PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 4031–4039
doi:10.1098/rstb.2008.0164
Published online 7 October 2008

Empirical evaluation of a prior for Bayesian phylogenetic inference

Ziheng Yang*

Department of Biology, University College London, Darwin Building, Gower Street,
London WC1E 6BT, UK

The Bayesian method of phylogenetic inference often produces high posterior probabilities (PPs) for trees or clades, even when the trees are clearly incorrect. The problem appears to be mainly due to large sizes of molecular datasets and to the large-sample properties of Bayesian model selection and its sensitivity to the prior when several of the models under comparison are nearly equally correct (or nearly equally wrong) and are of the same dimension. A previous suggestion to alleviate the problem is to let the internal branch lengths in the tree become increasingly small in the prior with the increase in the data size so that the bifurcating trees are increasingly star-like. In particular, if the internal branch lengths are assigned the exponential prior, the prior mean μ_0 should approach zero faster than $1/\sqrt{n}$ but more slowly than $1/n$, where n is the sequence length. This paper examines the usefulness of this data size-dependent prior using a dataset of the mitochondrial protein-coding genes from the baleen whales, with the prior mean fixed at $\mu_0 = 0.1n^{-2/3}$. In this dataset, phylogeny reconstruction is sensitive to the assumed evolutionary model, species sampling and the type of data (DNA or protein sequences), but Bayesian inference using the default prior attaches high PPs for conflicting phylogenetic relationships. The data size-dependent prior alleviates the problem to some extent, giving weaker support for unstable relationships. This prior may be useful in reducing apparent conflicts in the results of Bayesian analysis or in making the method less sensitive to model violations.

Keywords: Bayesian inference; clade probabilities; prior; species sampling; substitution model

1. INTRODUCTION

A number of studies have noted that the Bayesian method of phylogeny reconstruction (Rannala & Yang 1996; Mau & Newton 1997; Yang & Rannala 1997; Li *et al.* 2000) often produces very high posterior probabilities (PPs) for trees or clades (Suzuki *et al.* 2002; Cummings *et al.* 2003; Douady *et al.* 2003; Erixon *et al.* 2003; Simmons *et al.* 2004). For example, in the very first calculation of Bayesian PPs for trees, Rannala & Yang (1996) obtained a PP of 0.9999 for the best tree using a small dataset of 11 mitochondrial tRNA genes from five ape species. Analyses of modern larger datasets using the program MrBayes (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) similarly produced high PPs for trees or clades. Certain biological processes can cause the true gene trees to differ from the species tree. For instance, horizontal gene transfer may cause different genes or proteins to have different histories, and gene duplications followed by gene losses may result in paralogues being mistaken as orthologues, again causing the gene tree to differ from the species tree (see Rannala & Yang (2008) for a review). The gene tree–species tree mismatch may also be caused by lineage sorting due to polymorphisms in the common ancestors (e.g. Takahata 1989; Rannala & Yang 2003). In such cases, the different phylogenetic relationships obtained from different genes may have a biological basis, and one may consider the high PPs as an accurate assessment of the information content in the data. However, in many cases, this interpretation is not available, as it is apparent that the phylogenies are untenable biologically even if they are supported by high PPs. For example, the maximum posterior probability (MAP) tree may depend on the substitution model, species sampling (e.g. Bourlat *et al.* 2006) or the type of data being analysed (DNA or protein sequences). In such cases, there is only one true tree, even if it is unknown, and the different MAP trees cannot all be correct, so that the high PPs for the wrong trees are spurious.

As the PP for a tree is the probability that the tree is correct given the prior and data, there can only be three possible reasons for the spuriously high PPs for trees: (i) errors, (ii) violation of the substitution (likelihood) model, and (iii) the impact of the prior. Numerically incorrect PPs may be caused by errors in the theoretical formulation or the computer program or by computational problems in the Markov chain Monte Carlo (MCMC) algorithm, such as lack of convergence or poor mixing. While errors are possible in isolated cases, they are not the fundamental reason for the problem. Misspecification of the substitution model is always a concern in real data analysis. Computer simulations suggest that the use of a simplistic and unrealistic model in the Bayesian analysis may lead to inflated PPs for trees (e.g. Buckley 2002; Huelsenbeck & Rannala 2004; Lemmon & Moriarty 2004). Nevertheless, high

[*] z.yang@ucl.ac.uk

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

PPs for trees were observed in simulations even when the correct substitution model was assumed (Cummings *et al.* 2003; Lewis *et al.* 2005; Yang & Rannala 2005), suggesting that the phenomenon may have to do with the properties of the Bayesian methodology and the impact of the prior. This is indeed the position argued for by Lewis *et al.* (2005), Yang & Rananla (2005) and Yang (2007*a*).

In particular, the case of *star-tree paradox* has recently attracted much attention. Suzuki *et al.* (2002) noted that the PPs for the bifurcating trees for four species can occasionally be quite high even if the data were simulated assuming the star tree. Further analyses suggest that even with arbitrarily long sequences, the PPs for the three bifurcating trees do not converge to one-third each, as common intuition may suggest (Yang & Rannala 2005; Steel & Matsen 2007; Yang 2007*a*; Susko 2008). This phenomenon has been called the star-tree paradox. Instead, the PPs can be very small or very large even though, in fact, no information is available to resolve the tree one way or another (Yang 2007*a*). Yang & Rannala (2005) discussed the connection of the star-tree paradox to Bayesian model comparison, in which the prior on unknown parameters can have a major impact on posterior model probabilities (Lindley 1957; O'Hagan & Forster 2004, pp. 77–79).

