Software for Systematics and Evolution

Syst. Biol. 62(4):611–615, 2013

DOI:10.1093/sysbio/syt022
Advance Access publication April 5, 2013

PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment

NICOLAS LARTILLOT[1,2,*], NICOLAS RODRIGUE[3,4], DANIEL STUBBS[5], AND JACQUES RICHER[5]

[1]*Centre Robert Cedergren pour la Bioinformatique, Département de Biochimie, Université de Montréal, C.P. 6128, succursale Centre-ville. Montréal, Québec H3C 3J7, Canada;* [2]*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS Université de Montpellier 2, UMR 5506-CC477, 161 rue Ada, 34095 Montpellier Cedex 5, France;* [3]*Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada;* [4]*Department of Biology, University of Ottawa, 30 Marie Curie Pvt., Ottawa, Ontario K1N 6N5, Canada; and* [5]*Calcul Québec, Université de Montréal, C.P. 6128, succursale Centre-ville. Montréal, Québec H3C 3J7, Canada*

[*]*Correspondence to be sent to: Centre Robert Cedergren pour la Bioinformatique, Département de Biochimie, Université de Montréal, C.P. 6128, succursale Centre-ville. Montréal, Québec H3C 3J7, Canada; E-mail: nicolas.lartillot@umontreal.ca.*

Received 6 November 2012; reviews returned 23 December 2012; accepted 27 March 2013
Associate Editor: David Posada

Abstract.—Modeling across site variation of the substitution process is increasingly recognized as important for obtaining more accurate phylogenetic reconstructions. Both finite and infinite mixture models have been proposed and have been shown to significantly improve on classical single-matrix models. Compared with their finite counterparts, infinite mixtures have a greater expressivity. However, they are computationally more challenging. This has resulted in practical compromises in the design of infinite mixture models. In particular, a fast but simplified version of a Dirichlet process model over equilibrium frequency profiles implemented in PhyloBayes has often been used in recent phylogenomics studies, while more refined model structures, more realistic and empirically more fit, have been practically out of reach. We introduce a message passing interface version of PhyloBayes, implementing the Dirichlet process mixture models as well as more classical empirical matrices and finite mixtures. The parallelization is made efficient thanks to the combination of two algorithmic strategies: a partial Gibbs sampling update of the tree topology and the use of a truncated stick-breaking representation for the Dirichlet process prior. The implementation shows close to linear gains in computational speed for up to 64 cores, thus allowing faster phylogenetic reconstruction under complex mixture models. PhyloBayes MPI is freely available from our website www.phylobayes.org. [Bayesian inference; Dirichlet process; mixture models; phylogenetics; phylogenomics.]

Phylogenetic studies often rely on genetic sequences that are sufficiently conserved to be alignable at large evolutionary scale. The sequences of interest are therefore typically under a regime of strong purifying selection, characterized by site-specific constraints for subsets of acceptable nucleotides or amino acids. Such selective regimes in turn result in a substantial variation across sites of the substitution process which, if not properly modeled, may represent a major cause of systematic errors in phylogenetic reconstruction of deep evolutionary relationships.

Mixture models have been proposed as a natural and simple approach for modeling site-specific effects and have been applied to nucleotides (Pagel and Meade 2004; Evans and Sullivan 2012), amino acids (Koshi and Goldstein 1998, 2001; Lartillot and Philippe 2004; Le et al. 2008; Quang et al. 2008; Wang et al. 2008) and, more recently, codon alignments (Rodrigue et al. 2010). The limiting case of countably infinite mixtures, such as Dirichlet processes (Ferguson 1973), can be seen as nonparametric approaches (Green and Richardson 2001). They have been used in phylogenetics for fitting general distributions of random effects across sites (Lartillot and Philippe 2004; Huelsenbeck et al. 2006; Huelsenbeck and Suchard 2007; Rodrigue et al. 2010).

