36 estimates at the three modified settings, is unchanged by either and is the same for the three topologies within error. The residual offset is unchanged by reversing the path, doubling either burn-in or computing the likelihood in double precision, and at the sampled branch lengths the log-likelihood and log-prior that MrBayes reports agree with ours to 10^{-5} and 10^{-9} (Supplement S4.4). It disappears when one of the three branch-length proposals MrBayes uses by default, the node slider, is disabled: -0.0002 against -0.0268 log units (standard errors 0.0013 and 0.0014; 70 estimates per setting), and -0.003 against -0.026 (0.002 and 0.003; 72 per setting) on eight of the exact problems of this and the companion paper. The proposal ratio coded for this move from version 3.2.4 onward contains one more factor of the ratio of new to old length sum than the value that leaves the branch-length prior invariant, which is the value versions 3.1.2 to 3.2.3 used; run without data and with the node slider as its only proposal, MrBayes 3.2.7a samples trees 1.4 to 3 times longer than the prior mean, depending on the prior (Supplement S4.4). In the default mixture of proposals the posterior tree length

changes by about 0.1 per cent on these alignments, too little to affect inferred trees. An offset of the same sign appears in the MrBayes estimates on both tRNA-Phe alignments of Section 5.3 (Fig. 4b,c). This residual offset is far below any threshold used to interpret a Bayes factor (Kass and Raftery, 1995) and cancels in the Bayes factors of these analyses, which all lie within 0.015 log units of the exact ones. It also cannot be detected from any quantity that MrBayes itself reports, including the agreement between independent runs.

RevBayes at matched settings gives deviations similar in magnitude to those of MrBayes but of both signs, with seven of the twelve estimates below the exact values. Its mean stepping-stone deviations are $+0.022$, -0.019 and $+0.009$ log units, and its path-sampling deviations $+0.012$, -0.025 and $+0.001$. Hence, with no common term to cancel, its baseline stepping-stone Bayes factors differ from the exact values by 0.041 and 0.012 log units. As the budget grows the RevBayes deviations shrink toward zero, with no common offset at the resolution of the largest runs (0.006 log units), and with Bayes factors within 0.01 log units of the exact ones. The MrBayes offset is unchanged at these larger budgets, and only the scatter of its estimates across runs decreases.

The LoRaD estimates, which use only the posterior sample, agree with the exact values within their spread across chains. Their baseline mean deviations are $+0.006$, $+0.008$ and $+0.001$ log units, with chain-to-chain standard deviations of roughly the same size, and they give both Bayes factors within 0.006 log units of the exact ones. The LoRaD estimate becomes unstable, however, at extreme values of the coverage fraction.

On the tRNA-Gln quartet MrBayes and RevBayes agree to a few hundredths of a log unit, and each passes its convergence diagnostics across replicate runs. A much longer MrBayes run, the usual reference standard (Fourment et al., 2020), nevertheless converges to a value with the same offset as the shorter runs and, used as the reference, would make RevBayes and LoRaD appear biased by that amount. These deviations are far too small to change an inferred tree or a model choice, but against the exact values the systematic error of each estimator can be distinguished from its Monte Carlo error.

5 Five taxa and the hominoid quintet

A five-taxon tree is the smallest tree containing more than one quartet, and hence the first case in which the ranking of topologies by exact evidence can be compared with the tree that quartet amalgamation would assemble from its five quartets. Here, we add gibbon to the hominid tRNA-Phe alignment of Section 4.2 and compute the exact marginal likelihoods of all fifteen topologies under the same model and prior. The resulting ranking is compared with the splits that the five four-taxon subsets of the alignment support under exact evaluation, and stepping-stone estimates for the leading topologies are then tested against the exact values.

5.1 The five-taxon integral

Under a time-reversible model without a clock the likelihood does not depend on the position of the root (Felsenstein, 1981), and the evidence is a function of the unrooted topology. Five labeled taxa admit 15 unrooted topologies with seven edges each (two of them internal). Under nucleotide relabeling the 4^5 leaf patterns fall into 51 classes, the set partitions of five leaves into at most four blocks. By pruning over the three internal nodes each pattern probability p_k is written as a multilinear polynomial in the seven edge variables, $1024\,p_k$ having integer

coefficients. Writing $\mathbf{y} = (y_1, \ldots, y_{51})$ for the pattern counts on topology T, and expressing the exponential branch-length prior in the edge variables as at four taxa ($\alpha = \lambda/\mu$), the evidence is

$$Z(T; \mathbf{y}, \alpha) \;=\; \alpha^7 \int_{[0,1]^7} \prod_{k=1}^{51} p_k(x)^{y_k} \prod_{e=1}^{7} x_e^{\alpha-1} \, \mathrm{d}x. \tag{18}$$

Lemma 1 carries over unchanged, with denominator bound $1024^N \prod_{e=1}^{7} \prod_{j=0}^{N} (jq+p)$ for the bare integral $\alpha^{-7} Z$ of an N-site alignment at rational $\alpha = p/q$. The direct expansion has up to $(N+1)^7$ monomials rather than $(N+1)^5$, but constant columns again factor out, and the work is governed mainly by the number of variable columns, which at five taxa is the mixed-site count m.

5.2 The tRNA-Phe quintet

We formed the five-taxon alignment by adding the tRNA-Phe gene of the gibbon mitochondrial genome (GenBank NC_002082) to the four hominid tRNA-Phe sequences, which gives $N = 67$ gap-free columns, 54 of them constant and the other $m = 13$ mixed. The model and priors are those of the quartets (Section 2.2, $\alpha = 1$), with a uniform prior on the fifteen topologies.

Several pairs of topologies have equal evidence because of a symmetry of the pattern counts: human and gorilla differ at two columns only, at one of which gorilla shares its state with gibbon and at the other human does. Exchanging the human and gorilla rows therefore swaps the patterns of those two columns and leaves the 51 pattern counts unchanged. Hence two topologies related by that exchange have identical evidence under JC69, whatever the branch-length prior, provided it is exchangeable across branches. Three of the fifteen topologies contain the pair (human, gorilla) and are mapped to themselves, whereas the other twelve form six exchanged pairs, leaving at most nine distinct values. Computed separately, the two marginal likelihoods of each of the six pairs are identical as ratios of integers.

