Phylogenetics

AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics

Johan A.A. Nylander[1], James C. Wilgenbusch[1,*], Dan L. Warren[2] and David L. Swofford[1]

[1]School of Computational Sciences, Florida State University, Tallahassee, Florida 32306 and [2]Section of Evolution and Ecology, University of California, Davis, California 95616, USA

ABSTRACT

Summary: A key element to a successful Markov chain Monte Carlo (MCMC) inference is the programming and run performance of the Markov chain. However, the explicit use of quality assessments of the MCMC simulations—convergence diagnostics—in phylogenetics is still uncommon. Here, we present a simple tool that uses the output from MCMC simulations and visualizes a number of properties of primary interest in a Bayesian phylogenetic analysis, such as convergence rates of posterior split probabilities and branch lengths. Graphical exploration of the output from phylogenetic MCMC simulations gives intuitive and often crucial information on the success and reliability of the analysis. The tool presented here complements convergence diagnostics already available in other software packages primarily designed for other applications of MCMC. Importantly, the common practice of using trace-plots of a single parameter or summary statistic, such as the likelihood score of sampled trees, can be misleading for assessing the success of a phylogenetic MCMC simulation.

Availability: The program is available as source under the GNU General Public License and as a web application at http://ceb.scs.fsu.edu/awty

Contact: jwilgenb@scs.fsu.edu

1 INTRODUCTION

Despite the growing popularity of MCMC methods in phylogenetics, the use of MCMC convergence diagnostics is still relatively uncommon. Tools for assessing convergence are already available for many statistical models (e.g. Plummer *et al.*, 2005) but they are rarely used in phylogenetic studies [a notable exception is the Tracer software (Ramber and Drummond, 2004) designed for analyzing time-series plots of substitution model parameters]. This is probably due to the fact that convergence diagnostics for parameters specific to phylogenetic trees, such as splits and branch lengths, are few and their performance relatively unexplored.

The difficulties involved with diagnosing convergence in MCMC inference are well documented in the statistical literature (e.g. Brooks and Gelman, 1998; Geweke, 1992) and application of MCMC to Bayesian phylogenetics is no exception. As an example, the most frequently used method for assessing convergence in the phylogenetic literature involves examining trace plots of the likelihood scores for trees sampled by the Markov chain. This approach can, however, be misleading for diagnosing convergence (or lack thereof) as illustrated in the upper row of Figure 1. The plot in the first column shows the output from a Bayesian MCMC simulation where the likelihood trace for two independent runs reaches the same level of apparent stationarity. Posterior probabilities of splits continue, however, to change over the length of the simulation (Fig. 1, third column).

This example emphasizes the fact that trees with similar likelihoods are not necessarily close in parameter (tree) space and judging the success of a MCMC from the likelihood trace alone might lead to inaccurate and misleading results (Huelsenbeck *et al.*, 2002; Nylander *et al.*, 2004). It also emphasizes that using a range of MCMC diagnostics is important and that graphical exploration of tree-specific parameters is a crucial complement to existing diagnostics tools and should routinely be applied in phylogenetic analyses using MCMC.

2 THE AWTY PROGRAM

2.1 Program features

The AWTY program takes as input the phylogenetic trees generated as output by other phylogenetic MCMC programs; MrBayes (Ronquist and Huelenbeck, 2002), BEAST (Drummond and Rambaut, 2006) and BAMBE (Simon and Larget, 2000) formats are currently supported. A number of diagnostic analyses can then be performed on the trees and visualized graphically. The main focus is on splits or clades and Figure 1 shows some examples where properties related to splits are compared within and among separate MCMC runs of the same data. Other features available are, e.g. Geweke's diagnostic (Geweke, 1992) and Brooks and Gelman's $\widehat{R}$-interval diagnostic (Brooks and Gelman, 1998) for branch lengths.

Many of the diagnostics implemented in AWTY are based on a *post hoc* approach where the output from a MCMC analysis is

---

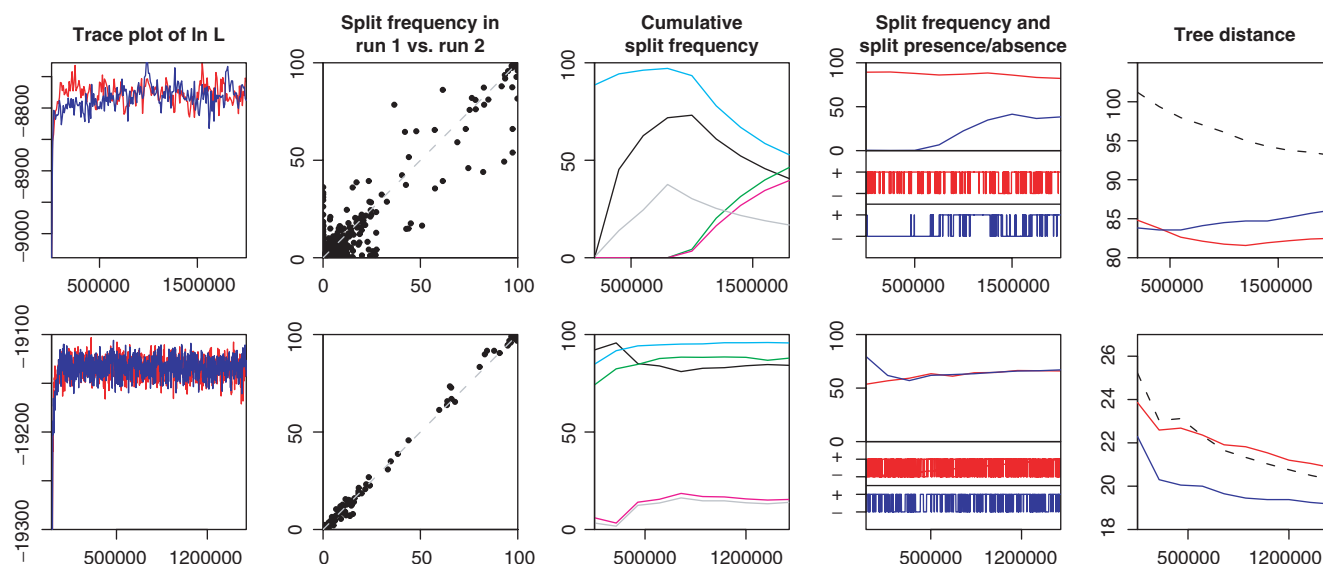*To whom correspondence should be addressed.

