Genome analysis

**markophylo: Markov chain analysis on phylogenetic trees**

Utkarsh J. Dang* and G. Brian Golding

Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on April 29, 2015; revised on September 2, 2015; accepted on September 8, 2015

**Abstract**

Summary: Continuous-time Markov chain models with finite state space are routinely used for analysis of discrete character data on phylogenetic trees. Examples of such discrete character data include restriction sites, gene family presence/absence, intron presence/absence and gene family size data. While models with constrained substitution rate matrices have been used to good effect, more biologically realistic models have been increasingly implemented in the recent literature combining, e.g., site rate variation, site partitioning, branch-specific rates, allowing for non-stationary prior root probabilities, correcting for sampling bias, etc. to name a few. Here, a flexible and fast R package is introduced that infers evolutionary rates of discrete characters on a tree within a probabilistic framework. The package, markophylo, fits maximum-likelihood models using Markov chains on phylogenetic trees. The package is efficient, with the workhorse functions written in C++ and the interface in user-friendly R.

**Availability and implementation:** markophylo is available as a platform-independent R package from the Comprehensive R Archive Network at https://cran.r-project.org/web/packages/markophylo/. A vignette with numerous examples is also provided with the R package.

**Contact:** udang@mcmaster.ca

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 Introduction

A Markov chain framework is frequently adopted to estimate evolutionary rates of discrete characters on phylogenetic trees. Some examples of previously investigated data using such models include restriction sites (Felsenstein, 1992), gene family size data (Hahn et al., 2005), gene family presence/absence patterns (Hao and Golding, 2006), intron presence/absence data (Kim and Hao, 2014), etc. A good introduction to such models can be found in O’Meara (2012) and Yang (2014). It has been noted before that many variants of such models are just restrictions of more general models for investigation of specific hypotheses or different kinds of data (O’Meara, 2012). Hence, there is a clear need for a flexible and efficient software to provide a unified interface that would allow fitting of varied discrete character datasets.

Recently, Kim and Hao (2014) put forward a unified R (R Core Team, 2015) package, namely **DiscML**, that provided users the option of biologically realistic features like gamma rate variation (Yang, 1994), estimation of character prior root probabilities (Cohen et al., 2008), correcting for ancient characters being lost from all examined extant taxa (cf. Felsenstein, 1992), etc. As noted in Kim and Hao (2014), while many of these features are found in existing programs like BayesTraits (Pagel et al., 2004), CAFE3 (Han et al., 2013), BdirRate (Librado et al., 2012) and GLOGML (Cohen et al., 2010), the sum total of these above-mentioned desirable features was not available in a single flexible software. However, **DiscML** is not computationally inexpensive. Moreover, **DiscML** does not currently allow for partition analyses or for accounting for sampling bias arising due to multiple unobserved phylectic patterns (see Supplementary Material).

Here, we present markophylo, an R package that is both fast and flexible. The markophylo package allows for estimating evolutionary rates using a user-specified, i.e. hypothesis driven,
substitution rate matrix in a continuous-time Markov chain model on phylogenies. The package is computationally efficient, with the workhorse functions written in C++, using the Rcpp (Eddelbuettel et al., 2011) and RcppArmadillo (Eddelbuettel and Sanderson, 2014) packages. The package is flexible, capable of modelling a myriad array of processes (described below in Section 2.2).

2 Description

2.1 Input

A phylogenetic tree with branch lengths in units of expected substitutions per site and a matrix containing phyletic patterns of discrete characters are the primary external inputs needed for markophylo. The tree must be in the phylo format of the ape package (Paradis et al., 2004), while each row of the matrix represents a phyletic pattern, where the columns denote closely related taxa of interest.

2.2 Features

The following features are available in the primary function markophylo::estimaterates:

1. Custom substitution rate matrices can be easily specified. These can be as simple as matrices where each non-diagonal entry is hypothesized to be the same (discrete character analogue of Jukes and Cantor, 1969), or birth-death matrices, or symmetric matrices (where the instantaneous rate of change from character $i$ to $j$ is the same as character $j$ to $i$), to matrices where each non-diagonal entry is different.

2. A custom character-only or numeric-only alphabet for the discrete characters can be specified with no limit on the number of possible states.

3. Clades or groups of branches hypothesized to follow substitution rates different from other branches can be easily specified. This is often essential when a clade is very diverged from other branches based on evolutionary time.

4. Sites can also be split into different partitions, where each partition of sites is a group of sites following their own rates different from the sites in the other partitions. This is useful, e.g. when different rates are hypothesized for mitochondrial versus nuclear genes (O’Meara, 2012).

5. Gamma rate variation (Yang, 1994) can be specified: a common gamma distribution (with $\alpha = \beta$, where $\alpha$ and $\beta$ are the shape and rate parameters, respectively) over all partitions or separate gamma rates within each partition separately. The latter option is useful when the hypothesized evolutionary processes among the different partitions are different enough to warrant separate gamma distributions for each partition.

6. Prior root probabilities for each discrete character can be either user-specified (based on some known constraint at the root), equal for each character state, follow Markov chain stationary probabilities or be estimated in the maximum likelihood framework when it is not reasonable to assume stationarity.

7. Correcting for multiple unobservable phyletic patterns, i.e. sampling (aka acquisition or ascertainment) bias. It is often important to correct for sampling bias. If sampling bias exists during data collection, some phyletic patterns cannot be observed in the data and as a result, not correcting for this in the statistical model can lead to biased estimates. Such a correction has been applied previously. For example, this correction has been used in the analysis of restriction sites (Felsenstein, 1992), when only variable characters are recorded (Lewis, 2001), correcting for unobservable ancient genes that are lost and not observed at the tips (Hao and Golding, 2006) or because gene families appearing in the COG database cannot occur in less than three genomes (Cohen and Pupko, 2010). DiscML only provides the option of correcting for observations of a zero character for each taxa; however, multiple user-specified phyletic patterns can be easily corrected for in markophylo.

The package also contains five example (simulated) datasets and a vignette that contains numerous examples. These data illustrate the kinds of models that the package is capable of fitting to discrete character data recorded for multiple taxa (with a user phylogeny).

2.3 Output

The output from the primary function contains parameter estimates for the user-specified substitution rate matrix, standard errors, time taken, a reduced dataset containing unique patterns and their frequencies and model selection criteria values, namely the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion (Schwarz, 1978).

3 Discussion

Here, a flexible and efficient R package named markophylo for estimating evolutionary rates of discrete characters is introduced. When compared with existing packages like DImcML, markophylo is at least an order of magnitude faster. Moreover, markophylo implements a wide variety of features increasingly seen in more biologically realistic Markov chain models run on discrete character datasets.

Funding

This work was supported by the National Sciences and Engineering Research Council of Canada [140221-10 to G.B.G].

Conflict of Interest: none declared.

References


