MATICCE: mapping transitions in continuous character evolution

Journal:	*Bioinformatics*
Manuscript ID:	Draft
Category:	Applications Note
Date Submitted by the Author:	
Complete List of Authors:	Hipp, Andrew; The Morton Arboretum, Herbarium Escudero, Marcial; Pablo de Olavide University, Department of Molecular Biology and Biochemical Engineering
Keywords:	Phylogenetics, Data analysis, Evolution

APPLICATION NOTE

Phylogenetics

MATICCE: mapping transitions in continuous character evolution

Andrew L. Hipp[1,*] and Marcial Escudero[2]

[1]Herbarium, The Morton Arboretum, 4100 Illinois Route 53, Lisle IL 60532-1293, USA, [2] Department of Molecular Biology and Biochemical Engineering, Pablo de Olavide University, Ctra. Utrera km 1, 41013 Seville, Spain

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: MATICCE is a new software package in the R language for inferring shifts in continuous character distribution on phylogenetic trees. MATICCE also provides simulation functions for visualizing analysis results and functions for accounting for phylogenetic and model uncertainty.

Availability and Implementation: MATICCE is open source, written solely in R, and free (http://r-forge.r-project.org).

Contact: ahipp@mortonarb.org.

1 INTRODUCTION

Inferring the evolutionary history of species traits is an issue of practical concern for any researcher testing biological hypotheses using phylogenetic data. Numerous methods utilizing phylogenetic generalized least squares have been utilized to model shifts in the dynamics of continuous character evolution (Martins and Hansen, 1997; Pagel, 1999; Butler and King, 2004; O'Meara *et al*., 2006; Hansen *et al*., 2008; Revell and Collar, 2009). However, there is no standardized approach for reconstructing the evolution of shifts in the distribution of continuous characters across a phylogenetic tree and quantifying the relative support for those transitions.

A recent study (Hipp, 2007) utilized an information-theoretic approach to identify where on a phylogeny there have been shifts in the stationary distribution of a continuous character. The method utilizes the Ornstein-Uhlenbeck (O-U) model of character evolution as implemented in *ouch* (King and Butler, 2009), modeling character transitions as shifts in character equilibrium. Under an O-U process, a continuous character evolves stochastically toward a stationary distribution with mean θ and (at stationarity) variance $\sigma^2 / 2\alpha$, where α determines the rate of evolution toward the stationary distribution (Butler and King, 2004). The use of O-U models in a phylogenetic context typically focuses on testing adaptive scenarios (Butler and King, 2004; Hansen *et al*., 2008) or phylogenetic constraints (Blomberg *et al*., 2003), with the phylogenetic locations of continuous character transitions defined *a priori* with respect to shifts in hypothesized selective regimes. The approach introduced in Hipp (2007) and implemented in MATICCE evaluates the relative support for alternative locations of continuous character distribution shifts by (1) identifying n nodes at which a change may have occurred, (2) evaluating support for all 2^n models that allow all permutations of change at those nodes, and (3) using a information theoretic approach to estimate the relative support for a shift in character distribution at each node as the cumulative information criterion weight for all models entailing a shift at that node. The method is reminiscent of stepwise AIC model-selection (e.g., Alfaro *et al.*, 2009). However, by taking a model-averaging approach (cf. Burnham and Anderson, 2002) rather than a model-selection approach, it is not susceptible to the significance threshold effects that introduce uncertainty into model-selection.

2 DESCRIPTION

MATICCE is a package written in the R statistical language (R Development Core Team, 2009). The package automatically generates the potentially large number of model specifications required to test character transition hypotheses and provides a set of tools for flexibly defining nodes for analysis; performs and summarizes analyses over sets of trees, integrating over phylogenetic and model uncertainty and allowing analyses to be performed even for nodes that are not found in all trees in a set; provides tools for simulating and visualization character evolution under a user-specified model or model-averaged parameters from analysis; and provides an easy interface by which to test numerous whole-tree and partitioned-tree models for a particular phylogenetic branch. The analysis framework implemented in MATICCE utilizes and complements the *ouch* package (King and Butler, 2009) while providing needed functions for bookkeeping, model definition, and data visualization. Worked examples, sample data, and a discussion of assumptions underlying MATICCE are available in the online documentation and vignette.

By default, models are defined in MATICCE as all possible permutations of models allowing changes in character distribution up to a defined number of nodes, such that all nodes being evaluated are present in the same number of models. To relax computational limitations, users can limit the maximum number of nodes at which a character is allowed to change distribution (for example, 30 candidate nodes could be investigated for models allowing up to a maximum of three character distribution transitions; this is a manageable 4526 models, compared to the 1.07E09 models defined by all possible permutations of 30 candidate nodes). Because sets of phylogenetic trees (e.g., bootstrap sets) are not typically identical in topology, nodes are defined in MATICCE by the entire set of taxa descendent from them. Lists of taxa defining a set of nodes can be generated quickly in R, and MATICCE utilizes these user-defined lists to create the subset of models applicable to each

*To whom correspondence should be addressed.

