BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 16 2006, pages 2047–2048
doi:10.1093/bioinformatics/btl175

Phylogenetics

BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny

Marc A. Suchard[1,2,*] and Benjamin D. Redelings[3]

[1]Department of Biomathematics and [2]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA and [3]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606, USA

Received on February 17, 2006; revised on April 28, 2006; accepted on May 1, 2006
Advance Access publication May 5, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Summary: BAli-Phy is a Bayesian posterior sampler that employs Markov chain Monte Carlo to explore the joint space of alignment and phylogeny given molecular sequence data. Simultaneous estimation eliminates bias toward inaccurate alignment guide-trees, employs more sophisticated substitution models during alignment and automatically utilizes information in shared insertion/deletions to help infer phylogenies.

Availability: Software is available for download at http://www.biomath.ucla.edu/msuchard/bali-phy.

Contact: msuchard@ucla.edu

INTRODUCTION

Phylogenetic methods to reconstruct the evolutionary tree relating molecular sequence data traditionally condition on a single, sometimes poorly estimated multiple sequence alignment (Holder and Lewis, 2003). This alignment specifies which residues across the sequences are derived from a common origin. Conditioning on a poor alignment derived from an inaccurate guide-tree can cause bias and inappropriate inference in evolutionary studies (Lake, 1991). This concern is particularly poignant for highly diverse sequences, where the complete alignment is not obvious. We provide a novel Bayesian program **BAli-Phy** that simultaneously estimates the alignment and phylogenetic tree that relate molecular sequences. This sidesteps the bias issue inherent in sequential estimation.

SOFTWARE OVERVIEW

Redelings and Suchard (2005) introduce a joint estimation model for alignment and phylogeny and describe a Markov chain Monte Carlo (MCMC) approach to generate random samples from the joint model posterior. We briefly review the salient features of the model here. Conditional on a given alignment, the model employs standard continuous-time Markov chain (CTMC) processes to describe residue substitution along the branches of an unknown tree relating the sequences. To remove this conditioning and treat alignments as unknown parameters, the model further assumes a prior distribution over all possible alignments. We construct this distribution from a set of hidden Markov models with affine gap penalties that describe the pairwise alignments along each branch of the tree.

To generate posterior samples, **BAli-Phy** employs a Metropolis-within-Gibbs (Tierney, 1994) approach. We construct the random-scan Gibbs cycle from straightforward Metropolis-Hastings proposals for updating branch lengths and substitution and indel parameters and several unique steps for updating the alignment and topology. These latter steps rely simultaneously on subtree transfer operators and dynamic programming through the Forward–Backward algorithm (Scott, 2002) and extend the work of Holmes and Bruno (2001) to provide good convergence properties to the sampler.

BAli-Phy also contains a number of tools to summarize the joint posterior samples. The most important among these is **Alignment-gild** that produces alignment uncertainty (AU, pronounced 'gold') plots. AU plots depict an estimate of the maximum a posteriori alignment annotated to identify features (residues or gaps) with positional variability in the posterior samples. **Alignment-gild** renders these plots in HTML for cross-platform viewing.

BAli-Phy can also sample from more traditional phylogenetic models in which the alignment of the leaf sequences is fixed. When this alignment is fixed, users have the option of including the indel process prior and sampling the alignment states of the internal nodes or excluding this prior completely. Indels shared by common descent will influence the posterior when the indel prior is included, while excluding the prior results in a residue-substitution-only model such as that sampled by MrBayes (Huelsenbeck and Ronquist, 2001).

BAli-Phy currently implements several CTMC processes for residue-substitution, including the JC69, HKY85 and TN93 nucleotide models, several codon-based models and empirically estimated amino acid models. Gamma-distributed rate variation and invariant sites extensions are available. Sequential estimation generally assumes a naive model for residue-substitution during the alignment phase. In contrast, joint estimation may employ any of these more sophisticated processes to inform the alignment. We distribute **BAli-Phy** as **C++** source code and precompiled binaries. **BAli-Phy** should run on all hardware with a modern operating system such as **Linux**, **Windows** or **Mac OS X**.

EXAMPLE

Considerable debate surrounds the early history of life on Earth. Molecular sequences across the Tree of Life are highly diverse and troublesome to align. Phylogenies based on some arbitrary alignments suggest three major domains of all organisms: Eubacteria, Archaea and Eukaryotes; other alignments support four, in

*To whom correspondence should be addressed.

