Software for Systematics and Evolution

MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space

FREDRIK RONQUIST[1,*], MAXIM TESLENKO[1], PAUL VAN DER MARK[2], DANIEL L. AYRES[3], AARON DARLING[4], SEBASTIAN HÖHNA[5], BRET LARGET[6], LIANG LIU[7], MARC A. SUCHARD[8], AND JOHN P. HUELSENBECK[9]

[1]*Department of Biodiversity Informatics, Swedish Museum of Natural History, SE-10405 Stockholm, Sweden;* [2]*Department of Scientific Computing, Florida State University, FL 32306, USA;* [3]*Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA;* [4]*Genome Center, University of California, Davis, CA 95616, USA;* [5]*Department of Mathematics, Stockholm University, SE-10691 Stockholm, Sweden;* [6]*Departments of Statistics and Botany, University of Wisconsin, Madison, WI 53706, USA;* [7]*Departments of Agriculture and Natural Resources, Delaware State University, Dover, DE 19901, USA;* [8]*Departments of Biomathematics, Biostatistics and Human Genetics, University of California, Los Angeles, CA 90095, USA; and* [9]*Department of Integrative Biology, University of California, Berkeley, CA 94720, USA;*
Correspondence to be sent to: Department of Biodiversity Informatics, Swedish Museum of Natural History, SE-10405 Stockholm, Sweden; E-mail: fredrik.ronquist@nrm.se.

Abstract.—Since its introduction in 2001, MrBayes has grown in popularity as a software package for Bayesian phylogenetic inference using Markov chain Monte Carlo (MCMC) methods. With this note, we announce the release of version 3.2, a major upgrade to the latest official release presented in 2003. The new version provides convergence diagnostics and allows multiple analyses to be run in parallel with convergence progress monitored on the fly. The introduction of new proposals and automatic optimization of tuning parameters has improved convergence for many problems. The new version also sports significantly faster likelihood calculations through streaming single-instruction-multiple-data extensions (SSE) and support of the BEAGLE library, allowing likelihood calculations to be delegated to graphics processing units (GPUs) on compatible hardware. Speedup factors range from around 2 with SSE code to more than 50 with BEAGLE for codon problems. Checkpointing across all models allows long runs to be completed even when an analysis is prematurely terminated. New models include relaxed clocks, dating, model averaging across time-reversible substitution models, and support for hard, negative, and partial (backbone) tree constraints. Inference of species trees from gene trees is supported by full incorporation of the Bayesian estimation of species trees (BEST) algorithms. Marginal model likelihoods for Bayes factor tests can be estimated accurately across the entire model space using the stepping stone method. The new version provides more output options than previously, including samples of ancestral states, site rates, site d_N/d_S rations, branch rates, and node dates. A wide range of statistics on tree parameters can also be output for visualization in FigTree and compatible software. [Bayes factor; Bayesian inference; MCMC; model averaging; model choice.]

Bayesian Markov chain Monte Carlo (MCMC) methods quickly gained in popularity after they were introduced in statistical phylogenetics in the late 1990's (Mau and Newton 1997; Yang and Rannala 1997; Larget and Simon 1999; Mau et al. 1999). This was due to the inherent advantages of the approach but also to the availability of easy-to-use software packages, such as MrBayes (Huelsenbeck and Ronquist 2001). Originally, MrBayes only supported simple phylogenetic models, but the model space expanded considerably in version 3.0 (Ronquist and Huelsenbeck 2003). In addition to a wide range of models on binary, "standard" (morphology), nucleotide and amino acid data, version 3.0 also supported mixed models. The latter allow different data partitions to be combined in the same model, with parameters linked or unlinked across partitions according to user specifications. MrBayes 3.0 was apparently the first statistical phylogenetics package to support such models (Rannala and Yang 2008).

Bayesian phylogenetic inference using MCMC has developed in leaps and bounds since the release of MrBayes 3.0. In particular, the relative ease with which complex models can be tackled using the MCMC machinery has led to an explosion in the development of probabilistic evolutionary models (for a review, see Ronquist and Deans 2010). We have also seen the appearance of better MCMC algorithms and more sophisticated convergence diagnostics for phylogenetic models, and methods for Bayesian model choice have improved considerably.

