PhyloBayes MPI. Supplementary information

Nicolas Lartillot, Nicolas Rodrigue, Daniel Stubbs, Jacques Richer.

Centre Robert-Cedergren pour la Bioinformatique, Département de Biochimie, Université de Montréal, Québec, Canada;

nicolas.lartillot@umontreal.ca

(+1) 514 343 6111 (2721)

1 Methods

1.1 Model

The data consist of a matrix of characters $D = (D_{ij})$, for $i = 1..N$ aligned positions, and $j = 1..P$ taxa, each non missing cell D_{ij} being in one among S possible states ($S = 4$ in the case of nucleotides, 20 in the case of amino-acids). Sites of the alignment are assumed to be independent and identically distributed (i.i.d.) from a Dirichlet process mixture of substitution processes running along a phylogenetic tree τ. In addition, a discretized gamma distribution is assumed for modeling among-site rate variation (Yang, 1994).

All substitution processes considered here are time-reversible. The pulley-principle therefore applies (Felsenstein, 1981) and trees are unrooted. A uniform prior over all possible bifurcating tree topologies is assumed and, conditional on the topology, branch lengths are i.i.d from an exponential of mean μ. The hyperparameter μ is itself endowed with an exponential prior of mean 0.1. The discretized gamma distribution of rates across sites is parameterized by a shape parameter α,

endowed with an exponential prior of mean 1.

Under the most general model configuration (CAT-GTR), all sites share a same set of exchangeability parameters between pairs of states $r = (r_{ab})_{1 \leq a < b < S}$, which are i.i.d. from an exponential of mean 1. Each component of the mixture $k > 0$ has its own equilbirium frequency profile $\pi_k = (\pi_{ka})_{a=1..S}$, such that the substitution rate matrix for component k is

$$Q^k_{ab} = r_{ab}\,\pi_{kb},$$

where we use the convention that $r_{ab} = r_{ba}$ whenever $a > b$, thus ensuring time-reversibility of the process.

A Dirichlet process mixture over equilibirium frequency profiles can be seen as an infinite mixture $(V_k, w_k, \pi_k)_{k \geq 0}$, where (Papaspiliopoulos and Roberts, 2008):

$$\begin{aligned} V_k &\sim Beta(1, \kappa), \\ w_k &= \prod_{l<k} (1 - V_l)\, V_k, \\ \pi_k &\sim Dirichlet(\nu). \end{aligned}$$

Here, $\kappa > 0$ is the concentration parameter of the Dirichlet process, and ν is a S-dimensional vector of hyperparameters $\nu_a > 0$, $a = 1..S$. An exponential prior of mean 10 is defined for κ, and a product of exponential priors of mean 1 for the $\nu_a > 0$, $a = 1..S$. In older versions (up to version 1.3), the prior on ν was truncated so that $\sum_a \nu_a > S/5$ (for numerical stability reasons). Thanks to improved numerical matrix diagonalization, this constraint has been removed starting from version 1.4.

Denoting, for $i = 1..N$, the allocation status of position i as c_i, then $c_i = k$ with prior probability w_k. Finally, conditional on c_i, the equilibrium frequencies of the amino-acid replacement process

at site i are described by π_{c_i}.

The weigths w_k are decreasing geometrically in expectation (as ρ^k, where $\rho = \kappa/(1+\kappa)$), which implies that the weights in the right tail of the infinite series $(w_k)_{k>0}$, as well as their sum from k to ∞, are rapidly decreasing and converging toward 0 as k increases. This suggests that the mixture can be truncated by letting $V_{K_{max}} = 1$, thus setting $w_k = 0$ for $k > K_{max}$. The resulting finite model is similar to previously proposed truncated Dirichlet processes (Ishwaran and James, 2001). Here we choose $K_{max} = 5000$.

1.2 Gibbs sampling for subtree pruning and regrafting (SPR)

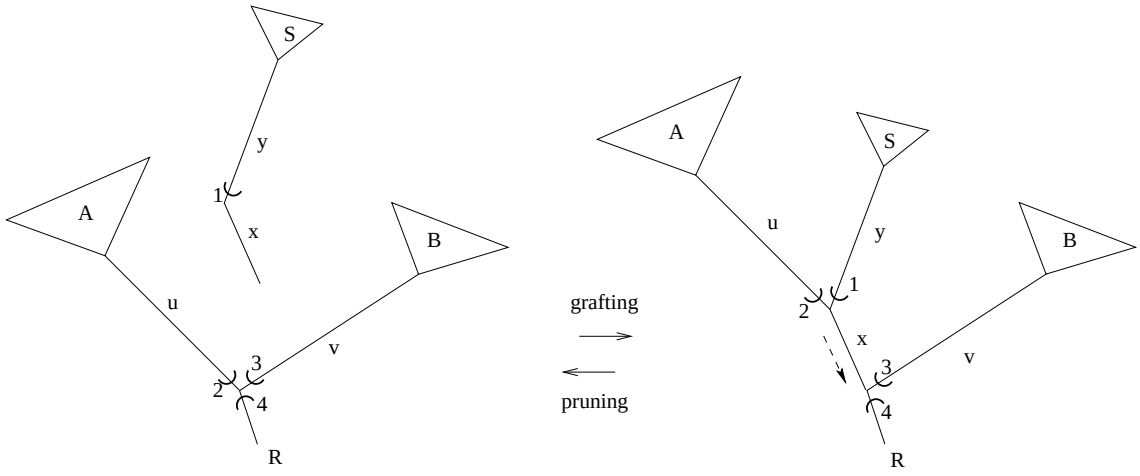


Figure S1. Pruning and regrafting subtrees (see text for details)

The Gibbs sampling algorithm proceeds as follows:

1. Choose an internal node uniformly at random and root the tree at that node.

2. Choose an internal node other than the root and its immediate descendants. Prune the pending subtree, as in figure S1 (pruning direction), taking away the stem branch (y on figure S1) as well as the branch upstream from y (x on figure S1), from the main tree, and leaving all branch lengths unchanged.