Two strategies have been suggested to alleviate the problem of high posterior tree probabilities, both of which manipulate the prior used in the Bayesian analysis. The first is the polytomy prior, which assigns non-zero prior probabilities for multifurcating as well as bifurcating trees (Lewis *et al.* 2005). The second is a prior on the internal branch lengths in the bifurcating trees, which depends on the data size, forcing the bifurcating trees to converge to the star tree when the amount of data increases (Yang & Rannala 2005). Yang (2007*a*) studied Bayesian phylogenetic inference in the case of three species under a molecular clock, using the exponential prior on the internal branch length with the mean $\mu_0 = cn^{-\gamma}$, where n is the number of sites in the sequence. Under the criteria that the PP for (i) each bifurcating tree should approach one-third if the star tree is true and (ii) the bifurcating tree should approach 1 if that bifurcating tree is true, it was determined that $(1/2) < \gamma < 1$. In other words, μ_0 should approach zero faster than $1/\sqrt{n}$ but more slowly than $1/n$.

In this paper, I apply this data size-dependent prior to an empirical dataset to evaluate its usefulness in reducing Bayesian PPs for trees. The dataset consists of mitochondrial protein-coding genes from the baleen whales. In this dataset, the MAP tree is found to be sensitive to the substitution model assumed, species sampling and the type of data analysed (DNA or protein sequences). The Bayesian analysis using the standard prior produced high support for contradicting phylogenetic relationships, while the data size-dependent prior appears to be useful in reducing such conflicts in Bayesian phylogenetic analysis.

2. DATA AND METHODS

(a) *Dataset*

The dataset includes 12 protein-coding genes on the H-strand of the mitochondrial genome from 14 species, including 12 baleen whales, the sperm whale and the hippopotamus (figure 1). There are 3535 codons or amino acids in the alignment. The data were published and analysed by Sasaki *et al.* (2005), where the GenBank accession numbers for the sequences can be found. The authors used the maximum-likelihood (ML) method under various substitution models to infer the phylogenetic relationships among the baleen whales, with the non-parametric bootstrap (Felsenstein 1985) used to assess confidence.

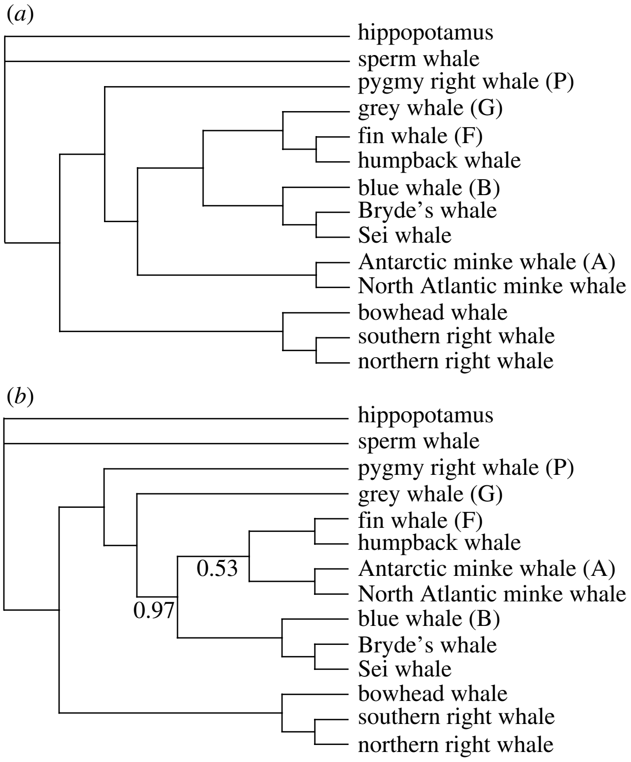

Figure 1. The MAP trees of 14 species under (*a*) the nucleotide model HKY+Γ_5 and (*b*) the amino acid model MTMAM+Γ_5. All clades in the two trees except for two in the tree of (*b*) have PP=1, and only those two are shown. In the text, a small dataset consisting of five species is analysed as well: Antarctic minke whale (A); fin whale (F); blue whale (B); grey whale (G); and pygmy right whale (P) used as the outgroup. The two trees shown in (*a*,*b*) are tree 5 (A((FG)B)) and tree 13 (G((FA)B)) in table 4.

Based on the ML analysis of Sasaki *et al.* (2005) and a pilot Bayesian analysis, it was noted that some clades are well supported by the data, while most phylogenetic uncertainties concern the relationships among the following four groups: (i) minke whales, (ii) the clade of fin whale and humpback whale, (iii) the clade of blue whale, Bryde's whale and Sei whale, and (iv) grey whale. These four clades are identified as lineages I, II, III and IV by Sasaki *et al.* (2005). Thus, I also analyse a smaller dataset consisting of only five species, with one representative species from each of the four clades and the pygmy right whale used as the outgroup. The four representative species are as follows: (i) Antarctic minke whale (A), (ii) fin whale (F), (iii) blue whale (B), and (iv) grey whale (G). The two datasets are referred to as the large and small datasets.