A particular case of infinite mixture, the CAT model (Lartillot and Philippe 2004), was devised for accomodating the rich structure of among-site variation in substitution patterns present in real data. The CAT model, implemented in PhyloBayes (Lartillot et al. 2009), has a better fit than classical empirical matrices on many phylogenomic data sets and is less sensitive to systematic errors (Lartillot et al. 2007; Philippe et al. 2009, 2011a, 2011b). For this reason, it is increasingly used in phylogenomic studies, especially for solving phylogenetic problems spanning deep evolutionary times (e.g., Cox et al. 2008; Kocot et al. 2011). However, the CAT model has several limitations. In particular, it is a mixture of F81 (Felsenstein 1981) processes, such that, upon a substitution, the probability of the final state does not depend on the initial state. This choice was primarily motivated by computational considerations (Lartillot and Philippe 2004; Lartillot 2006). Yet, F81 processes offer a poor representation of observed substitution patterns both in nucleotide sequences, characterized by unequal rates of transitions and transversions, and in amino acid sequences, where the constraint of the genetic

code results in unequal exchangeabilities between amino acids.

A more general version of the infinite mixture, the CAT–GTR model (Lartillot and Philippe 2004), combines the advantages of the CAT model (a Dirichlet process mixture on equilibirium-frequency profiles) with the richer expressivity of general time reversible Markov processes. However, estimating phylogenies under the CAT–GTR model, or even under the faster CAT settings in the case of very large phylogenomic data sets, remains computationally challenging and requires more efficient algorithmic strategies. To address this practical problem, here we introduce an message passing interface (MPI) parallelization of the PhyloBayes program, resulting in significant gains in computational speed and making CAT–GTR a viable alternative to both classical empirical mixtures and the CAT model for large phylogenomic data sets.

Description

Parallelization of the MCMC sampler is done by having one master process dispatching the computational tasks, collecting and summing the results, and $K-1$ slave processes, each in charge of a segment of the multiple sequence alignment or of a subset of the components of the mixture. Most of the computational burden is thus equally divided among slaves. The MCMC proceeds in two phases: a dynamic-programming phase, alternating with a data-augmentation phase.

During the dynamic-programming phase, the master broadcasts current parameter values and sends orders to the slaves, specifying tasks to be undertaken (e.g., rearranging the tree in specific ways or computing the likelihood). Each slave executes the orders and returns to the master the information needed for the master to decide whether to accept or reject the move. The tree topology is updated by subtree pruning and regrafting, according to a partial Gibbs sampling algorithm. This move is initiated by the master randomly choosing a subtree to be pruned and broadcasting the outcome of this choice to all slaves. Each slave then performs a complete scan of all possible topologies obtained by regrafting the subtree onto the main tree, computing the likelihood of each of the resulting topologies for the segment of the data set under its responsibility, and using a caching method (similar to Hordijk and Gascuel 2005) to accelerate likelihood computation. Once this is done, each slave sends back to the master an array contaning one single log likelihood for each regrafting point. The master collects the arrays, sums them up over all slaves for each regrafting position, and chooses a regrafting position based on the Gibbs-sampling decision rule.

The data-augmentation phase starts by having all slaves sample a substitution history, or mapping, from the conditional posterior distribution at each site under their charge (Nielsen 2002; Rodrigue et al. 2008), calculate sufficient statistics based on these substitution mappings (Lartillot 2006), and send the sufficient statistics to the master. Conditional on these sufficient statistics, the master then performs updates of the parameters of the model, comprising branch lengths, the alpha parameter of the discrete gamma distribution of rates across sites, and the relative exchangeabilities between states (Lartillot 2006). The updating of the Dirichlet process mixture over equilibrium frequency profiles is also conditional on the mapping-based sufficient statistics. A truncated stick-breaking representation of the Dirichlet process prior is used (Ishwaran and James 2001), which represents a parameter expansion more easily amenable to parallelization (Suchard et al. 2010) than the classical Chinese restaurant representation (Neal 2000). In addition, we developed a hybrid between Gibbs-sampling and Metropolis–Hastings, inspired by (Papaspiliopoulos and Roberts 2008), for resampling the allocations of the sites to the components of the mixture. Our method, which improves on the plain Gibbs sampling classically used in this context (Ishwaran and James 2001), hinges on the realization that most components of the truncated yet large mixture have negligible weights, so that an exhaustive evaluation of the likelihood of all possible allocations of a given site is most often not necessary (see Supplementary Information, available from Dryad under doi:10.5061/dryad.c459h).