The three topologies that separate orangutan and gibbon from the African apes have the three largest of the fifteen values (Table 4); the other twelve topologies lie at least 1.72 log units below the maximum. The maximum is attained by an exchanged pair,

$$\begin{gathered} ((\text{human}, \text{chimpanzee}), \text{gorilla}, (\text{orangutan}, \text{gibbon})) \\ \text{and} \quad ((\text{chimpanzee}, \text{gorilla}), \text{human}, (\text{orangutan}, \text{gibbon})), \end{gathered}$$

with $\log Z = -189.6451$ for both: the accepted hominoid tree and its exchanged image, which only a nonuniform topology prior could separate. The third resolution of the trichotomy among human, chimpanzee and gorilla, $((\text{human}, \text{gorilla}), \text{chimpanzee}, (\text{orangutan}, \text{gibbon}))$, lies 0.5625 log units below the tied pair (Fig. 5). That margin compares single five-taxon trees. Summed instead over the five placements of gibbon, that is, over the five topologies that contain each split of the four great apes, the evidence favors (H, C) | (G, O) and (H, O) | (C, G), which remain exactly tied, over (H, G) | (C, O) by 0.3518 log units; the posterior probabilities of the three splits are 0.370, 0.370 and 0.260 (Supplement S12). This is the five-taxon counterpart of the quartet's margin of 0.0013 log units in the opposite direction.

The (human, gorilla) resolution is nominally preferred at four taxa, where no column distinguishes the three resolutions. In the five-taxon alignment, however, the two columns at which human and gorilla differ become informative, and both now count against joining human with gorilla. The reversal is due to gibbon's sequence and not to the two constant columns by which

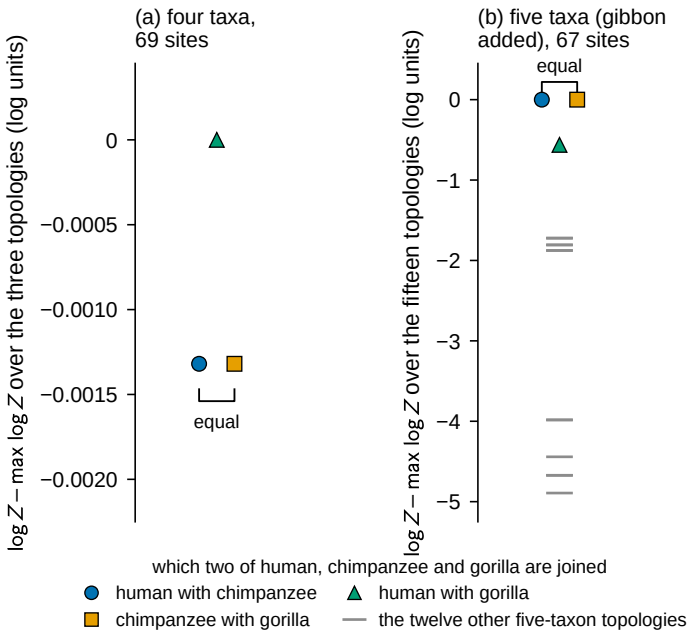

Figure 5: The resolution of the human, chimpanzee and gorilla trichotomy preferred at four taxa ranks third at five, while the two resolutions related by exchanging human and gorilla remain exactly tied. Exact $\log Z$ relative to the best topology (JC69, exponential branch-length priors of rate 4/3, uniform topology prior). (a) The quartet without gibbon ($N = 69$). (b) The fifteen five-taxon topologies ($N = 67$); gray ticks, the twelve topologies that are not resolutions of the trichotomy. "Equal" marks values identical as ratios of integers, as the exchange symmetry of the counts requires. The two vertical scales differ; so does N, because the gap-free columns are counted separately for each taxon set (on the 67 columns of (b) the quartet margin of (a) is 0.0014 log units; Supplement S12).

the alignments differ or to the change of prior: on the quintet's 67 columns the four-taxon margin for (human, gorilla) is 0.0014 log units, and under the branch-length prior that the five-taxon model induces on the quartet when gibbon's row is summed out (one of the five branches, each with probability 1/5, has the length of two exponential branches in series) it is 0.0016, both exact and both with the tie of Eq. (17) intact (Supplement S12). Thus the four-taxon tie persists as a tie between the two best five-taxon trees, and the resolution preferred at four taxa ranks third. At prior means of 0.1 and 0.01 substitutions per site the (human, gorilla) resolution still ranks below the tied pair, by 0.5660 and 0.5787 log units, slightly more than at 0.75 (Table 4, bottom block). These two margins are quadrature values, accurate to about 10^{-13} log units. The tie is specific to this locus: because the tRNA-Lys gene of the same five mitochondrial genomes admits no exchange of taxa that leaves its pattern counts unchanged, concatenating the two genes under one set of branch lengths separates the pair.

The tRNA-Phe quintet alignment contains five quartets (one per taxon left out), each with a preferred split under exact evaluation (middle block of the table). We include a quartet in the amalgamation of these splits into a five-taxon tree only when its log Bayes factor over the closest rival exceeds 2.0 log units, a threshold that was fixed before the comparison. Four qualify, by 2.02 to 2.40 log units, and are mutually compatible; the same four would qualify at a threshold

Table 4: The accepted hominoid tree and its image under exchange of human and gorilla share the largest evidence exactly; the (human, gorilla) resolution is third. Exact $\log Z$ (quadrature where marked †), tRNA-Phe quintet ($N = 67$, $m = 13$), JC69, exponential branch-length priors of rate 4/3 (mean 0.75), uniform topology prior. Upper block: the nine distinct values among fifteen topologies (tied topologies share a rank), the exchange partner of each ("itself" if mapped to itself) and Δ, log Bayes factor for the maximum. Middle block: the five leave-one-out quartets, preferred split and margin over the closest rival. Bottom block: the tied pair and the (human, gorilla) resolution at prior means 0.1 and 0.01 (exponential rates 10 and 100), with Δ the log Bayes factor for the pair. H human, C chimpanzee, G gorilla, O orangutan, Gi gibbon; last digit rounded.