3. Update all conditional likelihood vectors around each node of the tree, as indicated on figure S1 on the left. In figure S1, cups represent the conditional probability of the data spanned by the subtree to the open side of the cup. For instance, cup 1 represents the vector of conditional probabilities of the sequence data of group S, given the state at the node linking x and y. Cups 2, 3 and 4 (on the left) are the conditional probabilities for data of groups A, B and R, respectively, given the state at the node linking branches u and v (see also Guindon and Gascuel, 2003; Hordijk and Gascuel, 2005).

4. Recursively scan all possible regrafting of the subtree on the main tree, each time rearranging branches as indicated on figure S1 (grafting direction). For each regrafting position, use the locally cached conditional likelihoods to compute the likelihood of the tree resulting from the regrafting. On figure S1 (right hand side), this requires multiplying conditional likelihood vectors 1 and 2, propagating their product along branch x (dashed arrow), mulitplying the result with conditional likelihood vectors 3 and 4, and with the (site-specific) equilibrium frequencies, then summing over all states (and over all rates of the discretized gamma distribution) at each site. Conditional likelihood vectors 1, 2, 3 and 4 were already updated at step 3.

5. Choose among all possible regrafting points proportionally to their relative posterior probabilities. In the present case, these posterior probabilities are proportional to the relative likelihoods of each candidate tree. This is because the prior is uniform over all possible tree topologies, and because there is a one-to-one mapping of branch lengths across all of the candidate trees. Since the prior over branch lengths is i.i.d. across branches, the prior density is the same for all possible regrafting points.

The entire scan represents the equivalent of less than 3 likelihood computations over the entire

tree: one pre-order and one post-order traversal of the main tree to update all of the conditional likelihoods around each node, one post-order traversal of the subtree for updating the basal conditional likelihood (cup 1), and one traversal of the main tree for testing all possible regraftings.

In a parallel framework, where each slave is in charge of a specific segment of the complete sequence alignment, the master randomly chooses the root and the subtree to be pruned (step 1 and 2) and sends this information to all slaves, which then reroot the tree and prune it accordingly. The update of the conditional likelihood vectors (step 3) and the complete scan of all possible regrafting points (step 4) is done by each slave, after which each slave sends back to the master an array of log likelihoods, containing one log likelihood (one single real number) for each regrafting point. The master collects the arrays, sums them up over all slaves for each regrafting position and, finally, chooses a regrafting position based on the Gibbs-sampling decision rule (step 5). The frequency of communication between master and slaves, and the amount of information passing through the communication channel between master and slaves, is thereby minimized.

1.3 Gibbs-Metropolis over the truncated stick-breaking prior

Classical MCMC sampling methods for truncated Dirichlet processes (Ishwaran and James, 2001) alternate between Gibbs sampling over the allocations c_i, and Metropolis Hastings updates of the mixture variables (here the π_k) and the hyperparameters κ and ν. For large mixtures, however, this results in potentially long Gibbs sampling cycles, as each site $i = 1..N$ of the alignment has to be tentatively allocated to all possible components, and for each possible reallocation $k = 1..K_{max}$, the site-specific likelihood $p(X_i \mid \pi)$ has to be recomputed with $\pi = \pi_k$. In the present case, data-augmentation leads to simple and rapidly evaluated site-specific augmented likelihoods, which take

the following form:

$$p(X_i \mid \pi) \quad \propto \quad \prod_{a=1}^{S} \pi_a^{u_{ia}} \, e^{-x_{ia}\pi_a},$$

where, u_{ia} and x_{ia} are integral and real sufficient statistics computed based on the complete substitution history at site i (Lartillot, 2006). Yet, computing these likelihoods for all sites and for the whole mixture can quickly become a limiting factor, and furthermore, is most probably a waste of time for components that have negligibly small weights.

We therefore developed an alternative sampling method, which is a hybrid between Gibbs-sampling and Metropolis-Hastings and which was inspired by (albeit distinct from) Papaspiliopoulos and Roberts (2008). First, a threshold $K_0 \leq K_{max}$ is specified, such that the total weight of all components above K_0 is, in expectation, of the order of a pre-defined tuning parameter $\epsilon << 1$. Then, for each site $i = 1..N$, we propose a reallocation with probabilities determined by the *posterior* weights (as in the classical Gibbs sampling algorithm) for $k \leq K_0$, and proportional to the *prior* weights for $k > K_0$. In this way, we avoid recomputing the allocation-specific site-likelihoods for the right tail of the mixture ($k > K_0$), where the probability of accepting an allocation is effectively limited by the small weights. In a second step, this proposal has to be accepted or rejected according to a Metropolis-Hastings rule, so as to guarantee that the sampler leaves the posterior distribution invariant.

Specifically, for a given site i, with current allocation $c_i = k_1$, define

$$\begin{aligned} p_{ik} \quad &\propto \quad w_k \, p(X_i \mid \pi_k) \qquad \text{if} \quad k \leq K_0, \\ p_{ik} \quad &\propto \quad w_k \, M \qquad\qquad\quad \text{if} \quad k > K_0, \end{aligned}$$

where $M = \max_{k=1..K_0} p(X_i \mid \pi_k)$ and the p_{ik} are normalized, so that $\sum_{k=1..K} p_{ik} = 1$. Then, propose $c_i = k$ with probability p_{ik}. Denote the chosen value by k_2, and accept the move with

probability $\min(1, R_2/R_1)$ where, for $j = 1, 2$, $R_j = 1$ if $k_j \leq K_0$ and $R_j = p(X_i \mid \pi_{k_j})/M$ if $k_j > K_0$. It can be checked that this update rule verifies the detailed balance by separately considering each of the four sub-cases.