Benchmark

The MPI implementation was checked against the serial version of PhyloBayes (Supplementary Table S1 and Supplementary Fig. S2). The scaling properties of the parallel version were then evaluated on a dedicated cluster (Intel Quad Core Xeon E5450 3 GHz bi-processors, 8 cores per node, DDR2 667 Mhz, 16 Go RAM), under linux CentOS 5.8 and using low latency communication (Infiniband). Under either the CAT or the CAT–GTR model (Fig. 1), for large data sets ($>15\,000$ aligned positions) close to linear gains are obtained for up to 64 cores (eight nodes). For smaller data sets, 24–32 cores appear to represent a reasonable compromise between computational speed and occupation of resources. Similar gains were observed on other clusters, although less optimal hardware (e.g., Ethernet instead of Infiniband) or competition for the bandwidth between jobs could possibly result in decreased efficiency when the run is dispatched over several nodes of the cluster. We also observed consistent differences in efficiency between alternative architectures (e.g., Intel-Xeon vs. AMD-opteron processors).

Comparisons with the serial version of PhyloBayes are less straightforward, as multiple differences in the underlying implementations and algorithms, combined with the specific constraints inherent to parallelization, result in different mixing behavior and computational speed for the two implementations. From the data sets

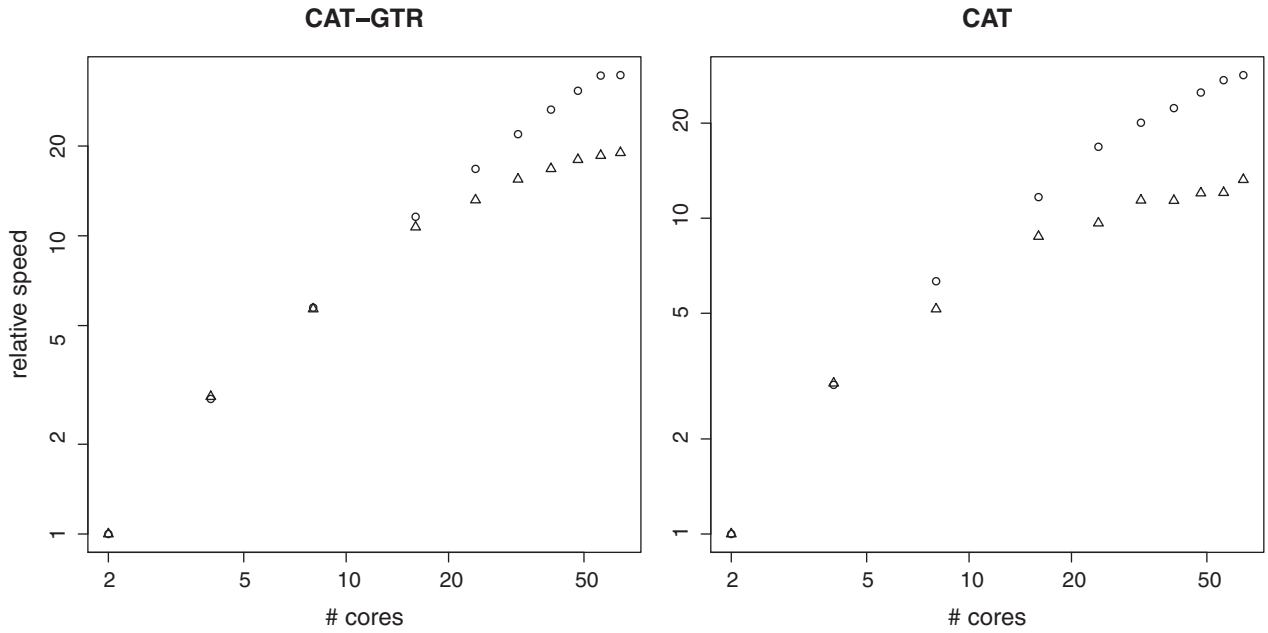

FIGURE 1. Gain in computational speed under the CAT–GTR (left) and the CAT (right) models. Gains are relative to a two-core computation (one master and one slave). Open circles: bilaterian data set (79 taxa, 14 909 positions Lartillot and Philippe 2008); diamonds: plastid data set (28 taxa, 10 137 positions Rodríguez-Ezpeleta et al. 2007).

that we studied and that are sufficiently small to be analyzed using both implementations, analyses under both CAT and CAT–GTR appear to reach convergence about 10 times faster under the parallel (with 32 cores) than under the serial version. This, however, may depend on the specific data set under investigation.

The program allows for standard Bayesian estimation of posterior consensus trees under CAT, CAT–GTR, finite models, and empirical matrices. As in the original serial version, it also implements posterior predictive checks and cross-validation methods for measuring model fit. It can be noted, however, that both CAT and single-matrix models are submodels of CAT–GTR. The use of CAT–GTR therefore amounts to an implicit model-averaging procedure, which should automatically select the most adequate configuration, thus suggesting that the CAT–GTR model could be used by default in many practical cases.