rank	topology	exchanged image	$\log Z$	Δ (log units)
1	((H,C),G,(O,Gi))	((C,G),H,(O,Gi))	−189.6451165	0
3	((H,G),C,(O,Gi))	itself	−190.2076498	0.5625
4	((C,O),H,(G,Gi))	((H,Gi),G,(C,O))	−191.3674404	1.7223
6	((H,C),O,(G,Gi))	((C,G),O,(H,Gi))	−191.4518421	1.8067
8	((H,O),C,(G,Gi))	((G,O),C,(H,Gi))	−191.5212229	1.8761
10	((H,C),Gi,(G,O))	((C,G),Gi,(H,O))	−193.6282837	3.9832
12	((H,G),Gi,(C,O))	itself	−194.0873426	4.4422
13	((C,Gi),H,(G,O))	((H,O),G,(C,Gi))	−194.3171364	4.6720
15	((H,G),O,(C,Gi))	itself	−194.5378508	4.8927

leave-one-out quartets of the same alignment

taxon left out	N	preferred split	log BF to closest rival
human	67	(C,G)|(O,Gi)	2.3013
chimpanzee	67	(H,G)|(O,Gi)	2.0225
gorilla	67	(H,C)|(O,Gi)	2.3013
orangutan	69	(H,C)|(G,Gi)[a]	2.3986
gibbon	69	(H,G)|(C,O)[b]	0.0013

the three resolutions of the trichotomy at smaller prior means

prior mean	topology	$\log Z$	Δ (log units)
0.1	((H,C),G,(O,Gi)) = ((C,G),H,(O,Gi))	−178.3150430†	0
	((H,G),C,(O,Gi))	−178.8810224†	0.5660
0.01	((H,C),G,(O,Gi)) = ((C,G),H,(O,Gi))	−179.7455451†	0
	((H,G),C,(O,Gi))	−180.3242311†	0.5787

[a] Exactly tied with (H,Gi)|(C,G), its image under the exchange; the margin is over the third split (H,G)|(C,Gi). [b] The four-taxon alignment of Section 4.2; its other two splits are the exact tie of Eq. (17). †Floating-point Gauss–Jacobi quadrature with 34 nodes per branch, which integrates the polynomial factor of the integrand (degree 67 in each variable) exactly against the prior, apart from double-precision rounding; error of order 10^{-13} log units (below 10^{-12}). At $\alpha = 1$, where the exact values are known, the same quadrature reproduces all fifteen to within 1.5×10^{-13} log units. One value is printed for the exchanged pair, whose two marginal likelihoods are the same integral; at mean 0.01 both were computed and agree to 10^{-13} log units.

of 1.0 log unit, and none at 2.5 log units. For the quartet without orangutan, two splits related by the exchange of human and gorilla are exactly tied, and the quartet qualifies by their margin over the third split. Whichever of the two is taken as preferred, a single five-taxon tree induces all four splits; the two resulting trees are the tied pair at the top of the table. The fifth, the quartet without gibbon, prefers (human, gorilla)|(chimpanzee, orangutan), which neither of the two tied trees induces; hence no five-taxon tree induces all five preferred splits, whatever the threshold. Quartet-based reconstruction is known from simulation to be less accurate than joint analysis (Ranwez and Gascuel, 2001); in this instance neither the quartet values nor the five-taxon values carry Monte Carlo error, so the disagreement is a property of the alignment itself under the model.

5.3 Stepping-stone estimates on the quintet

We ran stepping-stone sampling (Xie et al., 2011) in MrBayes (Ronquist et al., 2012) and in RevBayes (Höhna et al., 2016) on the three leading five-taxon topologies and on the three topologies of the quartet without gibbon (the hominid tRNA-Phe quartet of Section 4.2). Apart from the MrBayes chain length, the settings were those used for the tRNA-Gln quartet: JC69, topology fixed, branch lengths exponential with rate parameter 4/3, 50 stones with spacing parameter 0.4, and four independent runs per topology and program. The MrBayes chains were twice as long per stone as at the tRNA-Gln baseline, 5.1 million generations in all.

The MrBayes means lie 0.02 to 0.04 log units below the exact values, and the RevBayes means within 0.03 log units of them, with deviations of both signs (Fig. 4b,c). Both programs recover the margin of the tied pair over the third resolution, MrBayes giving 0.568 ± 0.027 and RevBayes 0.523 ± 0.066 (between-run standard deviations of the two topologies combined in quadrature), each within one such deviation of the exact margin.

Neither program can establish the exact tie of the leading pair or resolve the margin of the hominid quartet, since no finite sample can show a difference to be exactly zero or detect one far below its Monte Carlo error. The four-taxon margin is less than a tenth of either program's between-run standard deviation on that alignment (0.017 to 0.067 log units), and both programs return three overlapping estimates there. In the limit in which both internal branches have length zero the fifteen likelihood functions coincide. This is the star-tree setting of Yang and Rannala (2005) and of Steel and Matsen (2007), for which the polytomy prior of Lewis et al. (2005) was proposed. The African-ape part of this alignment appears to lie near that limit, close enough that the contribution of the branch-length prior can be distinguished from that of its two informative columns only by exact computation.

6 Multiple loci

When many loci are available for the same taxa, their single-locus marginal likelihoods can be combined in several ways, which differ in how much weight a single aberrant locus can receive.

6.1 Sums, products and votes

Linked loci on the same taxa that share branch lengths are concatenated: their site-pattern counts add and the evidence is evaluated once on the pooled counts. Loci that share a topology but not their branch lengths contribute independent marginal likelihoods, which multiply, and their log Bayes factors therefore add. Two hominid genes of Table 2 illustrate the difference

between the two rules: mitochondrial tRNA-Lys prefers the human–chimpanzee pairing over its nearest rival (the human–orangutan pairing) by 0.91 log units, whereas for tRNA-Phe those two topologies have identical marginal likelihoods. The product rule thus gives a combined log Bayes factor of 0.91 log units, to which tRNA-Phe contributes a term of exactly zero. Pooling the counts instead removes the symmetry under exchange of human and gorilla that forces the tie when tRNA-Phe is analyzed by itself.

Under the multispecies coalescent the topology itself may differ between loci. Each locus is then counted as one vote for the topology it prefers, optionally only when the preference exceeds a chosen threshold, and the vote frequencies are the input to quartet-based summary methods (Zhang et al., 2018). The sum of log Bayes factors over loci, in contrast, includes the term of each locus in full, and a single paralogous, contaminated or misaligned locus can dominate it. For example, six of the eight Rfam quartets in Table 2 other than the tRNA-Gln reference quartet consist of non-orthologous tRNAs and nonetheless support one split over the alternatives by up to 6.42 log units. A summed log Bayes factor is therefore best reported together with its largest single-locus term and a note of whether the sign of the sum changes when that locus is removed. Counting votes gives each locus the same weight however large its margin, with some loss of power, and vote frequencies collected only from loci whose margin exceeds a threshold must be calibrated by simulation before they are interpreted under the coalescent.

Voting with abstention, in which a locus whose margin falls below the chosen threshold is not counted, was applied to the placental root (Schwartz et al., 2026). The per-gene statistic in that study was, in place of the evidence, a maximized Jukes–Cantor likelihood bounded on both sides in exact arithmetic. A gene voted only when its bounds separated one rooted topology from all the others. Each of 13,088 genes was analyzed 24 times, each time with one representative species drawn from each of the four placental lineages. In each analysis the fifteen rooted arrangements of the four representatives were compared, three of which are the candidate positions of the placental root (Romiguier et al., 2013; Tarver et al., 2016). Of the 314,112 analyses, 3,884 voted, and by that study's criterion the votes do not order the three positions (Schwartz et al., 2026). If the exact five-taxon evidence were the per-gene statistic, a gene would vote when its log Bayes factor for one root position over each of the other two exceeded the chosen threshold.