With this note, we announce the official release of version 3.2 of MrBayes. Version 3.2 was originally intended as a relatively modest expansion of version 3.1, which added convergence diagnostics to the original features in version 3.0. Over the years, however, a number of significant new features were added to version 3.2, and large parts of the program were rewritten. When we now officially release version 3.2, it is every bit as significant in the evolution of the program as the release of version 3.0 almost a decade ago.

DESCRIPTION OF NEW FEATURES

Convergence

The phylogenetics community has come to accept as good practice that Bayesian MCMC results

be accompanied by a critical assessment of convergence. Arguably, the best way of accomplishing this is to compare samples obtained from independent MCMC analyses. It is typically the tree samples that are most divergent in phylogenetic analyses, and we therefore introduced the average standard deviation of split frequencies (ASDSF) in MrBayes to allow quantitative assessment of the similarity among such samples.

ASDSF is calculated by comparing split or clade frequencies across multiple independent MCMC runs that ideally should be started from different randomly chosen starting trees (Lakner et al. 2008). ASDSF should approach 0.0 as runs converge to the same distribution. The frequencies of rare splits or clades are difficult to estimate accurately and these groupings are usually of marginal interest. Therefore, it may be advantageous to exclude them from the diagnostic. MrBayes allows the user to set a cutoff frequency (default value 0.10); all splits or clades occurring minimally at that frequency in at least one of the runs will be incorporated in the ASDSF.

To allow users to monitor MCMC progress, MrBayes can run several analyses in parallel and report the average (ASDSF) or maximum standard deviation of split frequencies at regular intervals. More detailed diagnostics can be obtained using the "sump" and "sumt" commands after the run has completed. They include ASDSF across runs for each of the sampled clades in addition to the potential scale reduction factor (PSRF; Gelman and Rubin 1992) for branch lengths, node times, and substitution model parameters. PSRF compares the variance within and between runs and should approach 1.0 as runs converge. MrBayes 3.2 also reports the effective sample size, widely used for single-run convergence diagnostics.

MrBayes 3.2 also introduces several new features intended to improve MCMC convergence rates. A number of new tree proposal mechanisms have been added, including subtree-swapping moves and extending subtree-pruning-and-regrafting moves, and the default mix of proposals has been optimized (Lakner et al. 2008). MrBayes 3.2 further includes a completely new type of tree proposal that is guided using parsimony scores. The details of the parsimony-biased proposals will be presented elsewhere; however, tentative empirical results show that they can improve the speed of convergence by an order of magnitude on some problems (see also Höhna and Drummond 2012). For nontree proposals, MrBayes 3.2 implements auto-tuning that automatically adjusts tuning parameters such that a target acceptance frequency is reached (Roberts and Rosenthal 2009). Since previous versions, MrBayes supports Metropolis coupling (heated chains) to accelerate convergence. To simplify monitoring of convergence, MrBayes 3.2 prints ASDSF values, acceptance rates of moves, and acceptance rates of swaps between Metropolis-coupled chains to a separate file with a ".mcmc" suffix during runs.

Faster and More Convenient Computation

Much of the computational effort in a phylogenetic MCMC analysis is spent calculating likelihoods. To improve speed, MrBayes 3.2 now employs streaming single-instruction-multiple-data extensions (SSE) for all likelihood calculations. SSE instructions are supported by most current CPUs and provide low-level parallelization of arithmetic operations. Importantly, MrBayes 3.2 also supports the use of the BEAGLE library for likelihood calculations (Ayres et al. 2012). With BEAGLE, the likelihood calculations can be farmed out to one or more graphics processing units (GPUs) on compatible hardware, resulting in significant speedups for codon and amino acid models in particular. BEAGLE can also be used for likelihood computation on the CPU.

MrBayes 3.2 does not support multithreading, but it does implement the message passing interface (MPI) for efficient parallel processing across large computer clusters (Altekar et al. 2004). On many hardware platforms, including Mac OS and Linux, it is possible to use the MPI-enabled Unix version of MrBayes to take advantage of multiple cores. However, MPI parallelization is across chains, which means that the maximum number of cores or processors that can be used by MrBayes is the same as the total number of heated and nonheated chains across all simultaneous runs. For instance, two runs of four chains each would be maximally accelerated on a system with eight processors or cores. The MPI version can be combined with BEAGLE to further expand the opportunity for computational parallelization.