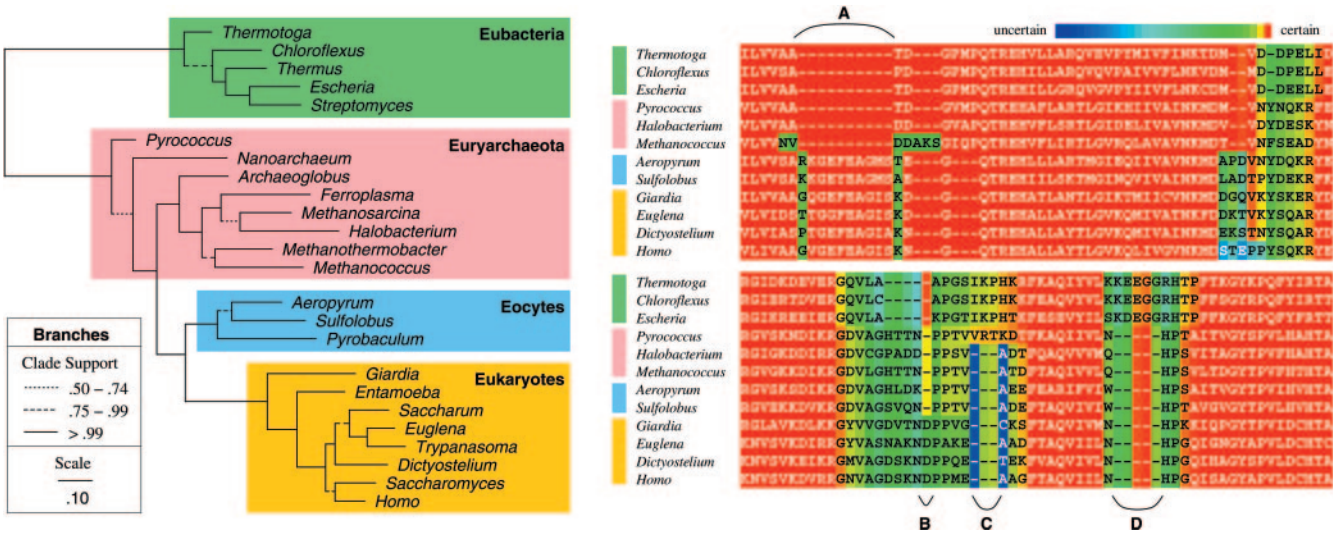

Fig. 1. Maximum a posteriori topology for 24 EF-1α/Tu sequences across the Tree of Life (left) and two separate portions of the alignment uncertainty (AU) plot (right) for a subsample of these sequences. Branch lengths equal posterior mean estimates and line-style depicts partition credibility. In the AU plot, well-resolved entries have a red background, whereas less certain entries have backgrounds tending towards violet based on an approximate probability that each entry is homologous with a residue at the root in each column. Four different types of topologically informative insertion/deletion events (A, B, C and D) are highlighted.

which the Archaea are subdivided into Euryarchaeota and Eocytes (Brown and Doolittle, 1997). We examine elongation factor 1α/Tu sequences from 24 species using the WAG+Γ+INV substitution model. Figure 1 presents the most probable evolutionary tree relating these sequences and a portion of the AU plot. This highly supported tree divides life into at least four domains, furthering the Eocyte hypothesis unconditional on alignment. The shared indel event labeled 'A' in the figure has been hypothesized previously (Rivera and Lake, 1992) and adds strength to the conclusion.

ACKNOWLEDGEMENTS

B.D.R. was supported by NIH training grant GM008185 and NSF training grant DGE9987641. M.A.S. is partially supported by NIH grant GM068955 and USPHS grant CA16042.

Conflict of Interest: none declared.

REFERENCES

Brown,J. and Doolittle,W. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.*, **61**, 456–502.

Holder,M. and Lewis,P. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, **4**, 275–284.

Holmes,I. and Bruno,W. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**, 803–820.

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

Lake,J. (1991) The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.*, **8**, 378–385.

Redelings,B. and Suchard,M. (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, **54**, 401–418.

Rivera,M. and Lake,J. (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science*, **257**, 74–76.

Scott,S. (2002) Bayesian methods for hidden Markov models, recursive computing in the 21st century. *J. Am. Stat. Asso.*, **97**, 337–551.

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