BIOINFORMATICS APPLICATIONS NOTE

Vol. 19 no. 12 2003, pages 1572–1574
DOI: 10.1093/bioinformatics/btg180

MrBayes 3: Bayesian phylogenetic inference under mixed models

Fredrik Ronquist[1,*] and John P. Huelsenbeck[2]

[1]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden and [2]Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0116, USA

Received on December 20, 2002; revised on February 14, 2003; accepted on February 19, 2003

ABSTRACT

Summary: MrBayes 3 performs Bayesian phylogenetic analysis combining information from different data partitions or subsets evolving under different stochastic evolutionary models. This allows the user to analyze heterogeneous data sets consisting of different data types—e.g. morphological, nucleotide, and protein—and to explore a wide variety of structured models mixing partition-unique and shared parameters. The program employs MPI to parallelize Metropolis coupling on Macintosh or UNIX clusters.

Availability: http://morphbank.ebc.uu.se/mrbayes.

Contact: fredrik.ronquist@ebc.uu.se

Computational complexity has long been a major obstacle in the development of statistical approaches to phylogenetic inference. Even moderate-sized empirical problems have posed serious challenges to computational biologists, forcing compromises in analytical accuracy. However, the recent introduction of Bayesian inference and Markov chain Monte Carlo (MCMC) techniques to phylogenetics has changed this situation. Early Bayesian phylogenetics papers showed that Markov chains based on the Metropolis–Hastings algorithm were computationally more efficient than the standard Maximum Likelihood (ML) bootstrapping approach (Larget and Simon, 1999). It is now known that problems with more than 350 sequences (taxa) can be analyzed successfully with moderate computational effort using Bayesian inference and an MCMC convergence acceleration technique known as Metropolis coupling (Huelsenbeck *et al.*, 2001). Such problems are set on tree spaces many orders of magnitude larger than those amenable to ML bootstrapping.

The increase in computational efficiency associated with the Bayesian MCMC approach makes it possible to analyze more complex and realistic evolutionary models than previously. Currently, an important but commonly invoked constraint on model complexity is the assumption of data homogeneity. Many phylogenetic data sets now include evidence from several different sources: morphology and molecules, amino acid and nucleotide data, or sequences from the mitochondrial, plastid and nuclear genomes. However, the available software commonly forces the investigator to either: (1) model the evolution of such data using a single stochastic model; (2) analyze the different data partitions or subsets separately and use *ad hoc* methods to obtain a summary result; or (3) resort to simple search algorithms or non-statistical methods. None of these alternatives is particularly attractive.

MrBayes 3 is a completely rewritten and restructured version of MrBayes, a command-driven program for Bayesian phylogenetic inference (Huelsenbeck and Ronquist, 2001). The hallmark of the new program is a powerful framework for phylogenetic inference under mixed models accommodating data heterogeneity. This framework will help the user to specify mixed models and exploit the computational efficiency of Bayesian MCMC analysis in dealing with composite data sets.

Bayesian phylogenetic inference is based on Bayes's rule. Applied to the phylogeny problem, the rule can be expressed as follows

$$f(\tau, v, \theta|X) = \frac{f(\tau, v, \theta)\, f(X|\tau, v, \theta)}{f(X)}$$

where X is the data matrix, τ is the topology of the tree, v is a vector of branch (or edge) lengths on the tree, and θ is a vector of substitution model parameters. The distribution $f(\tau, v, \theta)$ is referred to as the *prior*, and specifies the prior probability of different parameter values; $f(X|\tau, v, \theta)$ is the *likelihood function*, describing the probability of the data under different parameter values; and $f(X)$ is the total probability of the data summed and integrated over the parameter space. Bayesian inference is based on the so-called *posterior distribution* $f(\tau, v, \theta|X)$.

Typically, it is not possible to calculate the posterior probability distribution analytically; instead, MCMC tech-

[*]To whom correspondence should be addressed.

niques are used to obtain samples from it. MrBayes 3 uses a Metropolis–Hastings sampler and updates single parameters or blocks of related parameters in each move. Assume that, in the current generation, the Markov chain has parameter values τ, v, and θ and that we are considering a change in θ to the new value θ^* picked from some proposal distribution $q(\theta^*|\theta)$. Then we accept the change with probability

$$r = \min\left(1, \frac{f(\theta^*)}{f(\theta)} \frac{f(X|\tau, v, \theta^*)}{f(X|\tau, v, \theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right)$$

Bayesian inference is easily extended to deal with heterogeneous models. Assume that we have two subsets of our data, X_a and X_b, such that $X = X_a + X_b$. Assume further that it is likely that these data subsets evolved on the same phylogenetic tree but according to entirely different substitution models with parameters θ_a and θ_b, respectively, such that $\theta = \theta_a + \theta_b$. Bayes's rule now becomes:

$$\begin{aligned} f(\tau, v, \theta_a, \theta_b|X) &= \frac{f(\tau, v, \theta_a, \theta_b)\, f(X|\tau, v, \theta_a, \theta_b)}{f(X)} \\ &= [\, f(\tau, v, \theta_a, \theta_b)\, f(X_a|\tau, v, \theta_a) \\ &\quad \times f(X_b|\tau, v, \theta_b)] \big/ [\, f(X)] \end{aligned}$$