BIOLOGICAL EXAMPLES

As a first example, we ran the program on a concatenation of 62 nuclear protein-coding genes (13 087 coding positions) from 80 species representing all major groups of Panarthropoda (Regier et al. 2010). The nucleotide data set (21 823 aligned nucleotide positions, excluding third codon positions and those first codon positions encoding at least one leucine or arginine codons, as in the original article) was analyzed under GTR and CAT–GTR. The amino acid recoded matrix was analyzed under GTR, CAT and CAT–GTR. Two independent chains were run in each case. Convergence and mixing were assessed visually and quantified using convergence diagnostics based on discrepancy measures between the two chains, as well as empirical autocorrelations (Lartillot et al. 2009). Since runs were dispatched over two clusters differing in their architecture, in the following, we report all running times in Xeon-equivalent days (with 32 cores per run).

In the case of the nucleotide data set, a difference of at most 0.1 in bipartition support between the two chains, and an effective sample size greater than 300, were obtained after approximately 3 days under the two models. Excluding constant positions improved convergence and mixing. The tree inferred under CAT–GTR (Fig. 2) is globally compatible with the original analysis under GTR (Supplementary Fig. S3; Regier et al. 2010), although with several interesting and supported differences, in particular within Arachnida, suggesting that mixture models might be an interesting alternative to classical one-matrix models for nucleotide data (Pagel and Meade 2004; Evans and Sullivan 2012).

In the case of the amino acid data set, and excluding constant positions, differences in bipartition posterior probabilities not exceeding 0.15, and effective sample sizes greater than 100, were obtained after 5 days under CAT and 10 days under CAT–GTR. As for nucleotides, several supported incompatibilities are observed between the trees inferred under homogeneous and mixture models (Supplementary Figs S4 and S5), confirming previous observations collected across several examples of large-scale reconstruction using amino acid data (Lartillot et al. 2007; Philippe et al. 2009, 2011a, 2011b). On the other hand, the differences between the two infinite mixture models, CAT