The rationality result of Section 3.1 also covers Kimura's model with a transition and a transversion rate on every branch, under the prior specified there; the Jukes–Cantor and Kimura marginal likelihoods of an alignment can then both be evaluated exactly, each marginalized over the three topologies. The log Bayes factor between JC69 and this branch-heterogeneous Kimura model thus carries no estimator error. In the *Anopheles* genome scan described below, which scores each window under three quartets, fifteen window–quartet pairs have at most 30 gap-free sites and at most six distinct site patterns. We evaluated this Bayes factor on eight of them, drawn at random without replacement; the eight draws comprise six distinct windows, two of which were drawn under two quartets each. JC69 is preferred by 7.99 log units per draw on average, with a standard error over the eight draws of 0.63. At so few sites much of this margin is presumably the penalty for five additional rate parameters rather than evidence about the substitution process in these sequences. Supplement S9 gives the details, together with a larger comparison on loci of full tRNA length for which the Kimura marginal likelihoods had to be estimated by sampling (Schwartz, 2026).

Loci on one non-recombining molecule can still prefer different topologies, as the three mitochondrial tRNA genes of about 70 sites do for human, chimpanzee, gorilla and orangutan.

Although the three genes share one genealogy and their marginal likelihoods are exact, their support for the accepted human–chimpanzee pairing over the best alternative at this prior ranges from about −2.8 to +0.9 log units.

6.2 Two applications

Jensen et al. (2022) placed the extinct San Cristóbal giant tortoise sister to *Chelonoidis hoodensis* and *C. abingdonii* with posterior probability 1.0, from one mitochondrial control-region locus of about 700 nucleotides. We tested that placement with exact quartet marginal likelihoods under a design fixed in advance. Two quartets, each consisting of the extinct tortoise, a modern San Cristóbal tortoise, one of the two sister species and an outgroup, were scored separately on the six consecutive 118-column segments of our 708-column alignment of the locus. The locus was divided because quartet alignments of its full length exceeded what our programs accepted when the design was fixed. The statistic was the sum of the twelve exact log Bayes factors for the published split over the pairing of the extinct with the modern San Cristóbal tortoise. Because the segments share one genealogy and one set of branch lengths, this sum is not the log Bayes factor of the locus (Section 6.1) but a test statistic. It is used only for the decision rule fixed in the design and is never quoted as a combined evidence. The placement was to count as confirmed if the sum was at least 3.0, a threshold fixed in the written design, and the two quartet sums had the same sign. Because an ancient sequence may carry deamination damage, a null distribution was built from 150 replicates in which the modern tortoise's sequence, with such damage introduced, replaced that of the extinct tortoise (Supplement S7).

The observed sum of the twelve log Bayes factors is 0.90, with quartet sums of −3.61 and +4.51, whereas the damage replicates all fall below −43. Hence the locus does not confirm the placement at this threshold, and the extinct tortoise's sequence is not a damaged copy of the modern islander's. Simulations at the branch lengths fitted to the five sequences show that the shortfall is compatible with either placement and says chiefly that the internal branches of this locus are short (Supplement S7). This outcome does not show the published placement, which rests on a clock-model analysis of 93 haplotypes, to be wrong; it says only that this locus, read this way, does not confirm it. Five of the twelve terms are below one log unit in magnitude, but each is an exact value, so that their signs and the two quartet sums are determined without Monte Carlo error.

Fontaine et al. (2015) found that, in the *Anopheles gambiae* complex of malaria mosquitoes, *An. arabiensis* groups with different species in phylogenies of autosomal and of X-chromosome windows, and attributed the autosomal signal to gene flow. We rebuilt their sliding-window analysis, taking no position on gene flow, and scored 81,113 windows on three quartets each under JC69 in floating-point arithmetic. The rebuilt scan recovers the autosome-versus-X reversal of the *An. arabiensis* grouping that Fontaine et al. reported, on every autosomal arm and however exactly tied windows are counted: set aside, split fractionally among the tied topologies, or replaced by mean posterior topology probabilities (Supplement S8). Exact values were computed for the 5,375 window–quartet pairs that met two limits set in advance, at most seven mixed sites and at most 150 sites. Of these, 2,997 are exact ties for the largest evidence, which a floating-point scan can resolve only by rounding error. Among the 2,378 with a unique exact maximum the highest floating-point score falls on the same topology in 2,374, and the four exceptions involve separations of at most 0.016 log units.

7 Discussion

We have shown that the marginal likelihood of a four- or five-taxon tree under the Jukes–Cantor model, with independent exponential branch-length priors of rational rate, is a ratio of two integers. As an integral over the branch lengths it has the form of a parametric Feynman integral (Smirnov, 2012), known in algebraic statistics as an A-hypergeometric integral (Sturmfels and Telen, 2021; Borinsky et al., 2023), and the term-by-term evaluation in modular arithmetic used here was adopted from work on Feynman integrals. We have computed such values exactly for real alignments, extended the rationality result to other time-reversible models, and used the exact values as references for testing Monte Carlo estimators.

7.1 Uses of the exact values

Table B1 gives exact log marginal likelihoods of fourteen public alignments, and Table 3 adds quadrature values for two longer ones, under settings that MrBayes (Ronquist et al., 2012) and RevBayes (Höhna et al., 2016) accept as input, including the branch-length prior. These values allow an estimator to be tested against an exact reference on aligned sequences as it has been tested in conjugate problems (Xie et al., 2011), rather than against longer runs of other estimators (Fourment et al., 2020; Fan et al., 2011; Wang et al., 2023). In Section 4.3 such tests separated two systematic errors from Monte Carlo error: the discretization bias of thermodynamic integration on a coarse schedule, and a small offset common to all topologies in the MrBayes stepping-stone estimates at matched settings. With the exact values as reference we traced that offset to the proposal ratio of one of the program's branch-length moves (Section 4.3). Neither error is visible when estimators are compared only with one another, because each program was self-consistent within its Monte Carlo error.

