BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001
Pages 754–755

MRBAYES: Bayesian inference of phylogenetic trees

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

[1]Department of Biology, University of Rochester, Rochester, NY 14627, USA and [2]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden

Received on January 24, 2001; revised on March 23, 2001; accepted on March 28, 2001

ABSTRACT

Summary: The program MRBAYES performs Bayesian inference of phylogeny using a variant of Markov chain Monte Carlo.

Availability: MRBAYES, including the source code, documentation, sample data files, and an executable, is available at http://brahms.biology.rochester.edu/software.html.

Contact: johnh@brahms.biology.rochester.edu

In the *Origin of Species*, Darwin founded evolutionary biology on the idea that organisms share a common origin and have subsequently diverged through time. Phylogenies represent our attempts to reconstruct the evolutionary history of life and our ability to infer phylogeny has increased dramatically in the past decade; not only has it become relatively easy to quickly determine the DNA sequence of a gene, but computers have increased substantially in speed. Concomitant with the improvements in data and computers, a large number of methods have been proposed to infer phylogenetic trees, including the parsimony method, a number of distance methods, and maximum likelihood. Only recently, however, has Bayesian inference of phylogeny been proposed (Li, 1996; Mau, 1996; Mau and Newton, 1997; Mau *et al.*, 1999; Rannala and Yang, 1996; Yang and Rannala, 1997). Bayesian inference has several advantages over other methods of phylogenetic inference, including easy interpretation of results, the ability to incorporate prior information (if such information is available), and some computational advantages (see Larget and Simon, 1999).

In a Bayesian analysis, inferences of phylogeny are based upon the posterior probabilities of phylogenetic trees. The posterior probability of the ith phylogenetic tree (τ_i) conditional on an alignment of DNA sequences ($\boldsymbol{X}$) can be calculated using Bayes theorem:

$$f(\tau_i|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\tau_i)f(\tau_i)}{\sum_{j=1}^{B(s)} f(\boldsymbol{X}|\tau_j)f(\tau_j)}$$

where

$$f(\boldsymbol{X}|\tau_i) = \int_{\boldsymbol{v}}\int_{\theta} f(\boldsymbol{X}|\tau_i, \boldsymbol{v}, \boldsymbol{\theta}) f(\boldsymbol{v}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{\theta}.$$

The summation is over all $B(s)$ trees that are possible for s species [$B(s) = \frac{(2s-5)!}{2^{s-3}(s-3)!}$ for unrooted trees and $B(s) = \frac{(2s-3)!}{2^{s-2}(s-2)!}$ for rooted trees], and integration is over all combinations of branch lengths ($\boldsymbol{v}$) and substitution parameters ($\boldsymbol{\theta}$). The prior for phylogenetic trees is $f(\tau_i)$, and is usually set to $f(\tau_i) = \frac{1}{B(s)}$. The prior on branch lengths and substitution parameters is denoted $f(\boldsymbol{v}, \boldsymbol{\theta})$. Typically, the likelihood function [$f(\boldsymbol{X}|\tau_i, \boldsymbol{v}, \boldsymbol{\theta})$] is calculated under the assumption that substitutions occur according to a time–homogeneous Poisson process. The same models of DNA substitution used in maximum likelihood analyses (Swofford *et al.*, 1996) can be used in a Bayesian analysis of phylogeny.

The summation and integrals required in a Bayesian analysis cannot be evaluated analytically. MRBAYES uses Markov chain Monte Carlo (MCMC) to approximate the posterior probabilities of trees (Metropolis *et al.*, 1953; Hastings, 1970; Green, 1995). MCMC is a method for taking valid, albeit dependent, samples from the probability distribution of interest (in this case, the posterior probabilities of phylogenetic trees; Tierney, 1994). The basic MCMC algorithm works as follows: first, a new state for the chain is proposed using a stochastic mechanism. Second, the acceptance probability for this new state is calculated. The acceptance probability is equal to the minimum of one or the likelihood ratio times, the prior ratio times, the proposal ratio, where the likelihood ratio is the ratio of the likelihoods of the new state to the old state, the prior ratio is the ratio of the prior probability of the new state to the old state, and the proposal ratio is the ratio of the probability of proposing the old state to the probability of proposing the new state. Third, a uniform (0, 1) random variable is drawn. If this number is less than the acceptance probability, then the new state is accepted and the state of the chain is updated. Otherwise the chain

remains in the old state. This process of proposing and accepting/rejecting new states is repeated many thousands or millions of times. The proportion of the time any single tree is visited during the course of the chain is a valid approximation of its posterior probability.

MRBAYES not only implements the standard MCMC algorithm, but also a variant of MCMC called Metropolis-coupled Markov chain Monte Carlo [or $(MC)^3$ for short; Geyer, 1991]. $(MC)^3$ runs n chains, $n-1$ of which are heated. A heated chain has the steady-state distribution $f(\tau_i \mid \boldsymbol{X})^{\beta}$. MRBAYES uses incremental heating, where the heat applied to the ith chain is $\beta = \frac{1}{1+(i-1)T}$ and T is a heating parameter that must be set by the user. After all n chains have gone one step, a swap is attempted between two randomly chosen chains. If the swap is accepted, then the two chains switch states. Inferences are based only on the states sampled by the cold chain ($\beta = 1$). The heated chains can more easily explore the space of phylogenetic trees; the effect of heating is to lower peaks and to fill in valleys. Importantly, the cold chain can effectively leap across deep valleys in the landscape of trees when a successful swap is made between the cold chain (perhaps stuck on a local optimum of trees) and a heated chain that is exploring another peak. Experience has shown that mixing of the chain is dramatically improved using $(MC)^3$.

MRBAYES has a command-line interface. The program reads in an aligned matrix of DNA or amino acid sequences in the standard NEXUS format (Maddison *et al.*, 1997). The user can change assumptions of the substitution model, the prior, and the details of the $(MC)^3$ analysis on the fly. Moreover, the user can delete and restore taxa and characters in the analysis. The program implements the most general 4×4 model of DNA substitution possible (the general non-reversible model; Yang, 1994), a number of 20×20 models of amino acid substitution, and codon models of DNA substitution. The program also implements several methods for relaxing the assumption of equal rates across sites, including gamma-distributed rate variation (Yang, 1993). Finally, the program can infer ancestral states while accommodating uncertainty about the phylogenetic tree and model parameters.

ACKNOWLEDGEMENTS

This research was supported by NSF grant DEB-0075406.

REFERENCES

Geyer,C.J. (1991) Markov chain Monte Carlo maximum likelihood. In Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.*. Interface Foundation, Fairfax Station, pp. 156–163.

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

Hastings,W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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

Li,S. (1996) *Phylogenetic tree construction using Markov chain Monte carlo*, PhD Dissertation, Ohio State University, Columbus.

Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621.

Mau,B. (1996) *Bayesian phylogenetic inference via Markov chain Monte carlo methods*, PhD Dissertation, University of Wisconsin, Madison.

Mau,B. and Newton,M. (1997) Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.*, **6**, 122–131.

Mau,B., Newton,M. and Larget,B. (1999) Bayesian phylogenetic inference via Markov chain Monte carlo methods. *Biometrics*, **55**, 1–12.

Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

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

Swofford,D., Olsen,G., Waddell,P. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis, Moritz and Mable (eds), *Molecular Systematics, 2nd edition.*. Sinauer, Sunderland, MA, pp. 407–511.

Tierney,L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.*, **22**, 1701–1762.

Yang,Z. (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10**, 1396–1401.

Yang,Z. (1994) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, **39**, 105–111.

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