As for the tuning, we use $\epsilon = 0.001$, and set K_0 equal to the integer part of $\ln \epsilon / \ln \rho$, where, as above, $\rho = \kappa/(1+\kappa)$. As an example, for a dataset of 30 000 positions, 50 species, at equilibrium, κ is of the order of 200, so that K_0 is of the order of 1500 to 2000. Since, at equilibrium, the rank of the last occupied component of the mixture rarely exceeds 1500, the sampler has mixing properties almost indistinguishable from the exhaustive Gibbs sampler, while substantially reducing the number of likelihood evaluations (by about a factor 5).

Under the stick-breaking representation, the parallelization of the computation is straightforward. For updating allocations, each site $i = 1..N$ can independently perform the algorithm mentioned above independently of all other sites. Similarly, profiles associated to each component of the mixture can be updated independently of each other, conditional on the current allocation vector. Finally, the weights of the mixture need to be resampled conditional on the current allocation, which can be done by Gibbs sampling (Papaspiliopoulos and Roberts, 2008):

$$\begin{aligned} V_k &\sim Beta(1 + m_k, \kappa + M_k), \\ w_k &= \prod_{l<k} (1 - V_l)\, V_k, \end{aligned}$$

where m_k is the number of sites allocated to component k, and $M_k = \sum_{l=k+1}^{K_{max}} m_l$.

Altogether, the overall MCMC schedule for updating the mixture cycles over the following updates:

- the weights of the mixture are resampled conditional on the current allocation vector c and on κ, and are broadcast to all slaves;

- conditional on the weights just received, on the current equilibrium frequency profiles $(\pi_k)_{k=1..K_{max}}$ and on the site-specific sufficient statistics, each slave performs the Gibbs/Metropolis algorithm introduced above for all sites under its charge;
- the new site allocations are collected by the master and broadcast to all slaves;
- non-empty components are equally distributed among slaves, and each slave performs a series of Metropolis updates of the equilibrium frequency profiles of these components (conditional on the new allocations), while the empty components of the mixture are resampled by the master from the prior.
- new profiles of non-empty components are collected by the master, and all new profiles are broadcast to all slaves (in preparation for the next cycle).

In practice, the entire series is cycled over 5 times, before moving on to other types of update mechanisms. The latter consist of label switching moves (Papaspiliopoulos and Roberts, 2008), which are an important ingredient for proper mixing under the stick-breaking prior, followed by updates of the relative exchangeabilities and updates of the hyperparameters κ and α.

Validation and benchmarking

A series of 8 datasets were gathered from previously published phylogenetic analyses and were used for conducting a detailed comparison between the old (serial) and the new (parallel) implementations under equivalent models and priors. Specifically, we used 3 alignments obtained from TreeBase (Sanderson et al., 1994), with reference numbers M1382 (9 taxa, 1560 sites), M1487 (52 taxa, 981 sites), M2477 (39 taxa, 888 sites), three phylogenomic datasets at the level of chordates (Delsuc et al., 2006), Algae (Rodríguez-Ezpeleta et al., 2007) and Bilateria (Lartillot and

Philippe, 2008) and, finally, two datasets kindly provided by Frédéric Delsuc, reproducing the concatenations of nuclear and mitochondrial genes in 42 mammalian taxa of Springer et al. (2003), with 4768 and 3507 aligned amino-acid positions. In the case of the three phylogenomic datasets, a random subset of 20 genes from the original concatenations were uniformly sampled, leading to three concatenations of 5197, 4743 and 4431 aligned positions, respectively.

For each dataset, the old and new implementations were run under the CAT-GTR model, using the same priors in both cases. The chains were run for a total of 22 000 cycles, with two replicates under each version. Burnins of 2 000 points were discarded and posterior means and credibility intervals were computed for several parameters and key summary statistics (total tree length, α, number of occupied components, mean entropy of the equilibirum frequency profiles across sites, sum of the Dirichlet hyperparameters $\sum_a \nu_a$, and entropy of the relative exchangeabilities between pairs of amino-acids). Means and credibility intervals for these statistics are reported, for one chain under each implementation, in table S1. Bipartition frequencies and branch lengths estimated from two runs, one under each implementation, were plotted against each other for visual comparison (figure S2).

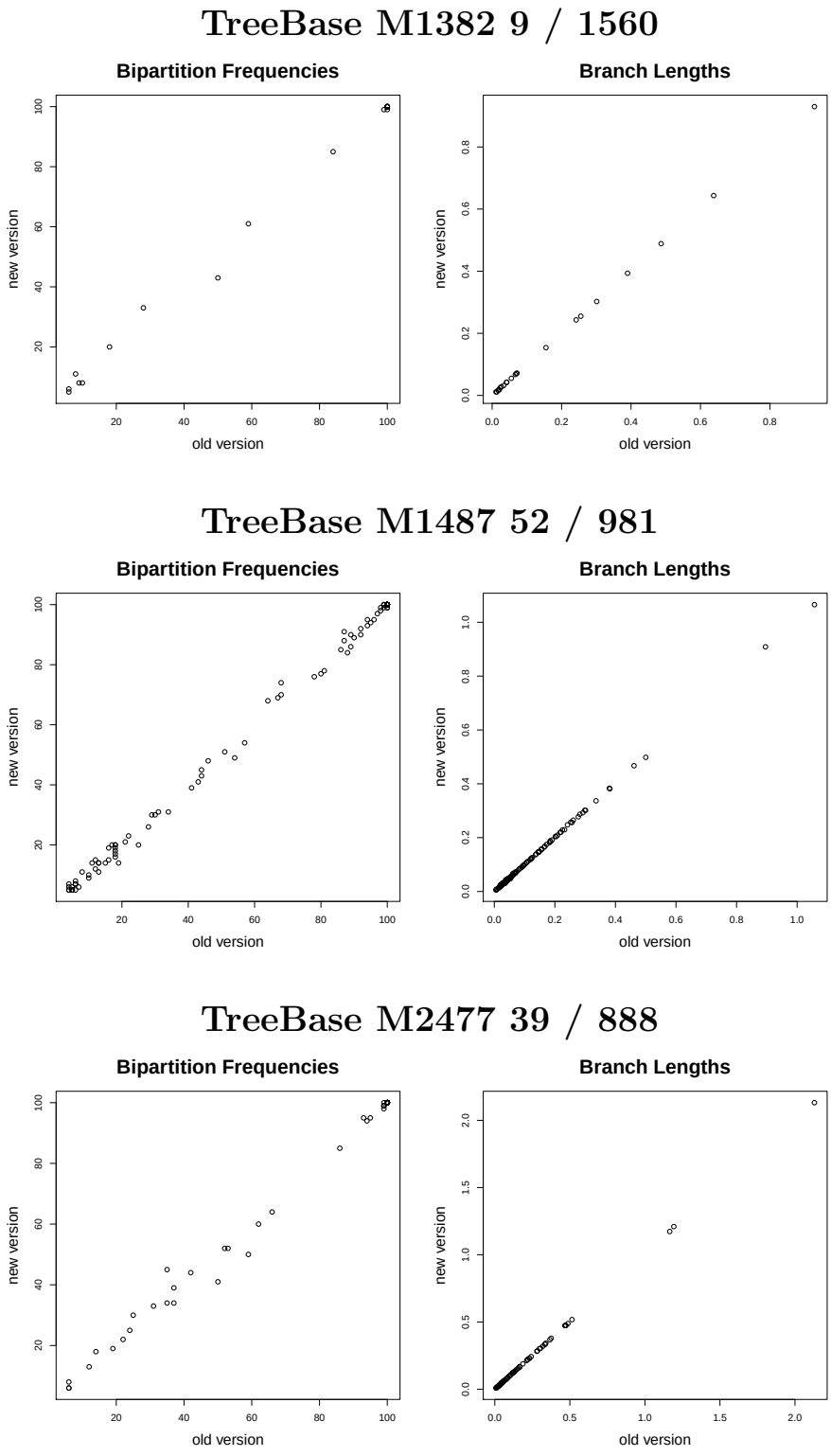
Table S1. Posterior mean and 95% credibility intervals for key statistics under the two implementations (CAT-GTR model)