Exact values may also serve as per-quartet weights, free of Monte Carlo error, in quartet-based tree assembly (Ranwez and Gascuel, 2001). At five taxa they allow quartet amalgamation to be compared directly with the joint analysis of all topologies, as in the hominoid quintet, where no single tree induces the preferred splits of all five quartets. Finally, an exact tie between two topologies cannot be broken by any reanalysis under the same model, and a margin that arises only in the integration over branch lengths, as on the hominid tRNA-Phe quartet, indicates that the locus has no informative site. Many of the *Anopheles* genome-scan windows small enough to evaluate exactly were such ties, where a floating-point scan reports a preference for one topology that is due only to rounding error.

7.2 Scope and limitations

The exact values on real data are four- and five-taxon marginal likelihoods under the Jukes–Cantor model, except the Kimura-model values for six *Anopheles* windows of at most thirty sites. The Jukes–Cantor values have independent exponential branch-length priors of mean 0.75 substitutions per site; one quartet was also evaluated at means 0.1 and 0.01 (Section 4.2). The Kimura-type values integrate over a transition and a transversion rate on every branch under the prior of Section 3.1, and no general time-reversible value has been computed on real data. The fixed-parameter Kimura and time-reversible statements were confirmed numerically on synthetic alignments of three to five sites. Rate variation across sites is treated only in the smallest cases. Hence for trees of realistic size, and under the substitution models in routine use, Monte Carlo estimation remains necessary.

The first restriction is the number of taxa. The rationality argument holds for any number of leaves, but the cost rises steeply with it. Each added taxon brings two more edge variables, and the direct expansion grows from $(N+1)^5$ terms at four taxa to $(N+1)^7$ at five and $(N+1)^9$ at six. We have not attempted six taxa, where 105 topologies would each need a value.

The second restriction is alignment size. At fixed taxon number the cost is governed by the mixed-site count m rather than by the alignment length N. The timed tRNA quartets, with up to 59 mixed sites, took from seconds to about seventy minutes per quartet for its three topologies together, on 32 to 72 cores of one workstation. The two mitochondrial alignments of about 900 sites have two hundred or more mixed sites and are out of reach on a workstation; a graphics-processor computation was tested at reduced scale but not completed (Supplement S3).

The third restriction is the substitution model. The Kimura model needs two variables per edge and the general time-reversible model a matrix inverse (Lemma 2) in place of monomial integration; the cost of the Kimura evaluation has not been measured at tRNA length, and the matrix-inverse identity proves rationality without giving a practical algorithm. Neither rationality statement covers substitution parameters integrated over their usual continuous priors, as these models are normally run. Under $+\Gamma$ the term-by-term argument fails even at three taxa and one site. The evidence is then a constant (presumably transcendental) that can be computed as precisely as required, and we have not gone beyond the smallest cases.

The last restriction concerns the branch-length prior. Rationality holds for any rational exponential rate, and under a gamma prior with rational parameters the evidence is an algebraic number (rational if the shape is an integer); hence the dependence of a Bayes factor on the prior mean can be computed exactly. For the hominoid tRNA-Phe alignments, lowering the prior mean from 0.75 to 0.1 or 0.01 leaves the sign of the quartet margin (exact) and the order of the three quintet resolutions (by quadrature) unchanged (Sections 4.2 and 5.2). Priors that couple the branches, such as compound Dirichlet priors on tree length (Rannala et al., 2012), break the edge-by-edge factorization on which rationality rests and are not treated.

7.3 Future directions

The evidence satisfies a linear recurrence in each pattern count, of order fourteen for the Jukes–Cantor quartet in the count of constant sites. For a general alignment, however, finding the recurrence has so far cost more than direct evaluation, and at five taxa none has been found. In the small tempered example treated here, at the stones where the site-pattern probabilities carry half-integer exponents, the power-posterior normalizing constant is known to any precision from its differential equation, though we have no complete closed form for it. We could evaluate the stepping-stone path deterministically only at half-integer exponents, short of the Beta-quantile spacings in common use (Xie et al., 2011), because no such equation was found for exponents with larger denominators. The natural next step is a method that constructs these recurrences and differential equations without a search. Such a method would bear on all three problems, and probably on six taxa as well.

For four- and five-taxon alignments of the sizes treated here, then, the Jukes–Cantor marginal likelihood need not be estimated. It can be computed exactly, and for a quartet of tRNA length on one multicore workstation the computation takes a few times the wall-clock time of a single-core stepping-stone run. We suggest that such exact values supplement conjugate test problems and long reference runs whenever a marginal-likelihood estimator is tested, and we offer the values of Table B1 for that purpose.

Acknowledgments

[To be supplied by the authors.]

Funding

[To be supplied by the authors.]

Data and code availability

The algorithms of this paper are implemented in the open-source package JACK & JILL (Schwartz, 2026). The data file of Supplement S10 gives the site-pattern counts and the numerator and denominator of each tabulated four-taxon exact value, with SHA-256 checksums, and lists the sequence accessions of every alignment; the alignments themselves accompany the supplement as NEXUS files. Three further machine-readable files give the integer coefficients of the order-14 recurrence of Section 3.2 and of the two differential equations of Section 3.4 (Supplements S1, S2 and S4 name the files). These files and the package source code will be deposited in the Dryad Digital Repository on acceptance.

A Derivations and algorithm

Throughout this appendix $T = 12|34$, $x_e = e^{-\mu t_e}$ is the variable of edge e (pendant edges 1 to 4, internal edge 5), y_k is the number of sites in pattern class k, and $N = \sum_k y_k$ is the alignment length. Lemma 1 is proved in the main text; its term-by-term formula and its denominator bound $D_N(\alpha)$ are used again in the algorithm of this appendix.

A.1 Proof of Lemma 2

Let Q be the rate matrix of a time-reversible model with stationary distribution π, scaled to expected substitutions per site; its entries are then rational functions of the exchangeabilities and of π. The transition matrix is $P(t) = e^{Qt}$, and each branch length carries an independent exponential prior with rate λ.

Consider one edge. Conditional on the states at its two ends, the N sites evolve independently along it. Their joint transition matrix on $\{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}^N$ is therefore the Kronecker power $P(t)^{\otimes N} = (e^{Qt})^{\otimes N} = e^{Q^{\oplus N}t}$, where $Q^{\oplus N} = \sum_{i=1}^{N} I \otimes \cdots \otimes Q \otimes \cdots \otimes I$ (with Q in position i) is the corresponding rate matrix on 4^N states. For any square matrix M whose eigenvalues have real part below λ, $\int_0^\infty e^{Mt}\,\lambda e^{-\lambda t}\,\mathrm{d}t = \lambda\,(\lambda I - M)^{-1}$, which is the matrix form of the scalar identity $\int_0^\infty e^{qt}\lambda e^{-\lambda t}\,\mathrm{d}t = \lambda/(\lambda - q)$. The eigenvalues of a reversible Q are real and nonpositive and those of $Q^{\oplus N}$ are sums of N of them; since $\lambda > 0$, the identity applies with $M = Q^{\oplus N}$ and gives Eq. (10). By Cramer's rule each entry of $\lambda(\lambda I - Q^{\oplus N})^{-1}$ is a rational function of λ and the entries of Q.