Finally, to facilitate long runs, MrBayes 3.2 implements checkpointing across all models. At a frequency determined by the user, all parameter samples are printed to a ".ckp" file. If desired, the analysis can later be restarted from the checkpoint file, and the final results will appear as if the run had never been stopped.

New Models

Many phylogenetic hypotheses concern the structure of the phylogenetic tree. To facilitate such analyses, MrBayes 3.2 implements three types of constraints on the tree: hard, negative, and partial. A hard constraint forces a split or clade to be present in all trees sampled in the MCMC analysis, whereas a negative constraint forces a split or clade to be absent. Unlike hard and negative constraints, a partial constraint (or backbone constraint) can leave the position of some taxa indeterminate. The indeterminate taxa are allowed to appear on either side of the specified split if the tree is unrooted, or either within or outside the specified clade if the tree is rooted. Several hard, negative, and partial constraints can be combined into complicated priors on the shape of the tree. However, constraints are either on or off; they cannot be associated with probabilities in the current version.

Unlike previous versions, MrBayes 3.2 supports relaxed clock models and dating. Three different relaxed clock models are available: the Compound Poisson

Process (CPP; Huelsenbeck et al. 2000), the Thorne–Kishino 2002 (TK02; Thorne and Kishino 2002), and the Independent Gamma Rate (IGR; Lepage et al. 2007) models.

The CPP model is a discrete autocorrelated model, in which rate multipliers appear on the tree according to a Poisson process. The MrBayes implementation uses a lognormal distribution for the rate multipliers instead of the modified gamma distribution proposed originally (Huelsenbeck et al. 2000). It also includes novel algorithms to allow sampling across tree space since the original paper only dealt with fixed trees.

The TK02 model is a continuous autocorrelated model. In the particular version we implemented (Thorne and Kishino 2002), the rate of a descendant node is drawn from a lognormal distribution, the mean of which is the same as the ancestral rate and the variance of which is proportional to the length of the branch (measured in expected substitutions per site at the base rate of the clock).

The IGR model is a continuous uncorrelated model. First published as the "white noise" model (Lepage et al. 2007), it is similar to the uncorrelated gamma model (Drummond et al. 2006) but is mathematically more elegant in that it truly lacks time structure. In the IGR model, effective branch lengths are drawn from a gamma distribution, in which the mean is the same as, and the variance proportional to, the branch length.

Dating can be achieved in MrBayes 3.2 by calibrating interior or tip nodes in the tree; calibrated interior nodes need to be associated with hard constraints to be valid. Calibration points can be either fixed or associated with uncertainty. The birth–death prior model on clock trees has been expanded to incorporate recent progress in the understanding of the linear constant birth–death process with complete sampling (Gernhard 2008), with random incomplete sampling (Stadler 2009), or with clustered or diversified sampling (Höhna et al. 2011). The tree moves on clock and relaxed clock trees have also been improved considerably over those that were available in previous versions.

Bayesian phylogenetic inference of species trees from multiple gene trees was first accomplished in the Bayesian estimation of species trees (BEST) software using a complex computational machinery, in which MrBayes was one of the components (Edwards et al. 2007; Liu and Pearl 2007). Despite later improvements to BEST, the analyses remained slow and computationally demanding. The multispecies coalescent model has now been fully integrated in MrBayes 3.2, and several of the original algorithms have been rewritten to speed up the calculations.

Model Averaging and Model Choice

It is standard practice today to select a substitution model for Bayesian phylogenetic inference using a priori model selection procedures (Goldman 1993; Posada 1998, 2008; Suchard et al. 2001). An alternative is to use Bayesian model jumping during the MCMC simulation to integrate out the uncertainty concerning the correct substitution model (Huelsenbeck et al. 2004). The latter procedure is now implemented in MrBayes 3.2. Rather than selecting a substitution model before the analysis, the user can now sample across all 203 possible time-reversible rate matrices according to their posterior probability. The model-jumping approach is available in all models where a four-by-four nucleotide model is a component, including doublet and codon models in addition to the ordinary nucleotide models.