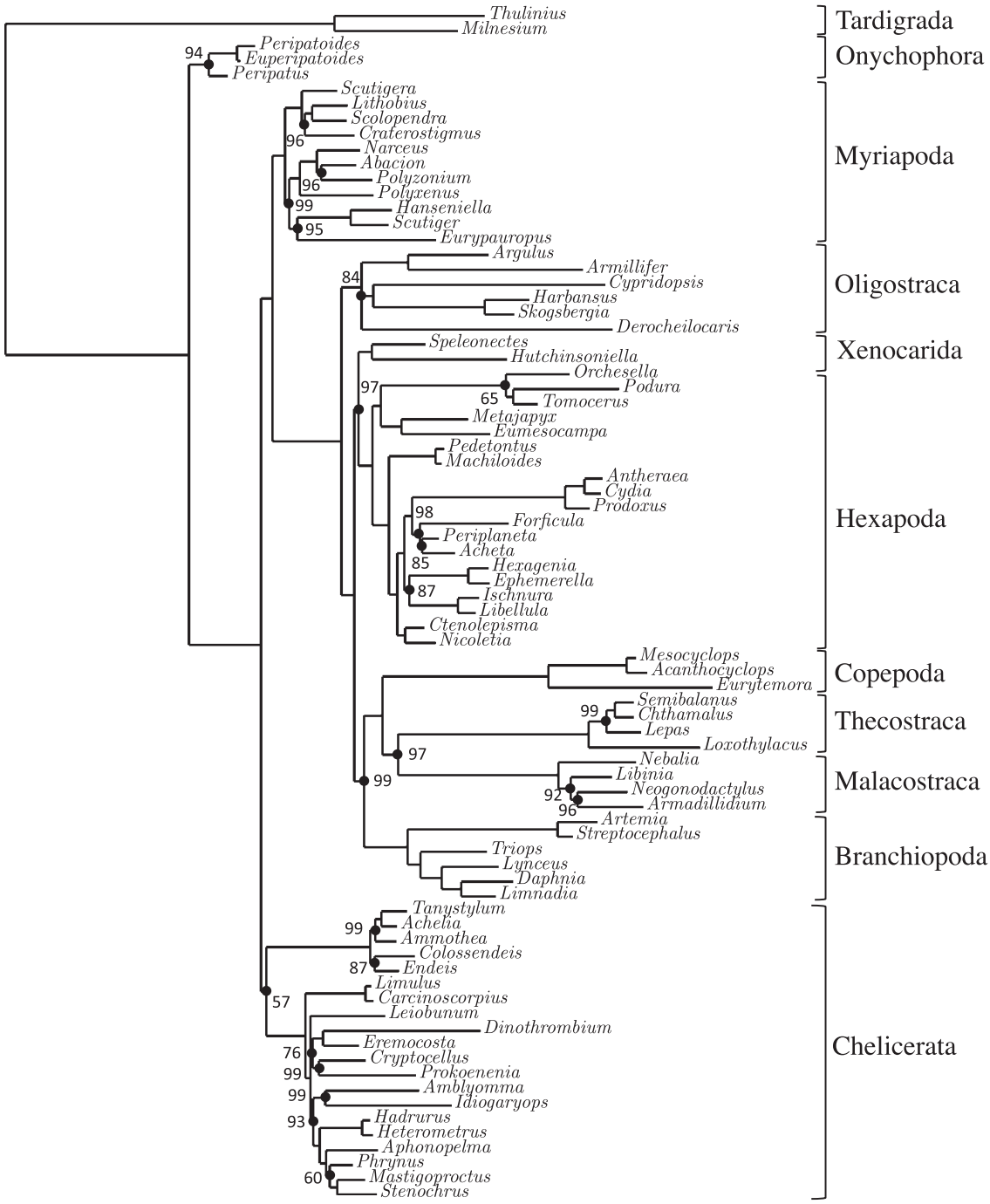

FIGURE 2. Posterior consensus tree obtained for a 62 nuclear protein-coding genes (13 087 coding positions) from 80 panarthropod species (Regier et al. 2010) under the CAT–GTR model. Posterior probability supports not distinguishable from 1 are not indicated.

and CAT–GTR, are more subtle and more weakly supported (Supplementary Figs S4 and S5, see also Stabelli et al. 2012).