(b) *Phylogenetic analysis*

The data were analysed as either DNA sequences under nucleotide and codon substitution models or translated protein sequences under amino acid substitution models. For a comparison with the Bayesian analysis, the data were also analysed using ML, with the RELL approximate bootstrap method (Kishino & Hasegawa 1989) used to compare the 15 possible trees relating the four clades. The BaseML and CodeML programs in the PAML package (Yang 2007*b*) were used, as some of the models used in this study were not available in other faster likelihood programs such as PhyML (Guindon & Gascuel 2003) or RaxML (Stamatakis *et al.* 2005). For the large dataset, the tree search was thus not exhaustive. The nucleotide substitution models used included JC (Jukes & Cantor 1969), HKY+Γ_5 (Hasegawa *et al.* 1985; Yang 1994) and HKY+C+Γ_5 (Yang 1995, 1996). The latter model assumed different rates, different transition/transversion rate ratios and different base compositions at the three codon positions. The amino acid sequences were analysed under the mtmam and mtmam+Γ_5 models (Yang 1994; Yang *et al.* 1998). The codon sequences were analysed under the codon model M0 (one-ratio), assuming the same non-synonymous/synonymous rate ratio ω for all sites and branches (Goldman & Yang 1994; Yang *et al.* 1998).

The Bayesian analysis was conducted using MrBayes (mb) v. 3.1.2, under the same models as in the ML method. Under the codon model, the codon frequencies are fixed at their observed values rather than being estimated in the MCMC. By default, the program specifies the same prior for all branch lengths, which is the exponential distribution with the mean being 0.1 changes per site.

To use the data size-dependent prior (Yang & Rannala 2005), mb v. 3.1.2 was modified to implement two independent exponential distributions with means μ_0 and μ_1 for the internal and external branch lengths, respectively. (In this notation, the default prior in mb is $\mu_0=\mu_1=0.1$.) Previously, I modified mb v. 3.0.2 for this analysis (Yang & Rannala 2005), and the same modifications were transferred to mb v. 3.1.2 by Will Fletcher (http://abacus.gene.ucl.ac.uk/software.html). Note that mb v. 3.1.2 made at least two improvements over mb v. 3.0.2, which may be important to this study: correction of the proposal ratio in the Local move (Holder *et al.* 2005) and correction of the handling of a lower bound on branch lengths (see below). The prior mean for the internal branch lengths is specified as $\mu_0=cn^{-\gamma}$, with $(1/2)<\gamma<1$. Smaller values of γ in the range (1/2, 1) lead to a more powerful method, as it produces high PPs for the true bifurcating tree. The value $\gamma=2/3$ (harmonic mean of 1/2 and 1) was used in this paper. The constant c was fixed at $c=0.1$, and the prior mean for external branch lengths at $\mu_1=0.1$. Thus, for the amino acid- and codon-based analysis, $\mu_0=0.1n^{-2/3}=0.1\times3535^{-2/3}=0.00043$, while for the nucleotide models, $\mu_0=0.1n^{-2/3}=0.1\times10\,605^{-2/3}=0.00021$. This prior is referred to also as the ‘2E’ prior for its use of two exponential distributions.

Each Bayesian MCMC analysis was run at least twice to confirm the consistency of the results. For most analyses in this study, 2×10^6 iterations were found to be sufficient to produce reliable results. The PPs of tree topologies were collected.

For comparison, I also applied the polytomy prior of Lewis *et al.* (2005) to the mitochondrial dataset, using the Phycas program (http://hydrodictyon.eeb.uconn.edu/projects/phypy/downloads/). Only the large dataset was analysed under two nucleotide substitution models: JC and HKY+Γ_5. The default polytomy prior with $C=\mathrm{e}=2.718$ was used. This means that a tree with $m-1$ internal nodes is 2.718 times more likely in the prior than a tree with m internal modes, with a fairly strong preference for multifurcating trees to binary trees.

3. RESULTS

(a) *Likelihood analysis*

The BaseML and CodeML programs in the PAML package were used to calculate the log-likelihood values for 15 fixed tree topologies. For the small dataset, this amounted to an exhaustive tree search. For the large dataset, only the 15 trees concerning the relationships among the four clades represented by A, F, B and G were evaluated. The RELL approximate bootstrap method was used to calculate the bootstrap proportions (BPs) for the 15 trees. The log-likelihood values and the BPs are shown in tables 1 and 2 for the small and large datasets, respectively.

Model complexity has far greater impact on the model's fit to data than has the tree topology, as judged by the log-likelihood values, and improving the model's fit by adding more parameters generally reduces the log-likelihood differences between trees. These appear to be common features in molecular phylogenetic analysis (Yang *et al.* 1994). For example, HKY+Γ_5 has five more parameters (κ, α and three base frequencies) than JC, and the log-likelihood difference between the two models is above 3000 in the small dataset and above 8000 in the large dataset. Similarly, HKY+C+Γ_5 has eight more parameters (two extra sets of κ and base frequencies) than HKY+Γ_5, and the log-likelihood difference between them is above 2000 (or above 3000) in the small (or large) dataset. Under the same model, the log-likelihood differences between the best and worst trees in the small dataset are 123, 9.4 and 26 under JC69, HKY+Γ_5 and HKY+C+Γ_5, respectively, and are 119, 24 and 14 in the large dataset. The log-likelihood differences between the best (ML) and second best trees are much smaller.

For the small dataset, the ML tree is tree 1 (A((FB)G)) under JC and HKY+Γ_5, but tree 2 (G((FB)A)) is very slightly preferred under HKY+C+Γ_5. Under the codon model M0 (one-ratio) and the two amino acid models mtmam and mtmam+Γ_5, tree 2 is the ML tree, although tree 1 is nearly equally good under mtmam+Γ_5.