Bayesian model choice using Bayes factors is rapidly gaining in popularity. Since earlier versions, MrBayes has reported the harmonic mean of the likelihoods from the MCMC sample, which can be used as a rough estimate of the model likelihood from which the Bayes factor is calculated (Newton and Raftery 1994). However, there are now considerably more accurate, albeit computationally more demanding, methods (Lartillot and Philippe 2006). Of these, MrBayes 3.2 implements the recently proposed stepping stone method (Xie et al. 2011) that uses MCMC to sample from a series of so-called power posterior distributions connecting the posterior distribution with the prior distribution. The samples across these distributions are then used to estimate the model likelihood. The stepping stone algorithm in MrBayes 3.2 uses the full MCMC machinery, including convergence diagnostics and Metropolis coupling, and can be applied to any model available in the program. For instance, it can be used to test various topological hypotheses or substitution models against each other.

More Output Options

MrBayes 3.2 provides more extensive output options than previous versions. The user can now request sampling of site rates, site selection coefficients, site positive selection probabilities, and ancestral states of particular nodes. A wide range of tree statistics, including the mean and variance of split or clade frequencies, node times, and branch rates, are now added as annotations to the consensus tree by the "sumt" command and can be displayed using FigTree and compatible tree viewers.

BENCHMARK AND BIOLOGICAL EXAMPLES

Benchmark data on the GPU-accelerated code are provided by Ayres et al. (2012). A number of example data sets are distributed with the program, and tutorials illustrating most of the new features are included in the program manual. Many of the dating features in MrBayes 3.2 are discussed in some detail and used in an empirical context in Ronquist et al. (2012).

AVAILABILITY

MrBayes 3.2 is freely available under the GNU General Public License version 3.0. The program web site (http://www.mrbayes.net) provides download links to both source code for compilation on Unix systems and to convenient installers for Windows and Mac OS

systems. The installers include both MrBayes and the required BEAGLE libraries, but the BEAGLE libraries can also be installed separately using the BEAGLE installer, available at http://beagle-lib.googlecode.com. The program comes with a manual and example files. Further help is available on the program web site, which also provides instructions for reporting bugs and signing up for the MrBayes e-mail list. Instructions for accessing the MrBayes source code repository can be found at http://sourceforge.net/projects/mrbayes/develop.

REFERENCES

Altekar G., Dwarkadas S., Huelsenbeck J. 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics. 20:407–425.

Ayres D.L., Darling A., Zwickl D.J., Beerli P., Holder M.T., Lewis P.O., Huelsenbeck J.P., Ronquist F., Swofford D.L., Cummings M.P., Rambaut A., Suchard M.A. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61:170–173.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U.S.A. 104:5936–5941.

Gelman A., Rubin D. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Höhna S., Drummond A.J. 2012. Guided tree topology proposal for Bayesian phylogenetic inference. Syst. Biol. 61:1–11.

Höhna S., Stadler T., Ronquist F., Britton T. 2011. Inferring speciation and extinction rates under different species sampling schemes. Mol. Biol. Evol. 28:2577–2589.

Huelsenbeck J., Larget B., Swofford D. 2000. A compound Poisson process for relaxing the molecular clock. Genetics. 154:1879–1892.

Huelsenbeck J.P., Larget B., Alfaro M.E. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–1133.

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

Lakner C., van der Mark P., Huelsenbeck J., Larget B., Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol. 57:86–103.

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

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

Lepage T., Bryant D., Philippe H., Lartillot N. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–2680.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

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

Mau B., Newton M.A., Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. 55:1–12.

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

Posada D. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.

Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Roberts G., Rosenthal J. 2009. Examples of adaptive MCMC. J. Comput. Graph. Stat. 18:349–367.

Ronquist F., Deans A.R. 2010. Bayesian phylogenetics and its influence on insect systematics. Annu. Rev. Entomol. 55:189–206.

Ronquist F., Huelsenbeck J.P. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. Forthcoming 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol.

Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.

Suchard M.A., Weiss R.E., Sinsheimer J.S. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

Thorne J.L., Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51:689–702.

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

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