	old implementation	MPI implementation
TreeBase M1382		
tree length	3.76 (3.39, 4.17)	3.73 (3.38, 4.15)
alpha	4.28 (3.01, 6.05)	4.30 (3.05, 6.01)
number of components	30.92 (21.00, 43.00)	29.62 (20.00, 41.00)
stationary entropy	1.87 (1.78, 1.96)	1.88 (1.79, 1.97)
dirichlet weight	6.24 (5.17, 7.70)	6.51 (5.30, 8.01)
exchangeabilities entropy	4.83 (4.77, 4.89)	4.83 (4.77, 4.89)
TreeBase M1487		
tree length	14.02 (12.69, 15.52)	14.00 (12.67, 15.44)
alpha	1.15 (1.02, 1.31)	1.15 (1.02, 1.30)
number of components	84.78 (61.00, 112.00)	82.96 (61.00, 109.00)
stationary entropy	2.12 (2.04, 2.19)	2.13 (2.06, 2.20)
dirichlet weight	8.88 (7.19, 10.74)	9.11 (7.58, 10.85)
exchangeabilities entropy	4.57 (4.50, 4.65)	4.57 (4.49, 4.64)
TreeBase M2477		
tree length	15.43 (13.93, 17.12)	15.41 (13.92, 17.03)
alpha	0.86 (0.78, 0.94)	0.86 (0.78, 0.94)
number of components	82.62 (64.00, 104.00)	83.04 (65.00, 103.00)
stationary entropy	2.20 (2.13, 2.26)	2.19 (2.13, 2.26)
dirichlet weight	11.45 (9.70, 13.37)	11.49 (9.74, 13.38)
exchangeabilities entropy	4.50 (4.43, 4.57)	4.50 (4.43, 4.57)

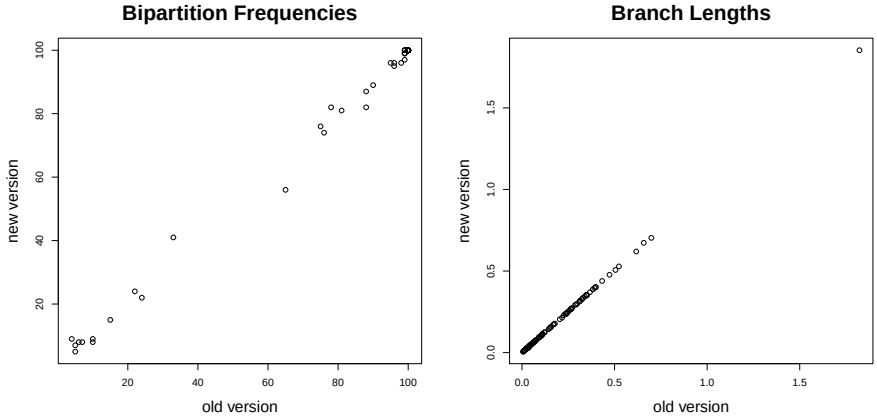
	old implementation	MPI implementation
Chordates		
tree length	19.63 (18.64, 20.63)	19.56 (18.55, 20.52)
alpha	1.06 (1.00, 1.11)	1.06 (1.01, 1.12)
number of components	275.51 (246.00, 307.00)	273.92 (244.00, 310.00)
stationary entropy	1.85 (1.80, 1.89)	1.86 (1.81, 1.90)
dirichlet weight	6.52 (5.93, 7.16)	6.77 (6.09, 7.52)
exchangeabilities entropy	4.41 (4.36, 4.46)	4.40 (4.35, 4.46)
Algae		
tree length	17.48 (16.60, 18.40)	17.39 (16.56, 18.25)
alpha	1.13 (1.07, 1.19)	1.13 (1.08, 1.19)
number of components	267.20 (233.00, 303.00)	268.23 (232.00, 305.00)
stationary entropy	2.02 (1.98, 2.06)	2.03 (2.00, 2.06)
dirichlet weight	8.81 (8.08, 9.56)	8.90 (8.18, 9.72)
exchangeabilities entropy	4.48 (4.42, 4.53)	4.47 (4.42, 4.53)
Metazoa		
tree length	19.58 (18.76, 20.47)	19.53 (18.66, 20.40)
alpha	1.29 (1.22, 1.36)	1.29 (1.22, 1.37)
number of components	296.83 (264.00, 333.00)	294.31 (258.00, 328.00)
stationary entropy	2.01 (1.98, 2.04)	2.01 (1.97, 2.04)
dirichlet weight	7.61 (7.02, 8.22)	7.63 (7.01, 8.35)
exchangeabilities entropy	4.48 (4.43, 4.53)	4.48 (4.43, 4.53)