Now write the likelihood of the N sites jointly by pruning: it is a finite sum, over joint internal-node states, of $\pi^{\otimes N}$ at the root times one entry of $P(t_e)^{\otimes N}$ per edge. Because this sum is multilinear in the per-edge matrices and the branch lengths are independent under the prior, the prior expectation may be taken edge by edge. Thus Z is the same sum with every $P(t_e)^{\otimes N}$

Table A1: The fifteen Jukes–Cantor site-pattern classes of the quartet 12|34 and their polynomials $P_k = 256\,p_k$, given by the coefficient $c_{k,a}$ of each monomial x^a; m_k is the number of the 256 leaf patterns in class k, and the last row is the identity $\sum_k m_k p_k \equiv 1$.

class k	m_k	1	x_1x_2	x_3x_4	$x_1x_2x_3x_4$	$x_1x_3x_5$	$x_1x_4x_5$	$x_2x_3x_5$	$x_2x_4x_5$	$x_1x_2x_3x_5$	$x_1x_2x_4x_5$	$x_1x_3x_4x_5$	$x_2x_3x_4x_5$	$x_1x_2x_3x_4x_5$
xxxx	4	1	3	3	9	3	3	3	3	6	6	6	6	12
xxyy	12	1	3	3	9	−1	−1	−1	−1	−2	−2	−2	−2	−4
xyxy	12	1	−1	−1	1	3	−1	−1	3	−2	−2	−2	−2	4
xyyx	12	1	−1	−1	1	−1	3	3	−1	−2	−2	−2	−2	4
xyyy	12	1	−1	3	−3	−1	−1	3	3	−2	−2	−2	6	−4
xyxx	12	1	−1	3	−3	3	3	−1	−1	−2	−2	6	−2	−4
xxyx	12	1	3	−1	−3	−1	3	−1	3	−2	6	−2	−2	−4
xxxy	12	1	3	−1	−3	3	−1	3	−1	6	−2	−2	−2	−4
xxyz	24	1	3	−1	−3	−1	−1	−1	−1	−2	−2	2	2	4
xyzz	24	1	−1	3	−3	−1	−1	−1	−1	2	2	−2	−2	4
xyxz	24	1	−1	−1	1	3	−1	−1	−1	−2	2	−2	2	0
xyzx	24	1	−1	−1	1	−1	3	−1	−1	2	−2	−2	2	0
xyyz	24	1	−1	−1	1	−1	−1	3	−1	−2	2	2	−2	0
xyzy	24	1	−1	−1	1	−1	−1	−1	3	2	−2	2	−2	0
xyzw	24	1	−1	−1	1	−1	−1	−1	−1	2	2	2	2	−4
$\sum_k m_k c_{k,a}$	256	256	0	0	0	0	0	0	0	0	0	0	0	0

replaced by its prior mean (10), and each term of the sum is then a rational function of the exchangeabilities, π and λ. □

The proof of Lemma 2 establishes that Z is rational but gives no practical algorithm for computing it, because the prior-averaged matrix $\lambda(\lambda I - Q^{\oplus N})^{-1}$ has dimension 4^N. The general time-reversible values mentioned in the main text (three taxa at three sites, and five taxa at four sites) were computed by forming that matrix directly. Under a gamma prior on t_e with shape k and rate λ, the edge variable $x_e = e^{-\mu t_e}$ has density $\alpha^k x_e^{\alpha-1}(-\log x_e)^{k-1}/\Gamma(k)$ on $[0,1]$, where $\alpha = \lambda/\mu$, and the gamma-prior identity (11) of the main text follows from the substitution $x = e^{-s}$ in its integral.

A.2 Pattern polynomials

Table A1 lists the fifteen pattern polynomials that appear in the integrand of the evidence integral (6). They follow from pruning on $T = 12|34$ with the Jukes–Cantor transition probabilities $\frac{1}{4} + \frac{3}{4}x_e$ (same state) and $\frac{1}{4} - \frac{1}{4}x_e$ (different states) and a uniform distribution at the root, whose position is immaterial by reversibility. A class label indicates which of leaves 1 to 4 share a state; the class `xxyx`, for example, comprises the twelve leaf patterns in which leaves 1, 2 and 4 agree and leaf 3 differs. The polynomials for 13|24 and 14|23 follow by exchanging leaves 2 and 3, or 2 and 4, in both the class labels and the edge subscripts.

A.3 Algorithm

Algorithm 1 states the term-by-term evaluation from the proof of Lemma 1 as a procedure in modular arithmetic; the Jukes–Cantor values in this paper were obtained from the same term-by-term formula, evaluated either directly, as in the algorithm, or in algebraically equivalent arrangements that need less memory. Let $P_k = 256\,p_k$ denote the integer-coefficient pattern polynomials, and let $F = \prod_k P_k^{y_k} = \sum_a C_a\, x_1^{a_1} \cdots x_5^{a_5}$, with integer C_a and every $a_e \le N$. At $\alpha = 1$ the integral of each monomial against the prior is the reciprocal of $\prod_e (a_e + 1)$, which divides L_{N+1}^5, where L_n denotes the least common multiple of $1, 2, \ldots, n$, and $Z < 4^{-N}$ because

Algorithm 1. Exact marginal likelihood of a quartet topology under JC69 with independent exponential branch-length priors of rate parameter 4/3 (mean 0.75 expected substitutions per site; $\alpha = 1$).

Input: topology T; pattern counts $\mathbf{y} = (y_1, \ldots, y_{15})$ in the frame of T (Table A1); $N = \sum_k y_k$. *Output:* $Z(T; \mathbf{y}, 1)$ as a reduced fraction.