The large dataset produced more variable results among models, with almost every model producing a different ML tree. Similar sensitivity to model assumptions was noted in the analysis of Sasaki *et al.* (2005), who commented that almost every possible rearrangement concerning lineages I, II, III and IV has been suggested. Under the nucleotide model JC and

Table 1. Log-likelihood values and RELL BPs (%) for trees for the small dataset. The number of parameters excluding branch lengths is shown in parentheses. The log-likelihood values for the ML tree are shown, while those for other trees are shown as a difference from that for the ML tree. The unrooted trees including the outgroup pygmy right whale are used in the likelihood calculation while they are shown as rooted trees for A, B, F and G. The BPs are approximated using the RELL method, evaluating all 15 trees. The MLEs for the substitution parameters are not shown. The log-likelihood values under the nucleotide and codon models are not comparable with those under the amino acid models, as different data are analysed.

	tree	JC (0)		HKY+Γ_5 (5)		HKY+C+Γ_5 (13)		codon.M0 (11)		MTMAM (0)		MTMAM+Γ_5 (1)	
1	(A((FB)G))	−29 522.9	40	−26 154.4	78	−0.7	34	−0.2	41	−7.5	16	−0.1	26
2	(G((FB)A))	−47.5	2	−3.7	4	−23 935.3	35	−23 643.2	41	−12 245.4	56	−12 166.7	30
3	((AG)(FB))	−81.4		−3.7	5	−7.3	1	−15.5	1	−13.0	3	−1.7	13
4	(A(F(BG)))	−2.6	34	−5.0	7	−14.0	2	−22.2	5	−18.8	9	−6.7	14
5	(A((FG)B))	−15.2	12	−5.2	2	−19.1		−36.1		−38.7		−16.6	
6	((BG)(FA))	−77.5		−9.0	1	−11.1	4	−31.7	1	−24.8	2	−8.7	9
7	((FG)(BA))	−123.2		−9.0	1	−26.3		−57.0		−50.3		−22.0	
8	(F((BA)G))	−90.5		−9.0		−24.5		−48.2		−42.1		−20.7	
9	(G((BA)F))	−85.5		−9.0	1	−18.2		−32.4		−26.3		−16.3	
10	(F((BG)A))	−55.4	1	−9.1		−19.4		−37.2		−25.6	3	−10.9	2
11	(B((FA)G))	−55.1	2	−9.2		−10.3	8	−34.0	1	−31.1	1	−15.0	
12	(B(F(AG)))	−64.7	1	−9.2		−19.9		−41.8		−32.6	1	−14.9	
13	(G((FA)B))	−47.6	5	−9.2		−7.2	14	−18.9	10	−15.4	9	−10.7	4
14	(F(B(AG)))	−62.7	2	−9.3		−22.5		−40.9		−31.7	1	−14.5	1
15	(B((FG)A))	−72.3		−9.4		−22.3		−53.0		−46.4		−21.4	

Table 2. Log-likelihood values and RELL BPs (%) for 15 trees for the large dataset. The data of 14 species are analysed to calculate the log-likelihood values for 15 unrooted trees. The trees differ in the relationships among A, B, F and G, but are otherwise identical, as shown in figure 1. See also legend to table 1.

	tree	JC		HKY+Γ_5		HKY+C+Γ_5		M0		MTMAM		MTMAM+Γ_5	
1	(A((FB)G))	−56 845.1	52	−6.0	12	−6.0	3	−3.1	13	−14.8		−8.0	
2	(G((FB)A))	−42.0	6	−20.2		−5.8	3	−4.6	14	−18 022.4	28	−0.3	19
3	((AG)(FB))	−73.3		−19.8		−10.6	1	−13.5	1	−9.3	7	−3.5	11
4	(A(F(BG)))	−21.3	6	−7.5	1	−4.2	12	−44 517.0	32	−14.5	2	−6.0	3
5	(A((FG)B))	−6.1	34	−48 504.6	84	−2.0	23	−4.2	12	−27.5		−14.5	
6	((BG)(FA))	−103.1		−23.7		−9.7	1	−17.0		−12.5	5	−3.6	11
7	((FG)(BA))	−93.1		−14.7	1	−10.0	1	−24.0		−23.4		−12.1	
8	(F((BA)G))	−112.6		−20.9		−13.0		−19.1		−16.5	1	−8.8	1
9	(G((BA)F))	−58.2	2	−20.6		−3.0	15	−6.2	13	−1.4	21	−1.6	13
10	(F((BG)A))	−91.2		−23.7		−14.3		−13.0	2	−10.9	11	−4.2	10
11	(B((FA)G))	−118.8		−20.5		−8.6	2	−23.1		−21.6		−8.9	1
12	(B(F(AG)))	−94.8		−19.4	1	−9.9	3	−17.0	2	−17.7	1	−7.5	3
13	(G((FA)B))	−66.3		−23.2		−45 410.5	33	−7.9	7	−1.6	18	−17 564.7	20
14	(F(B(AG)))	−95.8		−22.4		−11.3	2	−12.8	5	−12.9	6	−5.3	8
15	(B((FG)A))	−93.0		−13.8	2	−10.4	1	−21.7		−28.7		−13.7	

the amino acid model MTMAM, the ML trees are consistent between the small and large datasets, while, for all other models, they are incompatible, indicating that species sampling has considerable effect on phylogeny reconstruction.