	old implementation	MPI implementation
Mammals Mitochondrial		
tree length	23.44 (21.89, 25.07)	23.02 (21.44, 24.72)
alpha	0.80 (0.76, 0.84)	0.79 (0.75, 0.84)
number of components	176.46 (154.00, 199.00)	175.91 (156.00, 198.00)
stationary entropy	1.71 (1.67, 1.75)	1.71 (1.67, 1.75)
dirichlet weight	6.10 (5.51, 6.73)	6.15 (5.55, 6.83)
exchangeabilities entropy	4.22 (4.14, 4.30)	4.23 (4.15, 4.31)
Mammals Nuclear		
tree length	8.08 (7.85, 8.32)	8.09 (7.86, 8.33)
alpha	2.11 (1.96, 2.26)	2.11 (1.96, 2.28)
number of components	101.95 (83.00, 123.00)	101.63 (83.00, 122.00)
stationary entropy	2.30 (2.27, 2.33)	2.30 (2.26, 2.33)
dirichlet weight	10.64 (9.31, 12.09)	10.51 (9.16, 12.04)
exchangeabilities entropy	4.23 (4.18, 4.27)	4.23 (4.18, 4.27)

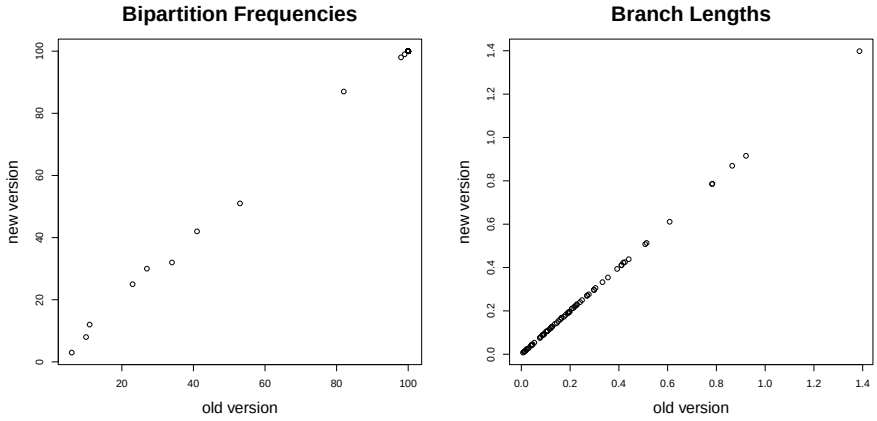
Figure S1. Bipartition frequencies (left) and posterior mean branch lengths (right) compared between the two implementations.