1. *Size.* Set $L = \mathrm{lcm}(1, \ldots, N+1)$ and $D'_N = 256^N L^5$. Choose distinct primes $p_1, \ldots, p_r$, each larger than $N+1$ and smaller than 2^{31}, with $p_1 p_2 \cdots p_r > D'_N$.
2. For each prime p:
   (a) *Expansion.* Starting from the constant array 1, multiply in the N factors of $F = \prod_k P_k^{y_k}$ one at a time, reducing modulo p. Each multiplication by a multilinear P_k is a convolution of the current five-index coefficient array with a $2 \times 2 \times 2 \times 2 \times 2$ kernel holding at most thirteen nonzero integers. The result is $C_a \bmod p$ for $a \in \{0, \ldots, N\}^5$.
   (b) *Integration.* Form $R_p = \sum_a C_a \prod_{e=1}^{5} \big(L/(a_e+1)\big) \bmod p$. Every $L/(a_e+1)$ is an integer, and $R_p \equiv D'_N Z \pmod p$ by the term-by-term formula of Lemma 1.
3. *Reconstruction.* By the Chinese remainder theorem find the unique integer $0 \le M < p_1 \cdots p_r$ with $M \equiv R_p \pmod p$ for every p. By Eq. (A1), $M = D'_N Z$. Return M/D'_N in lowest terms.

For a rational prior exponent $\alpha \neq 1$, replace $L/(a_e+1)$ by the corresponding integer multiple of $1/(a_e+\alpha)$ from the bound of Lemma 1 and restore the factor α^5. For 13|24 or 14|23, relabel the counts as in Section A.2.

no pattern probability exceeds $\frac{1}{4}$. Hence, with a bound D'_N sharper than the $D_N(1)$ of Lemma 1 and dividing it,

$$D'_N Z \in \mathbb{Z}, \qquad 0 < D'_N Z < D'_N, \qquad D'_N = 256^N L_{N+1}^5. \tag{A1}$$

This bound determines the number of word-size primes that are needed; for the tRNA quartets ($N = 63$ to 75), for example, 34 to 40 primes of 31 bits suffice.

The coefficient array in step 2(a) has $(N+1)^5$ entries per prime (about 8×10^8 at $N = 60$), which limits this direct form to alignments of about sixty sites. Most values on real alignments were computed with the mixed-site reduction of Section 3.2, in which only the m mixed sites occupy a five-index array, of side $m+1$, and the other sites a three-index array. The programs used to compute the exact values in each table are described in Supplement S3.

B Benchmark values

Table B1 collects the exact log marginal likelihoods of all three topologies for the fifteen four-taxon benchmark alignments, as reference values for testing marginal-likelihood estimators. The fifteen five-taxon values of Table 4 (upper block) are exact in the same sense: each marginal likelihood is a ratio of two integers, and only its printed logarithm is rounded. In the parameterization of MrBayes (unconstrained exponential) and of RevBayes, the branch-length prior of both tables has rate parameter 4/3 on each branch, or 10 and 100 in the two lines at smaller prior means. The topology prior does not enter $\log Z$. Supplement S10 gives, for each of the four-taxon alignments, the sequence accessions, the site-pattern counts that form the complete input to Algorithm 1, and the numerator and denominator of each value (48 to 1,682 digits). For the five-taxon values it gives the accessions and $\log Z$ to 60 digits. Gauss–Legendre quadrature

agrees with the $N = 20$ row to 58 or more significant digits at 70-digit working precision, and with each of rows 1–14 to about 16 digits.

Table B1: Exact log marginal likelihoods of the three topologies of fifteen four-taxon alignments (natural logarithms, rounded to ten decimals), under JC69 with independent exponential branch-length priors of rate 4/3 (mean 0.75 expected substitutions per site; $\alpha = 1$): the twelve tRNA quartets of Table 2, the lysozyme and 5S rRNA alignments of Table 3, and the twenty-site case of Eq. (12). Taxa are numbered in the order listed, m is the number of columns that are neither constant nor of class `xxyy`, and p/q gives the digit counts of numerator and denominator. Rows 1 and 15 contain exact ties; the last two lines repeat row 1 at prior means 0.1 and 0.01. The fifteen five-taxon values are in Table 4.

	alignment	taxa 1, 2, 3, 4	N	m	$\log Z(12\|34)$	$\log Z(13\|24)$	$\log Z(14\|23)$	digits p/q
1	mt tRNA-Phe	human, chimpanzee, gorilla, orangutan	69	10	−161.9379803099	−161.9366608916	−161.9379803099	217/287; 219/290; 217/287
2	mt tRNA-Lys	human, chimpanzee, gorilla, orangutan	70	11	−183.9485776844	−187.8629542754	−184.8590399779	223/303; 220/302; 220/300
3	mt tRNA-Leu(UUR)	human, chimpanzee, gorilla, orangutan	75	7	−162.2404586141	−159.4224712233	−162.2450370047	245/315; 244/313; 246/316
4	mt tRNA-Gln	mouse, rat, chicken, *Xenopus*	68	25	−261.4030641503	−275.9398266685	−276.9593356783	188/302; 183/302; 182/303
5	tRNA, primates	human, lemur, macaque, squirrel monkey	67	48	−292.0539158983	−284.5062632346	−291.7637313877	171/298; 179/302; 173/300
6	tRNA, vertebrates	chicken, mouse, rat, *Xenopus*	66	51	−345.3895438654	−346.0888693220	−340.5515437693	149/299; 147/297; 148/296
7	tRNA, bacteria	*B. sub.*, *E. coli*, *M. pneu.*, *T. therm.*	70	45	−344.8344419557	−347.4083575895	−345.9414813474	163/313; 161/311; 162/313
8	tRNA, fungi	*A. nid.*, *N. cra.*, *S. cer.*, *S. pom.*	71	53	−365.8107585974	−363.5684109499	−366.6350823253	157/316; 158/316; 160/320
9	tRNA, insects	mosquito, silkmoth, fly, locust	63	59	−350.4547300474	−349.9209712802	−349.8652403842	131/283; 132/284; 133/285
10	tRNA, plants[b]	*A. thal.*, *N. tab.*, *T. aest.*, *Z. mays*	70	50	−354.9284572642	−346.7915993289	−353.2121437076	157/311; 166/316; 159/312
11	tRNA, cross-domain[b]	*E. coli*, human, *M. vann.*, *S. cer.*	70	44	−335.1522846057	−334.1316747700	−329.1853760092	171/316; 171/316; 173/316
12	tRNA, archaea	*Haloarcula*, *Haloferax*, *M. vann.*, *T. acid.*	72	48	−354.3041117412	−349.9880386581	−353.0105152938	177/331; 177/329; 177/330
13	lysozyme[c], 390 bp	human, gibbon, colobus, marmoset	390	39	−829.7542319579	−844.3189101943	−846.1103094375	1320/1680; 1314/1681; 1315/1682
14	5S rRNA	mouse, chicken, *Xenopus*, *Drosophila*	116	38	−345.7452740049	−344.7711531629	−344.4186439167	362/512; 362/512; 366/515
15	synthetic, $N = 20$	(pattern counts only)	20	6	−87.6492433492	−90.8374830424	−90.8374830424	49/87; 48/87; 48/87
row 1 at smaller prior means (exponential rates 10 and 100, that is, $\alpha = 15/2$ and 75; Section 4.2); exact								
1	mt tRNA-Phe, mean 0.1	human, chimpanzee, gorilla, orangutan	69	10	−153.6812699266	−153.6802528459	−153.6812699266	286/352; 287/354; 286/352
1	mt tRNA-Phe, mean 0.01	human, chimpanzee, gorilla, orangutan	69	10	−153.8331832786	−153.8329798033	−153.8331832786	371/438; 375/441; 371/438

