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

Lin Himmelmann\(^1\) and Dirk Metzler\(^2\),

\(^1\)Goethe-University, Department for Computer Science and Mathematics, Frankfurt am Main, Germany, www.linhi.com
\(^2\)University of Munich (LMU), Biology Department, Munich, Germany, www.zi.biol.lmu.de/evol/StatGen.html

February 24, 2009

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://www.zi.bio.lmu.de/evol/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

Introduction

Many classical methods for the estimation of phylogenies from molecular data do not assume a molecular clock, especially those that estimate unrooted trees, cf. e.g. [SN87, Fel89]. 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 data sets. Several

---

\(^*\)to whom correspondence should be addressed
relaxed molecular clock models for the variation of mutation rates along the tree are currently discussed, cf. [LBPL07, Him09]. In the uncorrelated exponential model (UEX) and the uncorrelated log-normal model (ULN) [TKP98, DHPR06] 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 model (CPP) of [HLS00]. 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, [DR07]. The PhyloBayes software (version 2.3, [LP04]) provides four different rate change models but restricts a temporal analysis to one topology.

When large data sets 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.

Models and Methods

Like MrBayes and Beast [DR07, HR01], TreeTime is an implementation of a Metropolis-coupled Markov-Chain Monte-Carlo (MCMC) 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 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 to a geo-historical event that can be dated with certain error bounds. When fossils of age $t$ are known from species that must have lived after the split (cf. [MR05]), then the user of TreeTime can specify that the split was $t + G$ 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. Besides the strict molecular clock model, four relaxed clock models are already implemented in TreeTime: UEX, ULN, CPP, and a novel model, in which a Dirichlet distribution is used as a prior for the vector of rates factors for the edges of the tree. For the substitution process on DNA sequences the Jukes-Cantor model and the general time-reversible model are currently imple-
mented in TreeTime, and both can be combined with the assumptions of gamma distributed rate variation along the DNA and the existence of invariant sites, cf. [Tav86, Rog01].

On the TreeTime website 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 data sets with 9 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. TreeTime needed approximately 60% of the runtime of Beast and 40% of the runtime of MrBayes to perform the same number of MCMC steps, cf. [Him09]. It is possible that Beast and MrBayes converge faster than TreeTime for some data sets, 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 [HHMM09] TreeTime is applied to genetic data from 35 species of wall lizards Lacertidae to examine the evolutionary history of this family. Four different branching events in its phylogeny can be dated from the fossil record. The suitability of the strict molecular clock model and four different relaxed clock models for this data set is analysed using simulation studies and Bayes factors (cf. [LBPL07]).

Conclusions

The freely available TreeTime software for estimating phylogenies from molecular data can take paleontologic expertise into account. TreeTime uses the Nexus file format [MSM97], which facilitates the application for users who are already familiar with software like PAUP or MrBayes, cf. [Swo03, HR01, RH03].

The models that are already implemented in TreeTime include several different relaxed clock models and make TreeTime applicable for many data sets. In the case that a data set 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 [DR07], 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.
Acknowledgments

We thank Markus Pfenninger, Johannes Müller, Christy Hipsley and Lisha Naduvilezhath for stimulating discussions and helpful comments.

References