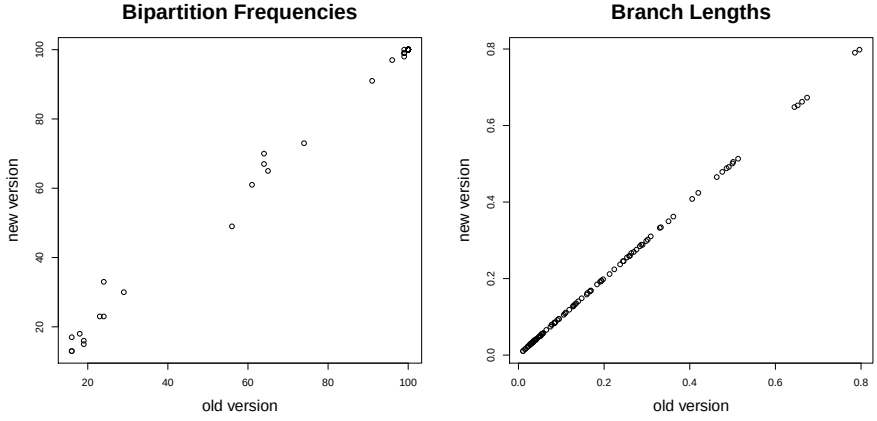
Chordates 51 / 5197



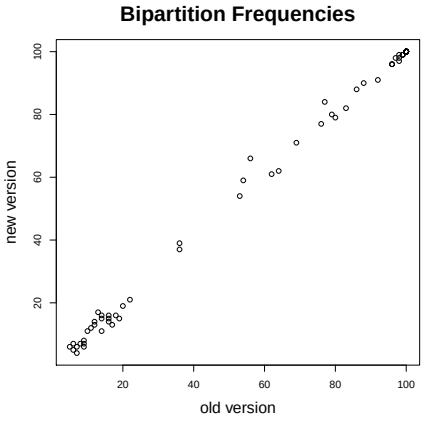
Algae 37 4743

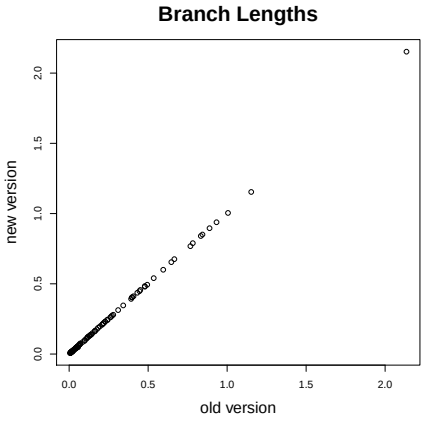


Metazoa 49 4431

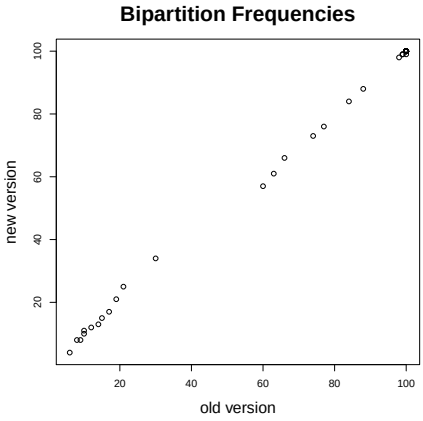


Mammals Mitochondrial 42 3507





Mammals Nuclear 42 4768



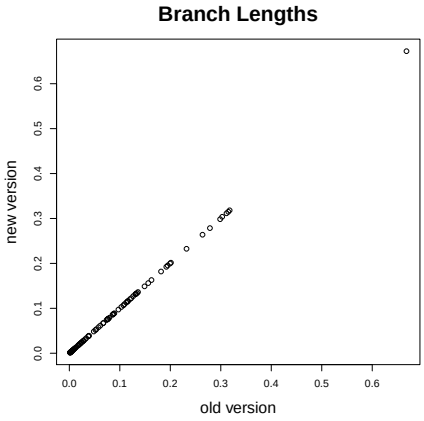


Figure S3. Posterior consensus tree obtained for the arthropod nucleotide dataset (Regier et al., 2010) under the GTR model. Posterior probability supports not distinguishable from 1 are not indicated.

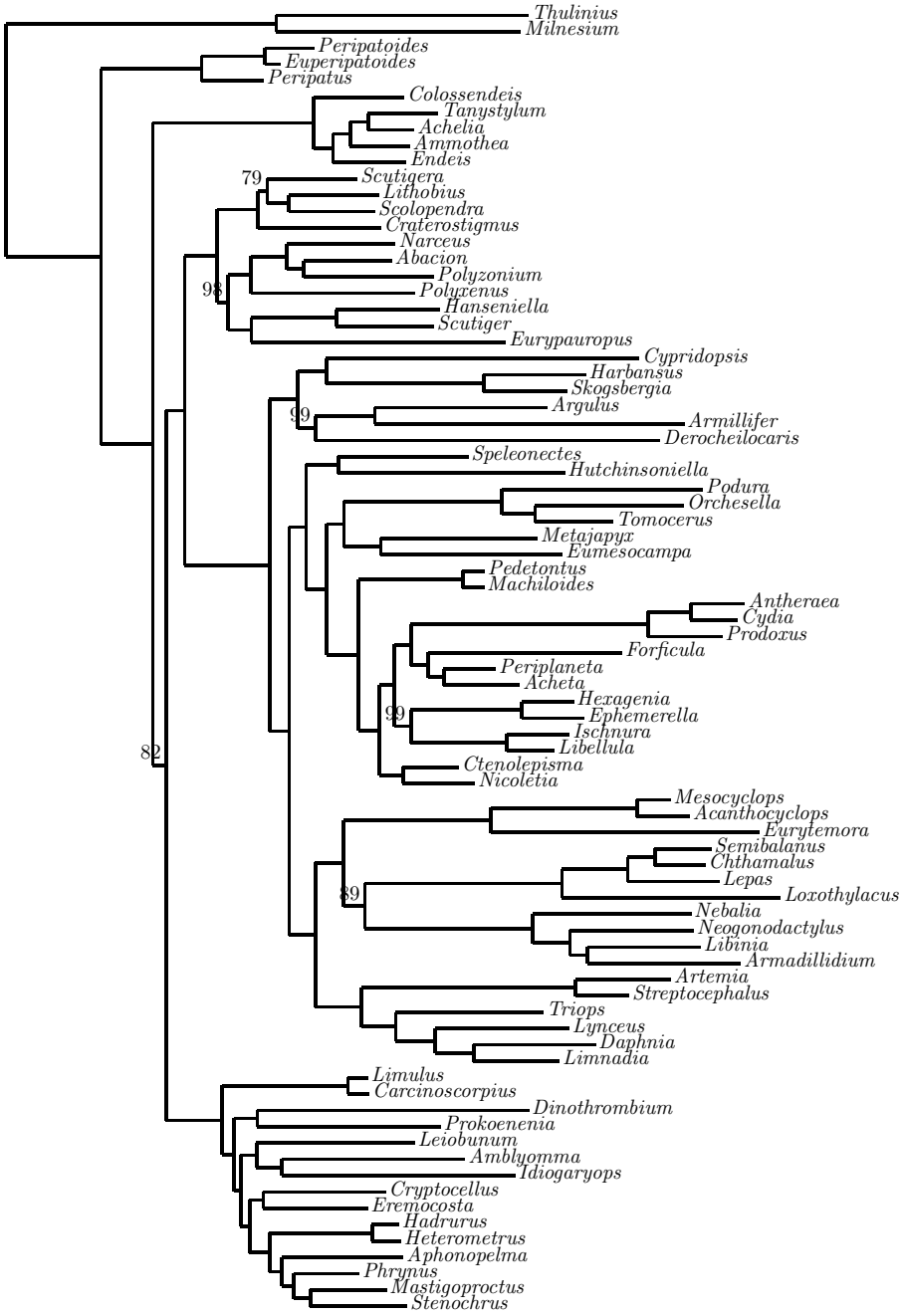


Figure S4. Posterior consensus tree obtained for the arthropod amino-acid recoded dataset (Regier et al., 2010) under the GTR model. Posterior probability supports not distinguishable from 1 are not indicated.