[b] Non-orthologous tRNAs (nuclear with organellar genes); the values refer to these alignments, not to the species tree.
[c] Four-species subset of the seven-species alignment of Messier and Stewart (1997) as distributed with PAML (the small lysozyme example data set). Sequence accessions (rows 1–3) and Rfam seed-row identifiers (rows 4–12, 14), with the site-pattern counts of each row, are listed in Supplement S10; row 15 is synthetic (counts only).

References

Allman, E. S. and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13:1101–1113.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., and Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.*, 29:2157–2167.

Borinsky, M., Sattelberger, A.-L., Sturmfels, B., and Telen, S. (2023). Bayesian integrals on toric varieties. *SIAM Journal on Applied Algebra and Geometry*, 7(1):77–103. arXiv:2204.06414.

Brown, W. M., Prager, E. M., Wang, A., and Wilson, A. C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, 18:225–239.

Chor, B., Hendy, M. D., Holland, B. R., and Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Biol. Evol.*, 17:1529–1541.

Chor, B. and Snir, S. (2004). Molecular clock fork phylogenies: closed form analytic maximum likelihood solutions. *Syst. Biol.*, 53:963–967.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2:e68.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011). Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol.*, 28:523–532.

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

Fontaine, M. C. et al. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347:1258524.

Fourment, M., Magee, A. F., Whidden, C., Bilge, A., Matsen, IV, F. A., and Minin, V. N. (2020). 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Syst. Biol.*, 69:209–220.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. B*, 70:589–607.

Gel'fand, I. M., Kapranov, M. M., and Zelevinsky, A. V. (1990). Generalized Euler integrals and A-hypergeometric functions. *Adv. Math.*, 84:255–271.

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

Hendy, M. D. and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38(4):297–309.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.*, 65:726–736.

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

Jensen, E. L., Quinzin, M. C., Miller, J. M., Russello, M. A., Garrick, R. C., Edwards, D. L., Glaberman, S., Chiari, Y., Poulakakis, N., Tapia, W., Gibbs, J. P., and Caccone, A. (2022). A new lineage of Galapagos giant tortoises identified from museum samples. *Heredity*, 128:261–270.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume III, pages 21–132. Academic Press, New York.

Kalvari, I. et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49:D192–D200.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, 90:773–795.

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.

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., and Pääbo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90:19–30.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 16:750–759.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55:195–207.

Lewis, P. O., Holder, M. T., and Holsinger, K. E. (2005). Polytomies and Bayesian phylogenetic inference. *Syst. Biol.*, 54:241–253.

Mau, B. and Newton, M. A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.*, 6:122–131.

Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, 385:151–154.

Milkey, A., Korte, E., and Lewis, P. O. (2023). *lorad: Lowest Radial Distance Method of Marginal Likelihood Estimation*. R package version 0.0.1.0, `https://CRAN.R-project.org/package=lorad`.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. B*, 56:3–48.

Oaks, J. R., Cobb, K. A., Minin, V. N., and Leaché, A. D. (2019). Marginal likelihoods in phylogenetics: a review of methods and applications. *Syst. Biol.*, 68:681–697.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.*, 5:568–583.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, 43:304–311.

Rannala, B., Zhu, T., and Yang, Z. (2012). Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Mol. Biol. Evol.*, 29:325–335.

Ranwez, V. and Gascuel, O. (2001). Quartet-based phylogenetic inference: improvements and limits. *Mol. Biol. Evol.*, 18(6):1103–1116.

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E. J. P. (2013). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.*, 30:2134–2144.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, 61:539–542.

Schwartz, M. D. (2026). JaCK & Jill: exact and certified Bayesian evidence for phylogenetic models. Manuscript in preparation.

Schwartz, M. D., Edwards, S. V., and Lewis, P. O. (2026). Resolving ancient branchings using precision phylogenetics. Manuscript in preparation.

Sinsheimer, J. S., Lake, J. A., and Little, R. J. (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, 52(1):193–210.

Smirnov, V. A. (2012). *Analytic Tools for Feynman Integrals*. Springer Tracts in Modern Physics. Springer, Berlin, Heidelberg.

Steel, M. and Matsen, F. A. (2007). The Bayesian "star paradox" persists for long finite sequences. *Mol. Biol. Evol.*, 24(4):1075–1079. arXiv:q-bio/0611007.

Sturmfels, B. and Sullivant, S. (2005). Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12:204–228.

Sturmfels, B. and Telen, S. (2021). Likelihood equations and scattering amplitudes. *Algebraic Statistics*, 12(2):167–186. arXiv:2012.05041.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.*, 4:vey016.

Susko, E. (2008). On the distributions of bootstrap support and posterior distributions for a star tree. *Syst. Biol.*, 57:602–612.

Székely, L. A., Steel, M. A., and Erdős, P. L. (1993). Fourier calculus on evolutionary trees. *Advances in Applied Mathematics*, 14(2):200–216.

Tarver, J. E., dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'Reilly, J. E., King, B. L., O'Connell, M. J., Asher, R. J., Warnow, T., Peterson, K. J., Donoghue, P. C. J., and Pisani, D. (2016). The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.*, 8:330–344.

von Manteuffel, A. and Schabinger, R. M. (2015). A novel approach to integration by parts reduction. *Phys. Lett. B*, 744:101–104. arXiv:1406.4513.

von zur Gathen, J. and Gerhard, J. (2013). *Modern Computer Algebra*. Cambridge University Press, Cambridge, 3rd edition.

Wang, Y.-B., Milkey, A., Li, A., Chen, M.-H., Kuo, L., and Lewis, P. O. (2023). LoRaD: Marginal likelihood estimation with haste (but no waste). *Syst. Biol.*, 72(3):639–648.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, 60:150–160.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401.

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

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14:717–724.

Yang, Z. and Rannala, B. (2005). Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.*, 54(3):455–470.

Zeilberger, D. (1990). A holonomic systems approach to special functions identities. *J. Comput. Appl. Math.*, 32:321–368.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19:153.