In sum mary, the ML trees are different under different substitution models, between the analyses of DNA and protein sequences and also between the small and large datasets. Nevertheless, the RELL bootstrap support values are not high (almost all of them being less than 0.8), so the differences in the ML trees in the different analyses may be explained by sampling errors or lack of resolution in the data. The K–H and S–H tests (Kishino & Hasegawa 1989; Shimodaira & Hasegawa 1999; see Goldman *et al.* (2000) for a review) were also used to compare trees, and the ML tree was not significantly supported under any of the models/analyses by any of the tests, consistent with the results from the RELL bootstrap analysis.

(b) *Bayesian analysis*

MRBAYES v. 3.1.2 (both the standard and modified versions) were used to calculate PPs for trees under the nucleotide, codon and amino acid substitution models.

The results for the small dataset are shown in table 3. Tree 1 (A((FB)G)) is the ML tree and also the MAP tree under the standard prior with one exponential for all branch lengths ($\mu_0 = \mu_1 = 0.1$, column '1E' in table 3) under the nucleotide models JC, HKY+Γ_5 and HKY+C+Γ_5. However, the PPs for the MAP tree are much higher than the BPs in the corresponding ML analysis, with PP=0.93, 0.98 and 1.00 under the three models compared with BP=0.40, 0.78 and 0.34 (table 1). Under the codon model M0 (one-ratio), trees 1 and 2 have nearly identical log-likelihood values (table 1) and PP values (table 3). Under the amino acid model MTMAM, tree 2 is the best tree by both ML/BP and Bayesian inference (BI)/PP, with BP=0.56 and PP=1.00. In summary, in the small dataset, ML and

Table 3. Bayesian posterior tree probabilities (%) for the small dataset. (1E means the default prior in mb ($\mu_0=\mu_1=0.1$), while 2E means the data size-dependent prior, with two exponential priors for the internal and external branch lengths ($\mu_0 \neq \mu_1$).)

		JC		HKY+Γ_5		HKY+C+Γ_5		codon.M0		MTMAM		MTMAM+Γ_5	
	tree	1E	2E	1E	2E	1E	2E	1E	2E	1E	2E	1E	2E
1	(A((FB)G))	93	69	98	19	100	37	49	5		2	34	49
2	(G((FB)A))			1	13		20	51	95	100	98	60	18
3	((AG)(FB))			1	15		23					6	33
4	(A(F(BG)))	7	31		6		2						
5	(A((FG)B))				5		2						
6	((BG)(FA))				4		2						
7	((FG)(BA))				4		1						
8	(F((BA)G))				5		1						
9	(G((BA)F))				4		2						
10	(F((BG)A))				4		1						
11	(B((FA)G))				4		2						
12	(B(F(AG)))				4		1						
13	(G((FA)B))				4		2						
14	(F(B(AG)))				4		1						
15	(B((FG)A))				4		1						

BI produced the same best tree under all models examined, although PP is much higher than BP. The very high PPs are not tenable biologically, as the MAP trees cannot all be correct.

The use of the data size-dependent prior (column 2E in table 3) was effective in reducing the PP values for almost all models in the small dataset. For example, PP=0.93, 0.98 and 1.00 under the three nucleotide models for the standard prior, and are 0.69, 0.19 and 0.37 for the 2E prior. However, the opposite result was observed under the codon model M0. Here, PP=0.49 and 0.51 for trees 1 and 2 for the standard prior, and become 0.05 and 0.95 for the 2E prior, which are even more extreme. Under the two amino acid models (MTMAM and MTMAM+Γ_5), the PP values are much less extreme for the 2E prior. Under MTMAM+Γ_5, the MAP tree was tree 2 for the standard prior but tree 1 for the 2E prior. Thus, the relationship between the prior mean μ_0 and the posterior tree probabilities is not a simple monotonic one.

The results for the large dataset are shown in table 4. Under the standard prior, the PPs for the MAP trees are very high, although the MAP trees are different for different models. For example, the MAP tree is tree 1 under JC and tree 5 under HKY+Γ_5 and HKY+C+Γ_5, but in each case PP=1.00. The codon model M0 and amino acid models MTMAM and MTMAM+Γ_5 produced yet different MAP trees, but with weaker support (PP$\leq$0.93). Overall, the Bayesian analysis produced more MAP trees in the large than the small datasets, indicating greater sensitivity to model assumptions in the large dataset. Note that the Bayesian analysis was conducted without any constraint on the tree topology, but only the trees listed in table 4 had non-negligible PPs.

The results under the 2E prior for the large dataset are listed in table 4 (columns headed 2E). Under the JC model, the 2E prior produced the same MAP tree as the standard prior, but with reduced support (PP=0.99 compared with 1.00). Under the HKY+Γ_5 and HKY+C+Γ_5 models, the 2E prior attached substantial probabilities for trees not listed in table 4. The placement of the sperm whale was unstable. The MAP tree grouped the sperm whale with Balaenidae (the clade of the bowhead whale and the right whales), with PP=0.63 under HKY+Γ_5 and 0.60 under HKY+C+Γ_5. The next tree groups the sperm whale with the clade of Neobalaenidae (pygmy right whale), Eschrichtiidae (grey whale) and Balaenopteridae, with PP=0.32 under HKY+Γ_5 and 0.37 under HKY+C+Γ_5. The third tree is the one shown in figure 1*a*. The relationships among A, F, B and G are (((FG)B)A) in all those three best trees, so that PP=1.0 for the clade (((FG)B)A). Under the codon model M0, the 2E prior produced the same MAP tree as the standard prior, but with slightly higher support (PP=0.97 compared with 0.93), so the effect is opposite to expectation. Under the amino acid substitution model MTMAM, the results are almost identical between the two priors. Under MTMAM+Γ_5, trees not listed in table 4 received substantial PPs. The MAP tree has PP=0.33, and groups the sperm whale with Balaenidae (the clade of the bowhead whale and the right whales) and favours tree 13 (G(FA)B) concerning the relationship among A, F, B and G.