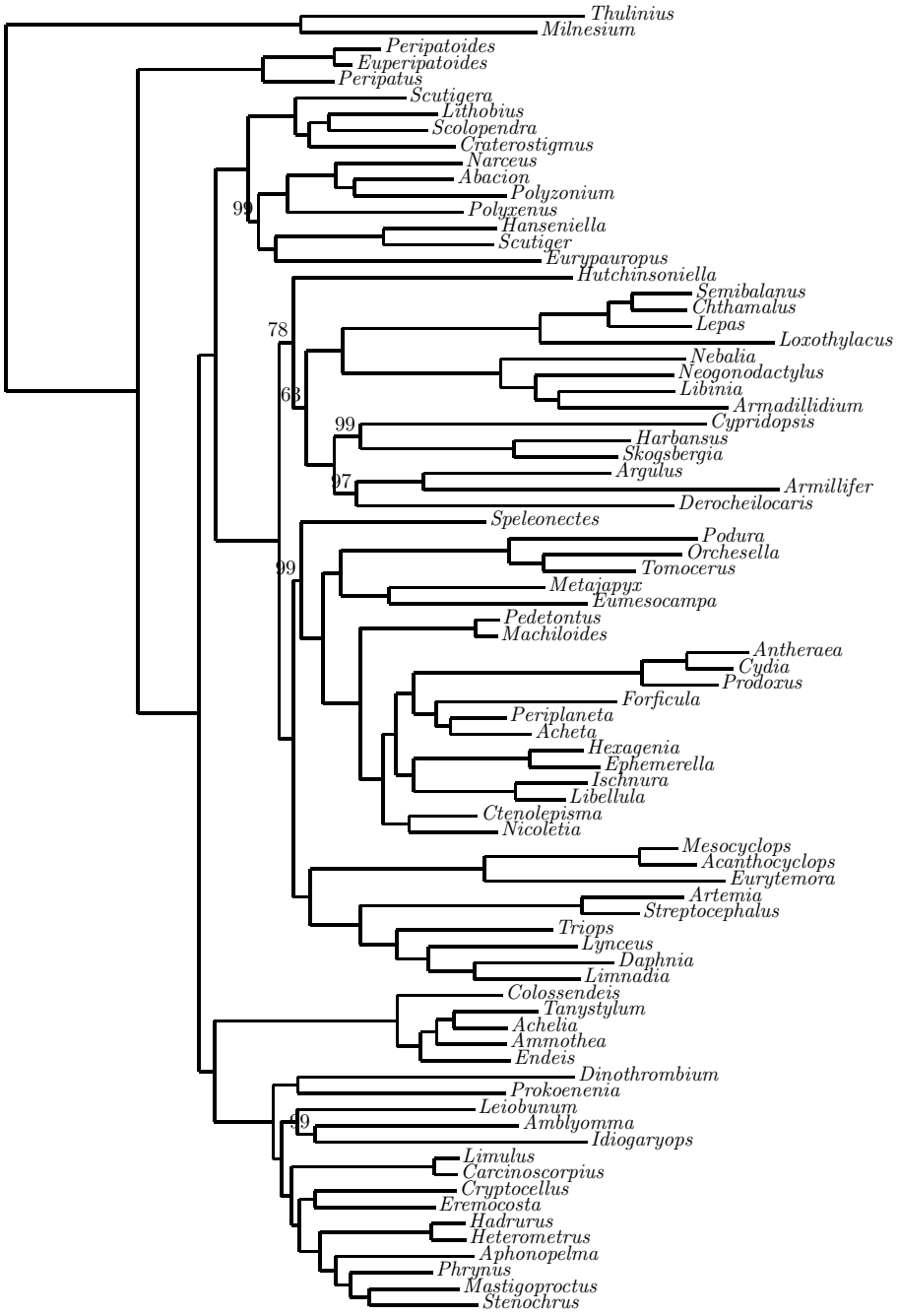


Figure S5. Posterior consensus tree obtained for the arthropod amino-acid recoded dataset (Regier et al., 2010) under the CAT model. Posterior probability supports not distinguishable from 1 are not indicated.