Fig. 1. Examples of output from AWTY. The figure shows the graphical exploration of the output from two different DNA data sets (upper row: 86 sequences, lower row: 62 sequences) analyzed in MrBayes v.3.1.2 (Ronquist and Huelsenbeck, 2002). Two separate simulations (indicated by red and blue colors) were run for each data set using the GTR+I+Γ model. The first graph shows the trace plot of the log likelihood and the sampled values reach apparent stationarity for both simulations in both data sets. The second graph is a bivariate plot of the split frequencies for the first and second run of the simulations. A low correlation (upper row) diagnoses lack of convergence. The third graph shows the cumulative split frequency for a number of selected splits for one of the individual simulations. A trend in frequencies (upper row) diagnoses lack of convergence. The fourth graph shows the cumulative split frequency (upper part) and the corresponding presence (+) and absence (−) for a single split (lower part) over the two simulations. A slow mixing—where the chain moves slowly in parameter space—resulting in a trend in the cumulative split frequency is apparent in the upper row. The last graph compares the symmetric tree-difference score (Penny and Hendy, 1985) within and between (dashed line) simulations. A between-run distance well separated from the within-run distance (upper row) diagnoses lack of convergence.

examined and compared over replicated runs. The underlying assumption is that simulations started from independent starting values should have similar properties at convergence (Brooks and Gelman, 1998). It must be emphasized, however, that this approach cannot guarantee convergence *per se* but is primarily a method for diagnosing lack of convergence in one (or several) runs. Furthermore, the success of the *post hoc* approach is dependent on the number of individual runs and the performance and behavior of each individual run. The information on the latter, such as proposal/acceptance ratios, should be included in the overall assessment of the success of an MCMC simulation and should be the focus in further research on MCMC applications in Bayesian phylogenetics.

2.2 Implementation details

The main routine in AWTY is written in Perl and uses the program PAUP* (Swofford, 2003) for handling phylogenetic trees. The graphical output is generated by GNUPLOT (Williams and Kelly, 2006). For some of the convergence diagnostics, AWTY uses the CODA package (Plummer *et al.*, 2005) written in R (R Development Core Team, 2006) through the R-from-Perl interface RSPerl (Temple-Lang, 2006). The program can be run using either a command-line UNIX-type interface, or via a Gtk2/Tk interface provided by the Perl modules Getopt:GUI:Long and QWizard

(Hardaker, 2006). In addition, the program comes with a web interface written in PHP4 and runs on any web server such as Apache.

REFERENCES

Brooks,S.P. and Gelman,A. (1998) General methods for monitoring convergence of iterative simulations. *J. Comput. Graphi. Stat.*, **7**, 434–455.

Drummond,A.J. and Rambaut,A. (2006) BEAST v1.4.http://beast.bio.ed.ac.uk/

Geweke,J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernado,J.M. *et al.* (eds.) *Bayesian Statistics 4*. Clarendon Press, Oxford UK.

Hardaker,W. (2006) Getopt:GUI:Long Version 0.62, and QWizard Version 3.03. http://www.cpan.org

Huelsenbeck,J.P. *et al.* (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.*, **51**, 673--688.

Nylander,J.A.A. *et al.* (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, **53**, 47–67.

Penny,D. and Hendy,M.D. (1985) The use of tree comparison metrics. *Systematic Zool.*, **34**, 75--82.

Plummer,M. *et al.* (2006) CODA: output analysis and diagnostics for Markov Chain Monte Carlo simulations. http://cran.Rproject.org

R Development Core Team (2006) R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rambaut,A. and Drummond,A.J. (2004) Tracer v1.3http://evolve.zoo.ox.ac.uk/software.html

Ronquist,F. and Huelsenbeck,J.P. (2003) MRBAYES 3: Bayesian phylogenetic-inference under mixed models. *Bioinformatics*, **19**, 1572–-1574.

Simon,D. and Larget,B. (2000) Bayesian analysis in molecular biology and evolution (BAMBE), version 2.03 beta Department of Mathematics and Computer Science, Duquesne University.

Swofford,D.L. (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts, USA.

Temple-Lang,D. (2006) RSPerl Version 0.83. http://www.omegahat.org/RSPerl

Williams,T. and Kelley,C. (2006) GNUPLOT, an interactive plotting program, Version 4.0. http://www.gnuplot.info