During a MCMC run, we might consider a change of the substitution model parameters affecting one of the subsets, for instance the subset X_a. Then we have a potential change from θ_a to θ_a^* and the jumping kernel of the Metropolis–Hastings sampler simplifies to

$$r = \min\left(1, \frac{f(\theta_a^*)}{f(\theta_a)} \frac{f(X_a|\tau, v, \theta_a^*)}{f(X_a|\tau, v, \theta_a)} \frac{q(\theta_a|\theta_a^*)}{q(\theta_a^*|\theta_a)}\right)$$

Because we only need to calculate the likelihood ratio (the second ratio in the product) for the affected data subsets, the run will actually be *faster* (per generation) under a mixed than under a homogeneous model. Of course, since there are more parameters in the mixed model, we will be visiting each parameter more rarely and this is likely to lead to slower mixing and a need to run the chain longer before an adequate sample of the posterior distribution is obtained. The net effect is dependent on the particulars of the analysis but there is no obvious reason why the computational complexity would necessarily be much worse under a parameter-rich mixed model than under a homogeneous model for the same data set. MrBayes 3 now provides the tools needed to examine this question empirically.

MrBayes 3 implements a wide variety of stochastic models for nucleotide, protein, restriction site, and morphological (standard) data. Single, doublet (for stem regions) and codon (with or without variation in the non-synonymous/synonymous rate ratio across sites) models are all available for nucleotide data. Protein data can be analyzed using a range of fixed or variable rate matrices. Both the standard and restriction site models can include correction for coding biases. The standard model, appropriate for analysis of morphological data, allows up to ten different states and there are models for both unordered and linearly ordered characters. Rate variation across sites can be accommodated using a standard gamma or, for nucleotide and protein models, an autocorrelated gamma distribution. For nucleotide and protein data, it is possible to allow rate variation across the tree using a variant of the covarion/covariotide model, in which sites independently switch between an ‘on’ and an ‘off’ state (Huelsenbeck *et al.*, 2001; Huelsenbeck and Ronquist, 2001). Available tree models include unconstrained, standard molecular clock, birth-and-death, and coalescent models.

MrBayes 3 reads data from a text file conforming to a modified NEXUS format allowing mixed data sets. By default, the data are partitioned according to data type. The user can further subdivide the data, most easily by specifying character sets. Once the data set has been partitioned appropriately, the user can set the model structure and priors of each data subset. Individual model parameters can be unlinked or linked across selected data subsets. The currently active subsets and the model parameters applying to these subsets can be listed, giving the user a means of checking that the mixed model is specified correctly.

In addition to allowing heterogeneity across data subsets in overall rate and in substitution model parameters, MrBayes 3 also allows the user to unlink topology and branch lengths. Different data subsets can thus have independent branch lengths or even different topologies.

Correct scaling of rate parameters is important in mixed models. MrBayes 3 scales rates such that branch lengths are measured in the expected number of changes per site. For instance, a change in a codon model is counted as one change per three sites. Partition-specific rate multipliers are scaled such that the mean rate per site across partitions is unity.

MrBayes 3 provides many options for summarizing and diagnosing the results of an MCMC analysis. Simple plots of overall likelihood and individual parameter values can be generated to determine burn-in and examine mixing, and the program will also estimate the model likelihood, used in Bayesian model testing. The parameter file is a tab-delimited text file, which can be imported into and analyzed with most standard statistical software packages. The program will summarize trees and branch lengths in the form of consensus trees, partition tables, and lists of trees with their estimated posterior probability. Consensus trees are written with both branch lengths and posterior clade probabilities, for easy graphical representation using software such as TreeView (Page, 1996).

By default, MrBayes 3 uses Metropolis coupling to accelerate convergence of the Markov chain (Huelsenbeck and Ronquist, 2001). Because of the relatively small amount of information communicated among the multiple chains during such a run, Metropolis coupling is well suited to parallel implementations in which chains are distributed among processors. Using the Message-Passing Interface (MPI), this type of parallelization has been implemented in MrBayes 3 for UNIX and Macintosh clusters. With large data sets, near linear speed-ups can be achieved using this approach (Altekar *et al.*, 2003).

MrBayes 3 is written in ANSI C and is available free of charge from http://morphbank.ebc.uu.se/mrbayes/. The site provides precompiled versions for the MacOS and Windows platforms, and the source code for compilation on UNIX machines. The MPI-enabled parallel version of MrBayes 3 is available both precompiled for Macintosh OS X and through setting the relevant compiler switch before compilation for UNIX. The parallel Macintosh version requires installation of POOCH (http://www.daugerresearch.com/pooch/whatis.html) on all participating machines.

ACKNOWLEDGEMENTS

We are deeply indebted to the beta testers and users of previous versions of MrBayes for identifying bugs and proposing various improvements. F.R. was supported by the Swedish Research Council grant 621-2001-2963 and J.P.H. by NSF grants DEB-0075406 and MCB-0075404.

REFERENCES

Altekar *et al.* (2003) Parallel metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, in press.

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

Huelsenbeck,J.P., Ronquist,F., Nielsen,R. and Bollback,J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

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

Page,R.D.M. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, **12**, 357–358.