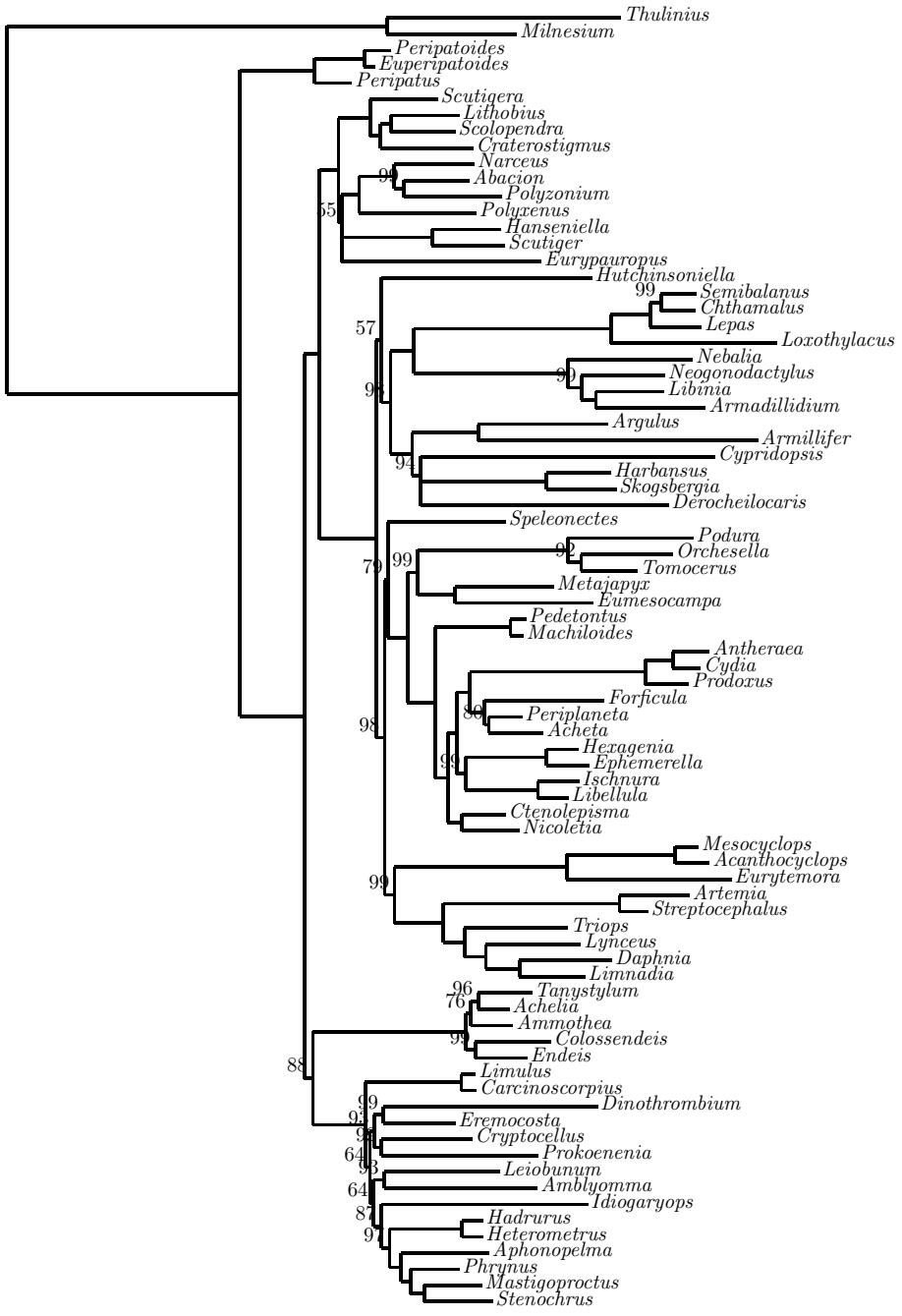
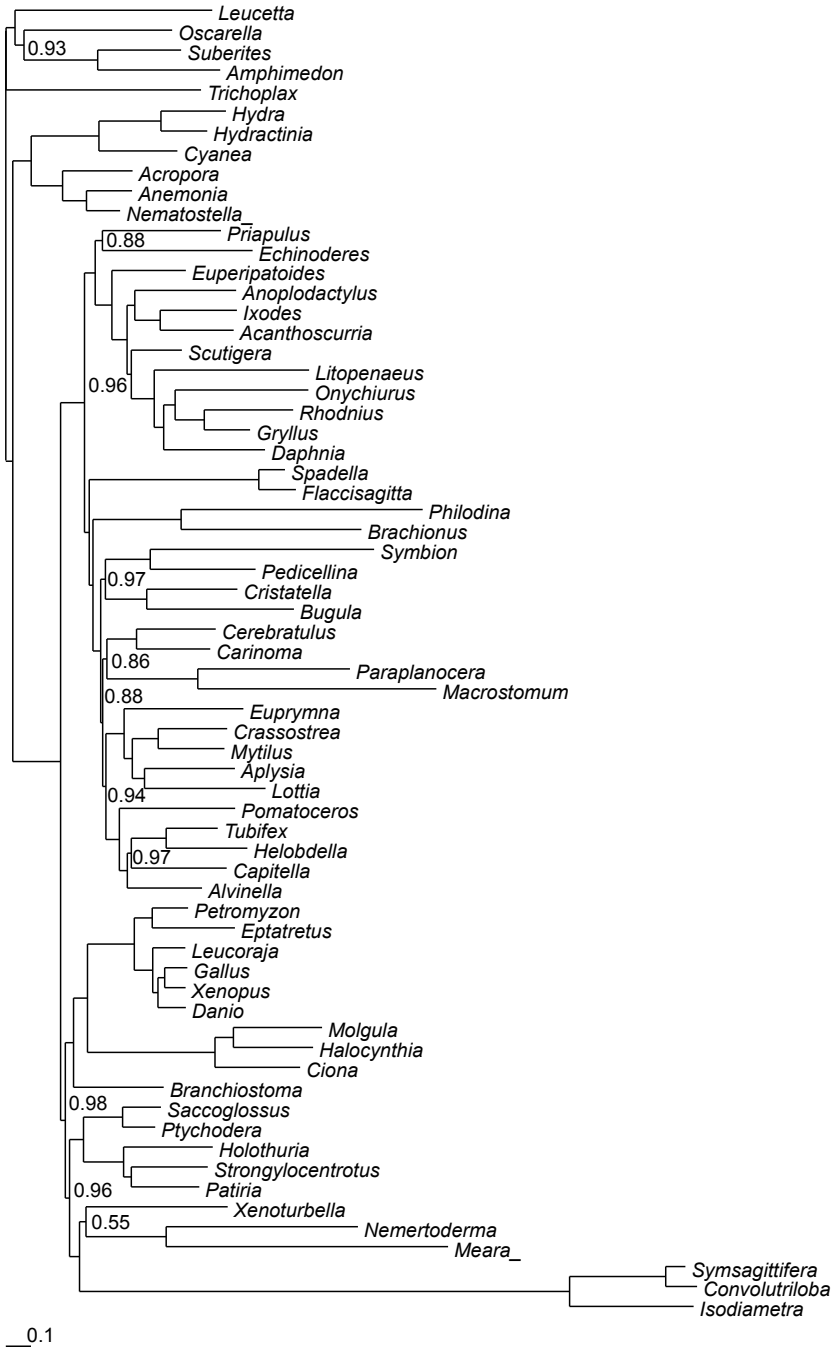


Figure S6. Posterior consensus tree obtained for a dataset comprising 38 330 aligned positions for 66 animal taxa (Philippe et al., 2011) under the CAT-GTR model. Posterior probability supports not distinguishable from 1 are not indicated.



References

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature. 439:965–968.

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

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

Hordijk W, Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics. 21:4338–4347.

Ishwaran H, James LF. 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association. 96:161–173.

Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. J Comput Biol. 13:1701–1722.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci. 363:1463–1472.

Papaspiliopoulos O, Roberts GO. 2008. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika. 95:169–186.

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature. 470:255–258.

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature. 463:1079–1083.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Sanderson MJ, Donoghue MJ, Piel WH, Eriksson T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany, Vol. 81, No. 6. (1994), 183. 81:183.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A. 100:1056–1061.

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