As a second and larger example, a concatenate of 38 330 aligned positions for 66 animal taxa (Philippe et al. 2011a) was reanalyzed here with the MPI version under the CAT and the CAT–GTR models. Constant positions were removed, leaving a set of 27 290 nonconstant positions for subsequent analysis. The estimation took 5 days under CAT, and 15 days under CAT–GTR, for a maximum difference in posterior probability support of 0.15. Effective sample sizes were greater than 100 for all summary statistics under CAT, but remained small under the CAT–GTR model, particularly for the summary statistics monitoring the mixing of the exchangeability parameters and of the Dirichlet process. This particular example is therefore at the limit of tractability under the CAT–GTR model, and perhaps the

main advantage of the use of the MPI version in the present case lies in the faster rate of convergence under CAT. In the present case, the consensus tree obtained under CAT–GTR (Supplementary Fig. S6) was similar to the tree inferred under CAT (Philippe et al. 2011a).

Altogether, when computational resources permit it, we suggest exploring the use of the CAT–GTR model, now rendered more tractable with the present MPI implementation, while benefiting from faster analysis under CAT and one-matrix models for the largest phylogenomic data sets.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.c459h.

Funding

The Natural Science and Engineering Research Council of Canada (to N.L.) and Agriculture and Agri-Food Canada (to N.R.).

Acknowledgments

We wish to thank Éric Fournier, Bastien Boussau, Andrew Roger, Matthew Brown and Hervé Philippe for their extensive testing of the code, as well as Jeremy Brown and Leonardo Martins for their useful comments on the manuscript. Computational resources were provided by Calcul Québec and Compute Canada and the Canadian Foundation for Innovation.

References

Cox C.J., Foster P.G., Hirt R.P., Harris S.R., Embley T.M. 2008. The archaebacterial origin of eukaryotes. Proc. Natl Acad. Sci. U. S. A. 105:20356–20361.

Evans J., Sullivan J. 2012. Generalized mixture models for molecular phylogenetic estimation. Syst. Biol. 61:12–21.

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

Ferguson T.S. 1973. A Bayesian analysis of some nonparametric problems. Ann. Stat. 1:209–230.

Green P.J., Richardson S. 2001. Modelling heterogeneity with and without the Dirichlet process. Scand. J. Stat. 28:355–375.

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

Huelsenbeck J.P., Jain S., Frost S.W.D., Pond S.L.K. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc. Natl Acad. Sci. U. S. A. 103:6263–6268.

Huelsenbeck J.P., Suchard M.A. 2007. A nonparametric method for accommodating and testing across-site rate variation. Syst. Biol. 56:975–987.

Ishwaran H., James L.F. 2001. Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. 96:161–173.

Kocot K.M., Cannon J.T., Todt C., Citarella M.R., Kohn A.B., Meyer A., Santos S.R., Schander C., Moroz L.L., Lieb B., Halanych K.M. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452–456.

Koshi J.M., Goldstein R.A. 1998. Models of natural mutations including site heterogeneity. Proteins 32:289–295.

Koshi J.M., Goldstein R.A. 2001. Analyzing site heterogeneity during protein evolution. Pac. Symp. Biocomput. 6:191–202.

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

Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7 (Suppl 1):S4.

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

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

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.

Le S.Q., Lartillot N., Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363:3965–3976.

Neal R.M. 2000. Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. 9:249–265.

Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51: 729–739.

Pagel M., Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

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

Philippe H., Brinkmann H., Copley R.R., Moroz L.L., Nakano H., Poustka A.J., Wallberg A., Peterson K.J., Telford M.J. 2011a. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470:255–258.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706–712.

Quang L.S., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24: 2317–2323.

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

Rodrigue N., Philippe H., Lartillot N. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics 24:56–62.

Rodrigue N., Philippe H., Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl Acad. Sci. U. S. A. 107:4629–4634.

Rodríguez-Ezpeleta N., Philippe H., Brinkmann H., Becker B., Melkonian M. 2007. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. Mol. Biol. Evol. 24:723–731.

Stabelli O.R., Lartillot N., Philippe H., Pisani D. 2012. Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst. Biol. 62:121–133.

Suchard M.A., Wang Q., Chan C., Frelinger J., Cron A., West M. 2010. Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures. J. Comput. Graph. Stat. 19:419–438.

Wang H.-C., Li K., Susko E., Roger A. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol. Biol. 8:331.