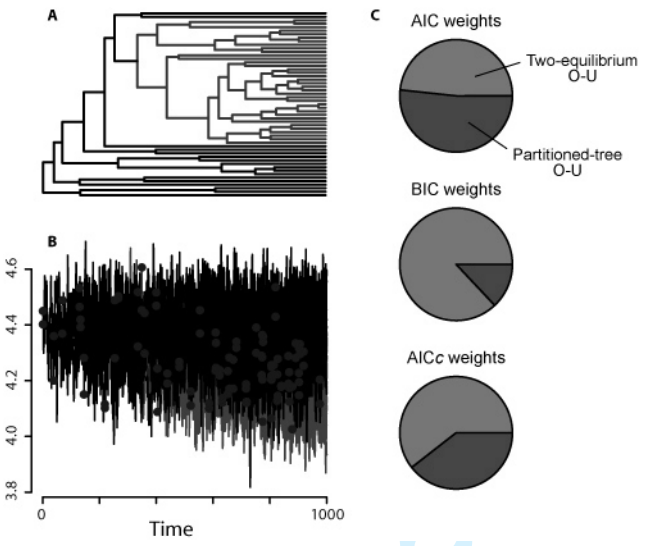

Fig. 1. (A) Phylogeny of 53 sedges, with clade of interest in light gray (Hipp 2007). Note that the vertical scale in this phylogeny, as in most phylogenies, is arbitrary, with branches ordered to avoid crossing. **(B)** Simulation of character evolution on same phylogeny using the `ouSim` function, with nodes indicated by blue dots and vertical scale indicating character value. Simulation parameters were estimated from the data. **(C)** Estimates of the relative support for two alternative models of character transition at the base of the red-colored clade using the `multiModel` function: a two-equilibrium O-U model versus a partitioned ("censored") tree model in which each partition evolves according to a single-equilibrium O-U model.

tree. Analyses are then conducted and summarized conditional on the models possible on each tree in the set. User-defined sets of models can also be created to test specific character evolution scenarios, or individual models created for testing in *ouch*. All model specifications are compatible with the analysis, simulation, and visualization functions in *ouch*, and objects created in the course of analysis can be analyzed directly in MATICCE or used as input for discrete-time character simulations (Figs. 1A, B).

MATICCE provides straightforward functions for calculating and summarizing information criterion (AIC, AIC*c*, and BIC; Burnham and Anderson, 2002) values and weights for any set of models, as well as a method for analyses conducted in MATICCE. Analyses are summarized for each node both over all trees and over just the trees in which the node occurs. When analyses are conducted over a set of trees sampling from the posterior probability distribution of the phylogeny (e.g., the trees visited in a Markov chain Monte Carlo phylogenetic analysis), the evidential support for a shift in character distribution occurring at a given node may be interpreted either as the relative support for that particular shift conditioned on the existence of the node, or as the product of the posterior probability of that node and the relative support for that particular shift conditioned on the existence of the node.

Based on these analyses, users will often want to evaluate the relative support for different models of character evolution at a particular node. MATICCE provides a function for evaluating whether an Ornstein-Uhlenbeck model with a change in distribution fits the observed change better than a partitioned-tree (or "censored" model *sensu* O'Meara *et al*., 2006) or no-change models (Fig. 1C). The function lends itself to customization and can easily accommodate additional models as desired. The effects of different model assumptions can also be visualized through a set of flexible discrete-time simulation functions (Fig. 1B) that integrate smoothly with the analysis functions; running a simulation on the results of a MATICCE batch analysis, for example, returns a simulation of character evolution using model-averaged parameter values for σ, α, and the character mean (θ) for each branch.

3 CONCLUSION

While there has been a substantial increase in statistical analysis of continuous character evolution in the past several years, an easy-to-use general framework for reconstructing shifts in continuous character evolution across a phylogeny has been lacking. MATICCE uses existing tools to implement such a framework, providing a needed tool for exploratory analysis of continuous character evolution.

ACKNOWLEDGEMENTS

We thank A.King, M.Butler, A.Platt, and B.O'Meara for helpful comments on this work.

Funding: This work was supported by a NESCent visiting scholar award to A.H.; the National Science Foundation [0743157 to A.H.]; and the Spanish Government [FPU AP2005–3715 to M.E.].

REFERENCES

Alfaro,M.E. *et al.* (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 13410–13414.

Blomberg,S.P. *et al.* (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Burnham,K.P. *et al.* (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, NewYork.

Butler,M.A. and King,A.A. (2004) Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am.Nat.*, **164**, 683–695.

Hansen,T.F. et al. (2008) A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, **62**, 1965–1977.

Hipp,A.L. (2007) Nonuniform processes of chromosome evolution in sedges (Carex: Cyperaceae). *Evolution*, **61**, 2175–2194.

King,A.A. and Butler,M.A. (2009) ouch: Ornstein-Uhlenbeck models for phylogenetic comparative hypotheses (R package), http://ouch.r-forge.r-project.org.

Martins,E.P. and Hansen,T.F. (1997) Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am.Nat.*, **149**, 646–667.

O'Meara,B.C. *et al.* (2006) Testing for different rates of continuous trait evolution using likelihood. *Evolution*, **60**, 922–923.

Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

R Development Core Team. 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Revell,L.J. and Collar,D.C. (2009) Phylogenetic analysis of the evolutionary correlation using likelihood. *Evolution*, **63**, 1090–1100.