In summary, in the small dataset, the 2E prior was very effective in reducing the high PPs for trees and the overconfidence of BI under the standard prior. The 2E prior also reduced the PPs for the MAP tree in the large dataset. However, the alternative trees that attracted some probabilities in the large dataset are apparently wrong. The fact that the 2E prior can lead to changes of the order of the binary trees may be an undesirable feature of the prior.

The polytomy prior (Lewis *et al.* 2005) is applied to the large dataset, using the PHYCAS program. Under the nucleotide substitution model JC, the MAP tree is tree 1, with PP=1.00, compared with PP=1.00 from the standard prior in mb and 0.99 from the 2E prior (table 4). Under HKY+Γ_5, the MAP tree is tree 5, with PP=0.94, compared with PP=1.00 from the

Table 4. Bayesian posterior tree probabilities (%) for the large dataset.

		JC		HKY+Γ_5		HKY+C+Γ_5		codon.M0		MTMAM		MTMAM+Γ_5	
τ	tree	1E	2E	1E	2E[a]	1E	2E[a]	1E	2E	1E	2E	1E	2E[a]
1	(A((FB)G))	100	99					5	1				
2	(G((FB)A))									72	72	36	23+11+1
3	((AG)(FB))											1	1+1
4	(A(F(BG)))							93	97				
5	(A((FG)B))		1	100	63+32+4	100	60+37+3	2	3				
6	((BG)(FA))											1	1+1
7	((FG)(BA))												
8	(F((BA)G))												
9	(G((BA)F))									14	17	8	7+4
10	(F((BG)A))												1
11	(B((FA)G))												
12	(B(F(AG)))												
13	(G((FA)B))									14	11	52	33+16+1
14	(F(B(AG)))												
15	(B((FG)A))												

[a] Under the 2E prior and models HKY+Γ_5, HKY+C+Γ_5 and MTMAM+Γ_5, the PPs shown are the sums of several tree topologies. See text for details, and also legend to table 3.

standard prior in mb and 0.63 from the 2E prior (table 4). Thus, the polytomy prior also ameliorates the problem of high PPs to some extent.

4. DISCUSSION

(a) *The usefulness of the data size-dependent prior in reducing apparent conflicts in Bayesian phylogenetic analysis*

In both the small and large datasets, the data size-dependent prior almost always led to reduced PPs for the MAP trees. It thus appears useful in reducing the apparent conflicts in Bayesian phylogenetic analysis. The implementation here involves some arbitrariness. The theoretical prediction (Yang 2007*a*) is that one should have $(1/2)<\gamma<1$ if the mean in the exponential prior for the internal branch lengths is $\mu_0=cn^{-\gamma}$. However, the constant c is arbitrary and is fixed at 0.1 in this study. Furthermore, the theory was based on the analysis of the simple case of estimating the rooted trees for three species under the molecular clock, and no proof yet exists that the same prediction should apply to the general case of estimating unrooted trees for many species, even though the structures of the problems appear similar. More tests using both real and simulated datasets may be necessary to confirm the usefulness of the prior in resolving the star-tree paradox and in reducing the overconfidence of BI of molecular phylogenies.

A technical issue in the current implementation of MrBayes may affect the use of the data size-dependent prior. Both versions 3.0.2 and 3.1.2 truncate branch lengths to the interval $(10^{-6}, 100)$. The exponential priors studied in this paper are truncated from both ends. In the likelihood calculation, one needs to constrain branch lengths away from 0 to avoid zero probability of observing the data (as it occurs if there is a difference between two sequences but their distance is 0). For analysis under the standard prior, this constraint may not cause any problem since the branch lengths are typically far away from the lower bound. However, for the data size-dependent prior explored in this study, the constraint may affect the results if the prior mean μ_0 is very small. In theory, it should be sufficient to apply the lower bound to the external branch lengths only, which will ensure a strictly positive distance between any two sequences.

(b) *Possible approaches to reduce spuriously high PPs for trees*

The concern about BI of phylogenetic trees is not so much that different MAP trees are produced depending on the assumed substitution model, the sampled species or the data type, but that the PPs for the MAP trees are often very high. As there is only one true tree underlying the data, not all the different MAP trees can be correct, so that one cannot escape the conclusion that the high PPs attached on the wrong trees are spurious. It should be stressed that the problem of high PPs for trees discussed in this study (and in Yang & Rannala 2005; Yang 2007*a*) is not due to implementation in the MrBayes program but rather reflects the statistical properties of the Bayesian method. Bayesian model selection, of which phylogeny reconstruction is a special case, is a controversial area. In the case of comparing two simple models, theory predicts that the model closer to the truth will dominate when the sample size increases (Dawid 1999), but in finite datasets, the Bayesian method may quite often attach high PPs to the wrong model.

