BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 18 2009, pages 2440–2441
doi:10.1093/bioinformatics/btp417

Phylogenetics

TreeTime: an extensible C++ software package for Bayesian phylogeny reconstruction with time-calibration

Lin Himmelmann[1,*] and Dirk Metzler[2,*]

[1]Department for Computer Science and Mathematics, Goethe-University, Frankfurt am Main, and
[2]Department of Biology, University of Munich (LMU), Munich, Germany

Received on March 11, 2009; revised on June 30, 2009; accepted on July 1, 2009

Advance Access publication July 3, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: For the estimation of phylogenetic trees from molecular data, it is worthwhile to take prior paleontologic knowledge into account, if available. To calibrate the branch lengths of the tree with times assigned to geo-historical events or fossils, it is necessary to select a *relaxed molecular clock* model to specify how mutation rates can change along the phylogeny.

Results: We present the software TreeTime for Bayesian phylogeny estimation. It can take prior information about the topology of the tree and about branching times into account. Several relaxed molecular clock models are implemented in TreeTime. TreeTime is written in C++ and designed to be efficient and extensible.

Availability: TreeTime is freely available from http://evol.bio.lmu.de/statgen/software/treetime under the terms of the GNU General Public Licence (GPL, version 3 or later).

Contact: lin@linhi.de; metzler@bio.lmu.de

1 INTRODUCTION

Many classical methods for the estimation of phylogenies from molecular data do not assume a molecular clock, especially those that estimate unrooted trees (e.g. Felsenstein, 1989; Saitou and Nei, 1987). When additional informations, e.g. from the fossil record or geo-historical events are used to time-calibrate a phylogeny, it is necessary to model the variability of mutation rates along the tree. The hypothesis of a strict molecular clock is significantly violated for most large datasets. Several relaxed molecular clock models for the variation of mutation rates along the tree are currently discussed (Himmelmann, 2009; Lepage *et al.*, 2007). In the uncorrelated exponential model and the uncorrelated log-normal model (UEX and ULN, (Drummond *et al.*, 2006)) rate factors for all edges in the tree are independently drawn from an exponential and log-normal distribution, respectively. Other relaxed clock models change the rates along the tree in an autocorrelated way, e.g. the compound Poisson process (CPP) model of Huelsenbeck *et al.* (2000) or the autocorrelated log-normal (CLN) model in Thorne *et al.* (1998). Some programs for inferring phylogenies provide one or two relaxed molecular clock models as options, e.g. BEAST (version 1.4.7) offers ULN and UEX (Drummond and Rambaut, 2007). The PhyloBayes software (version 2.3, Lartillot and Philippe, 2004) provides four different rate change models but restricts a temporal analysis to one topology.

When large datasets are to be analyzed, the choice of the relaxed clock model may have a significant impact on the result. The process of model fitting may include several known models as well as newly developed variants of these models. Software tools that are used for the analysis should therefore support the addition of new models. TreeTime is a Markov chain Monte Carlo (MCMC) framework for Bayesian phylogeny reconstruction. It is very flexible and extensible in many ways, including the selection of relaxed clock models, priors on tree topologies, sequence evolution models, specification of topological constraints and MCMC methodology.

2 MODELS AND METHODS

Like MrBayes and Beast (Drummond and Rambaut, 2007; Huelsenbeck and Ronquist, 2001), TreeTime is an implementation of a Metropolis-coupled MCMC (MCMCMC) method for Bayesian phylogeny sampling, i.e. the program outputs possible phylogenies according to their posterior probabilities given the sequence data, making also the uncertainty of the phylogeny estimation assessable. The user can give prior information about the phylogeny's topology to TreeTime by specifying two taxa sets, A and B, such that only trees are allowed, in which at least one branch separates A from B. Neither A nor B needs to be monophyletic. Furthermore, the user can specify normally or gamma distributed priors for the time of the split between A and B. The normal distribution may be appropriate when a speciation is associated with a geo-historical event that can be dated with certain error bounds. When fossils of age t are known to belong to the descendants of the split, and therefore must postdate the actual divergence event (Müller and Reisz, 2005), the user of TreeTime can specify that the split was at time $t+G$ before present, where G is gamma distributed with parameters that are also specified by the user. Such priors can be specified for one or more pairs (A, B) of taxa sets.

