Bioinformatics
BIO2002/268
Accepted November 26, 2002 – Second revision December 02, 2002

Supplementary Information:

How Bayes tests of molecular phylogenies compare with frequentist approaches

Stéphane Aris-Brosou

Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England.

Present address:

Stéphane Aris-Brosou, Bioinformatics Research Center, Box 7566, North Carolina State University, Raleigh NC, 27695-7566, USA.

Phone: (919) 513-1588

Fax: (919) 515-7315

Email: stephane@statgen.ncsu.edu

ABSTRACT

Motivation: The desire to compare molecular phylogenies has stimulated the design of numerous tests. Most of these tests are formulated in a frequentist framework, and it is not known how they compare with Bayes procedures. I propose here two new Bayes tests that either compare pairs of trees (Bayes hypothesis test, BHT), or test each tree against an average of the trees included in the analysis (Bayes significance test, BST).

Results: The algorithm, based on a standard Metropolis–Hastings sampler, integrates nuisance parameters out and estimates the probability of the data under each topology. These quantities are used to estimate Bayes factors for composite vs. composite hypotheses. Based on two data sets, the BHT and BST are shown to construct similar confidence sets to the bootstrap and the Shimodaira Hasegawa test, respectively. This suggests that the known difference among previous tests is mainly due to the null hypothesis considered.

Contact: stephane@statgen.ncsu.edu

Availability: http://statgen.ncsu.edu/stephane/

Supplementary Information: This supplemental material is available at http://statgen.ncsu.edu/stephane/bayestests.htm

COMPUTATIONAL DETAILS

The Bayes tests: BST and BHT

Posterior distributions of trees were approximated using a Markov chain Monte Carlo (MCMC) algorithm based on the Metropolis–Hastings sampler (e.g. Gilks et al, 1996). Prior distributions were chosen to be non-informative as described in Huelsenbeck and Imennov (2002). Parameters θ of the evolutionary process (branch lengths, transition-transversion rate ratio, base frequencies and the shape parameter of the gamma distribution modelling among-site rate variation) were drawn from normal proposal densities centred on the current parameter values (e.g., Aris-Brosou and Yang, 2002). The tree space was explored with the NNI algorithm (e.g. Yang and Rannala, 1997), and trees were indexed with respect to a left-right ordering of the leaves giving unique in-order transversal representations (Yang, 1997; Larget and Simon, 1999). At a given step of the MCMC where the chain is in state φ, a new state φ^* is proposed, consisting either in an update of one of the components of θ (branch length, etc.), or in a topology change. The proposed state is then accepted with probability:

$$\min\{1, p(\varphi^* \mid X)/p(\varphi \mid X)\} \qquad \text{(SI-1)}$$

For each MCMC run, the 10,000 first steps of the chain were discarded (burn-in) and the chain was then sampled every 100 steps until 10,000 samples were collected. Four chains starting from different initial values were run to check convergence as follows. For each run, time series outputs of each parameter (θ) were analysed and consistency of the estimates across the different runs, as well as with their maximum likelihood estimates, was checked. The posterior probability of tree T_i is then estimated by the proportion of trees T_i sampled at stationarity.

Unlike the computation of posterior probabilities, the algorithm estimating the BF keeps the topology constant, as specified by (5) (see main text). The integration is carried out by the

same MCMC implementation as above. Computation of (2) (see main text) is difficult and cannot be carried out directly. Instead, a general means of estimating integrals of the form:

$$I = \int g(\theta)\ p(\theta)\ d\theta \qquad \text{(SI-2)}$$

is to generate a sample from a distribution $p^*(\theta)$ and calculate the simulation consistent estimator:

$$I = \sum w_i\ g(\theta) \Big/ \sum w_i \qquad \text{(SI-3)}$$

where w_i is the importance sampling function $p(\theta) \big/ p^*(\theta)$. By taking:

$$\begin{cases} g(\theta) = p(X \mid \theta) \\ p^*(\theta) = p(\theta \mid X) = p(X \mid \theta)\ p(\theta) / p(X) \end{cases} \qquad \text{(SI-4)}$$

at a given T_i and substituting into (SI-3), the probability of the data under a given model tree T_i (2) (see main text) can be estimated by the harmonic mean of the likelihood sampled from the posterior distribution (Raftery, 1996):

$$\hat{p}(X \mid T_i) = \left(\frac{1}{N} \sum_t^N 1 \Big/ p(X \mid \theta_i^t) \right)^{-1} \qquad \text{(SI-5)}$$

where θ_i^t is sampled from the posterior distribution at the t^{th} step along the MCMC, and N is the number of steps sampled for inference. This estimator proved here to be stable across independent MCMC runs. Although exponentiation of the log-probabilities $\log\{p(X \mid \theta_i^t)\}$ could be achieved via a scaling procedure to avoid underflows (e.g., Yang, 1997), it was here taken with Mathematica®. The computer program implementing the algorithms described as well as the perl and Mathematica® scripts used are available at http://statgen.ncsu.edu/stephane/.

The frequentist tests: BP, SH and SOWH

Besides the BST and the BHT, I also compute and refer in the main text to the BP, estimated by RELL (Kishino et al., 1990), as well as *p*-values of the SOWH test and the SH test were computed. 10,000 replicates were used to estimate BP and the *p*-values of the SH test. 500 replicates were simulated with evolver (Yang, 1997) under each tree T_i tested for by the SOWH test. The "*posPpud* approximation under H_A" (Goldman et al., 2000) was used to lessen the computational burden: This saves on estimating parameters of the substitution model under the alternative hypothesis that a tree other than T_i is correct.

ACKNOWLEDGMENTS

I thank David Balding, Joe Bielawski, Lounès Chikhi, Nick Goldman, Jeff Thorne, Ziheng Yang and three anonymous reviewers for constructive comments. This work was funded by a Biotechnological and Biological Sciences Research Council grant to Ziheng Yang and a National Science Foundation grant DEB-0120635 to Jeff Thorne.

REFERENCES

Aris-Brosou,S. and Yang,Z. (2002) The effects of models of rate evolution on estimation of divergence dates with a special reference to the metazoan 18S rRNA phylogeny. *Syst. Biol.*, **51,** 703–714.

Gilks,W.R., Richardson,S., and Spiegelhalter,D.J. (1996) *Markov chain Monte Carlo in practice.* Chapman & Hall, Boca Raton.

Goldman,N., Anderson,J.P., and Rodrigo,A.G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.*, **49,** 652–670.

Huelsenbeck,J.P. and Imennov,N.S. (2002) Geographic origin of human mitochondrial DNA: accommodating phylogenetic uncertainty and model comparison. *Syst. Biol.*, **51,** 155–165.

Kishino,H., Miyata,T., and Hasegawa,M. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.*, **30,** 151–160.

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

Raftery,A.E. (1996) Hypothesis testing and model selection. In Gilks,W.R., Richardson,S., and Spiegelhalter,D.J. (eds), *Markov chain Monte Carlo in practice*. Chapman & Hall, Boca Raton, pp. 163–187.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13,** 555–556.

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