In the analysis of real sequence data, misspecification of the substitution (likelihood) model is a serious concern. Models are simplistic descriptions of reality and will never (and are not supposed to) catch all the complexities and nuances of the real biological process. This study has used a number of substitution models, some being the most sophisticated currently available. It is possible that the PPs for trees may become more moderate if the assumed model is made even more complex. In particular, by ignoring interactions and dependencies among sites in the sequence, the widely used i.i.d. models, which assume that the data at different sites have independent and identical distributions, may exaggerate the information content in the data, thus producing spuriously high PPs. For example, the paired sites in the stem regions of an RNA molecule do not evolve independently, and they may contain only about half as much information as if the sites are independent. However, this effect appears to be too small, even if the 'effective sequence length' is only one-tenth the real sequence length, i.e. even if our dataset is as informative as one of 10 per cent its size but with independent sites. In the star-tree simulations (Yang & Rannala 2005), high PPs occur commonly with only 200 or 1000 sites, while, nowadays, real datasets of 10 Kb or even 10 Mb are routinely analysed. We have too much data. It should also be noted that increasing model complexity may both decrease and increase posterior tree probabilities (tables 3 and 4).

Thus, I suggest that model improvement may not be the ultimate solution to the problem of high posterior tree probabilities. First, one has to analyse the data using existing models and methods, and it is impractical to claim that existing models are unrealistic and cannot be used. Second, extreme sensitivity to the assumed model is not a desirable property of any analytical method. The Bayesian MCMC machinery has the power to enable researchers to implement sophisticated multi-parameter models. It is somewhat ironic if the need to implement such models is not that they offer any insights into the biological process but that the Bayesian method is oversensitive to model assumptions. Similarly, model averaging may not be the solution to the problem. If the true model is not included in the set of models that the MCMC is averaging over, the model that is closest to the truth will dominate and the results will not be very different from those obtained using that model alone. For example, given the huge log-likelihood differences between JC and HKY+Γ_5 in the analysis of the mitochondrial data (tables 1 and 2), it seems certain that averaging over JC and HKY+Γ_5 in the MCMC will produce very similar posterior tree probabilities as using HKY+Γ_5 alone.

The problem appears to lie deeper than model violation. The polytomy prior and the data size-dependent prior may both be viewed as extreme measures for a difficult problem. Here, it is interesting to note their similarities and differences. First, the polytomy prior is slightly more complex to implement. As bifurcating and multifurcating trees have different numbers of branch lengths and thus different dimensions, algorithms such as reversible-jump MCMC (Green 1995) are necessary (Lewis *et al.* 2005). Second, both priors may be considered 'non-Bayesian'. The data size-dependent prior has the prior mean dependent on the size of the data. The polytomy prior does not depend on any aspects of the sequence data but assigns positive probabilities to multifurcating trees, which are not biologically meaningful models (see below). Third, asymptotic theory of Bayesian model selection predicts that the polytomy prior resolves the star-tree paradox, and the theory applies to phylogenies of any size (Dawid 1999; Yang 2007*a*). The performance of the data size-dependent prior in large trees is unknown.

Fourth, the use of the data size-dependent prior may change the order of the PPs for the binary trees, as seen in the analysis of this study. The situation with the polytomy prior is more complex and may depend on the measure of accuracy used. The marginal likelihood for the binary trees (i.e. the integral of the likelihood under the tree over the branch lengths and substitution parameters) is not affected by the introduction of the polytomy prior, nor is the order of the PPs for the binary trees. However, it is not sensible to consider only the binary trees because much of the PP may be attached to the multifurcating trees. Wheeler & Pickett (2008; see also Yang 2006, p. 176) argued that the PPs for clades, unlike PPs for binary trees, are not very meaningful measures of accuracy. Furthermore, it appears generally accepted that polytomies are an intuitive way of representing lack of resolution but do not represent biological truth. Thus, to calculate the PPs for binary trees, one should reapportion the PPs for the multifurcating trees among the compatible binary trees. In the case of four species, the PPs for the four trees (the star tree and three binary trees) $(P_0, P_1, P_2, P_3) = (0.9, 0.06, 0.03, 0.01)$ should be considered equivalent to $(P_1, P_2, P_3) = (0.36, 0.33, 0.31)$ for the three binary trees only. This discussion suggests that two measures may be useful when both binary and multifurcating trees are evaluated in the Bayesian analysis, as under the polytomy prior: (i) the PPs for clades, calculated by summing up PPs for all trees, both binary and multifurcating, that contain the clade, and (ii) the PPs for binary trees, calculated by apportioning the PPs for multifurcating trees among the compatible binary trees. The former is an intuitive measure, useful when the data are not informative and the PPs for binary trees are very low. Under both measures, the order of the PPs for clades or binary trees may change between the standard prior and the polytomy prior or with different Cs in the polytomy prior.

Lastly, the polytomy prior may also be viewed as a prior on the internal branch lengths; that is, the prior on each internal branch length is a mixture of a component at 0 (with probability π_0) and another component from the exponential distribution (with probability $1-\pi_0$). For unrooted trees of four species, the two formulations are equivalent with $\pi_0 = C/(C+3)$. For larger trees, the prior probabilities assigned to the multifurcating trees with different numbers of internal nodes may differ depending on the details of the prior specification. It may be useful to explore other forms of prior on the internal branch lengths.

I thank two anonymous reviewers for their comments and Paul Lewis for his assistance with PHYCAS. This study is supported by a grant from NERC to Z.Y.