TreeTime allows the input of multi-locus sequence data. Sequence evolution models and relaxed clock models can be specified separately for each locus. In addition to the strict molecular clock model, four relaxed clock models are currently implemented in TreeTime: UEX, ULN, CPP and a novel Dirichlet model (DM). Like ULN and UEX, the DM model relaxes the molecular clock by assigning an individual mutation rate to each branch of the tree. The joint prior distribution of these mutation rates is multivariate Dirichlet, such that all rates have the same expectation. The relaxation of the molecular clock is controlled by a meta-parameter that determines the variance of the Dirichlet distribution. This meta-parameter has an exponential prior, which can be adjusted by the user. For the substitution process on DNA sequences the Jukes–Cantor model and the General Time-Reversible model are currently implemented in TreeTime, and both can be combined with the

*To whom correspondence should be addressed.

assumptions of gamma distributed rate variation along the DNA and the existence of invariant sites (Rogers, 2001; Tavaré, 1986).

On the TreeTime web site a detailed example is given on how to add a substitution model or a relaxed molecular clock model to TreeTime. Extending TreeTime requires some basic knowledge of C++. We selected the C++ programming language for the development of TreeTime because C++ supports a clean object-oriented design, produces little overhead in runtime, and allows a precise control of memory consumption.

Much effort has been made to optimize the efficiency of TreeTime. We compared the performance of TreeTime, Beast and MrBayes with simulated datasets with nine taxa. The three programs were equally precise in the estimation of the tree topology after 1.6 million MCMC steps. TreeTime and Beast gave similar estimations for the branching times. However, TreeTime needed only ∼60% of the runtime of Beast and 40% of the runtime of MrBayes to perform the same number of MCMC steps (Himmelmann, 2009). It is possible that Beast and MrBayes converge faster than TreeTime for some datasets, because these programs allow more extensive MCMC steps for the tree topology, while the current version of TreeTime is restricted to nearest-neighbor interchanges. Since TreeTime is mainly designed as a framework for extensions, we focused on the efficiency in performing fundamental MCMC steps.

In Hipsley *et al.* (2009), TreeTime is applied to study the evolutionary history of the family of wall lizards *Lacertidae*, see also Himmelmann (2009).

3 CONCLUSIONS

The freely available TreeTime software for estimating phylogenenies from molecular data can take paleontologic expertise into account. TreeTime uses the Nexus file format (Maddison *et al.*, 1997), which facilitates the application for users who are already familiar with software like PAUP or MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Swofford, 2003).

The models that are already implemented in TreeTime include several different relaxed clock models and make TreeTime applicable for many datasets. In the case that a dataset requires a special model or a newly developed relaxed clock model is to be explored, it is possible to augment TreeTime with this model. While experienced Java programmers may prefer to implement novel relaxed clock models as extensions for Beast (Drummond and Rambaut, 2007), TreeTime is an alternative for those who prefer C++. TreeTime is designed to make the addition of new relaxed clock models very easy. Elementary C++ programming skills suffice to extend TreeTime.

ACKNOWLEDGEMENTS

For stimulating discussions and helpful comments we thank Markus Pfenninger, Johannes Müller, Christy Hipsley, Lisha Naduvilezhath and three anonymous referees.

Conflict of Interest: none declared.

REFERENCES

Drummond,A.J. and Rambaut,A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.

Drummond,A.J. *et al*. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.

Felsenstein,J. (1989) Phylip—phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.

Himmelmann,L. (2009) *Bayessche Methoden zur Schätzung von Stammbäumen mit Verzweigungszeitpunkten aus molekularen Daten*. PhD Thesis, Goethe-Universität Frankfurt am Main.

Hipsley,C.A. *et al*. (2009) Integration of Bayesian molecular clock methods and fossil-based soft bounds reveals early Cenozoic origin of African lacertid lizards. *BMC Evol. Biol.*, **9**, 151.

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

Huelsenbeck,J.P. *et al*. (2000) A compound poisson process for relaxing the molecular clock. *Genetics*, **154**, 1879–1892.

Lartillot,N. and Philippe,H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.

Lepage,T. *et al*. (2007) A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.*, **24**, 2669–2680.

Maddison,D.R. *et al*. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621.

Müller,J. and Reisz,R.R. (2005) Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *Bioessays*, **27**, 1069–1075.

Rogers,J.S. (2001) Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.*, **50**, 713–722.

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

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

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

Tavaré,S. (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.*, **17**, 57–86.

Thorne,J.L. *et al*. (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.