REFERENCES

Bourlat, S. J. *et al.* 2006 Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**, 85–88. (doi:10.1038/nature05241)

Buckley, T. R. 2002 Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* **51**, 509–523. (doi:10.1080/10635150290069922)

Cummings, M. P., Handley, S. A., Myers, D. S., Reed, D. L., Rokas, A. & Winka, K. 2003 Comparing bootstrap and posterior probability values in the four-taxon case. *Syst. Biol.* **52**, 477–487. (doi:10.1080/10635150390218213)

Dawid, A. P. 1999 The trouble with Bayes factors. Research report 202, Department of Statistical Science, University College London.

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. & Douzery, E. J. 2003 Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* **20**, 248–254. (doi:10.1093/molbev/msg042)

Erixon, P., Svennblad, B., Britton, T. & Oxelman, B. 2003 Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* **52**, 665–673. (doi:10.1080/10635150390235485)

Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791. (doi:10.2307/2408678)

Goldman, N. & Yang, Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736.

Goldman, N., Anderson, J. P. & Rodrigo, A. G. 2000 Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* **49**, 652–670. (doi:10.1080/106351500750049752)

Green, P. J. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)

Guindon, S. & Gascuel, O. 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704. (doi:10.1080/10635150390235520)

Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174. (doi:10.1007/BF02101694)

Holder, M. T., Lewis, P. O., Swofford, D. L. & Larget, B. 2005 Hastings ratio of the Local proposal used in Bayesian phylogenetics. *Syst. Biol.* **54**, 961–965. (doi:10.1080/10635150500354670)

Huelsenbeck, J. P. & Rannala, B. 2004 Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* **53**, 904–913. (doi:10.1080/10635150490522629)

Huelsenbeck, J. P. & Ronquist, F. 2001 MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)

Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. N. Munro), pp. 21–123. New York, NY: Academic Press.

Kishino, H. & Hasegawa, M. 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29**, 170–179. (doi:10.1007/BF02100115)

Lemmon, A. R. & Moriarty, E. C. 2004 The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* **53**, 265–277. (doi:10.1080/10635150490423520)

Lewis, P. O., Holder, M. T. & Holsinger, K. E. 2005 Polytomies and Bayesian phylogenetic inference. *Syst. Biol.* **54**, 241–253. (doi:10.1080/10635150590924208)

Li, S., Pearl, D. & Doss, H. 2000 Phylogenetic tree reconstruction using Markov chain Monte Carlo. *J. Am. Stat. Assoc.* **95**, 493–508. (doi:10.2307/2669394)

Lindley, D. V. 1957 A statistical paradox. *Biometrika* **44**, 187–192.

Mau, B. & Newton, M. A. 1997 Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* **6**, 122–131. (doi:10.2307/1390728)

O'Hagan, A. & Forster, J. 2004 *Kendall's advanced theory of statistics: Bayesian inference*. London, UK: Arnold.

Rannala, B. & Yang, Z. 1996 Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311. (doi:10.1007/BF02338839)

Rannala, B. & Yang, Z. 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656.

Rannala, B. & Yang, Z. 2008 Phylogenetic inference using whole genomes. *Annu. Rev. Genom. Hum. Genet.* **9**, 217–231 (doi:10.1146/annurev.genom.9.081307.164407)

Ronquist, F. & Huelsenbeck, J. P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)

Sasaki, T. *et al.* 2005 Mitochondrial phylogenetics and evolution of mysticete whales. *Syst. Biol.* **54**, 77–90. (doi:10.1080/10635150590905939)

Shimodaira, H. & Hasegawa, M. 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116.

Simmons, M. P., Pickett, K. M. & Miya, M. 2004 How meaningful are Bayesian support values? *Mol. Biol. Evol.* **21**, 188–199. (doi:10.1093/molbev/msh014)

Stamatakis, A., Ludwig, T. & Meier, H. 2005 RaxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463. (doi:10.1093/bioinformatics/bti191)

Steel, M. & Matsen, F. A. 2007 The Bayesian "star paradox" persists for long finite sequences. *Mol. Biol. Evol.* **24**, 1075–1079. (doi:10.1093/molbev/msm028)

Susko, E. 2008 On the distributions of bootstrap support and posterior distributions for a star tree. *Syst. Biol.* **57**, 602–612. (doi:10.1080/10635150802302468)

Suzuki, Y., Glazko, G. V. & Nei, M. 2002 Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Natl Acad. Sci. USA* **99**, 16 138–16 143. (doi:10.1073/pnas.212646199)

Takahata, N. 1989 Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122**, 957–966.

Wheeler, W. C. & Pickett, K. M. 2008 Topology-Bayes versus clade-Bayes in phylogenetic analysis. *Mol. Biol. Evol.* **25**, 447–453. (doi:10.1093/molbev/msm274)

Yang, Z. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. (doi:10.1007/BF00160154)

Yang, Z. 1995 A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005.

Yang, Z. 1996 Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587–596. (doi:10.1007/BF02352289)

Yang, Z. 2006 *Computational molecular evolution*. Oxford, UK: Oxford University Press.

Yang, Z. 2007*a* Fair-balance paradox, star-tree paradox and Bayesian phylogenetics. *Mol. Biol. Evol.* **24**, 1639–1655. (doi:10.1093/molbev/msm081)

Yang, Z. 2007*b* PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)

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

Yang, Z. & Rannala, B. 2005 Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* **54**, 455–470. (doi:10.1080/10635150590945313)

Yang, Z., Goldman, N. & Friday, A. 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**, 316–324.

Yang, Z., Nielsen, R. & Hasegawa, M. 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611.