ARCHIPELAGO

Jeet Sukumaran

November 1, 2015

1 Introduction

"Archipelago" is the name of a generative phylogeny-based model that simultaneously incorporates the diversification processes of speciation and extinction nad the biogeographical processes of area gain ("dispersal") and area loss ("extirpation"), with these processes being differentially regulated by ecological or other traits that are themselves co-evolving on the phylogeny. The theory and background to this model and its usage is described in the following paper:

This software project, "archipelago" presents a suite of programs to generate and analyze data under the Archipelago model. The primary objective of the analysis is *model selection*: i.e., statistically identifying the model that generated a particular set of data. "archipelago" is, thus, a computational biogeographical model selection analysis package that exploits the power and flexibility of the Archipelago model to allow you to ask and answer historical biogeographical questions of a nature and complexity that are not possible under any other approach. In particular, instead of asking questions about ancestral area patterns, you can ask questions about processes, and how some processes (e.g., ecology) affect other processes (e.g., dispersal or speciation).

2 Installation

2.1 Pre-requisites

- Python 2.7 or higher

- DendroPy Phylogenetic Computing Library, version 4 or above (http://dendropy.org).

- R (http://www.r-project.org/)

- The following R packages:
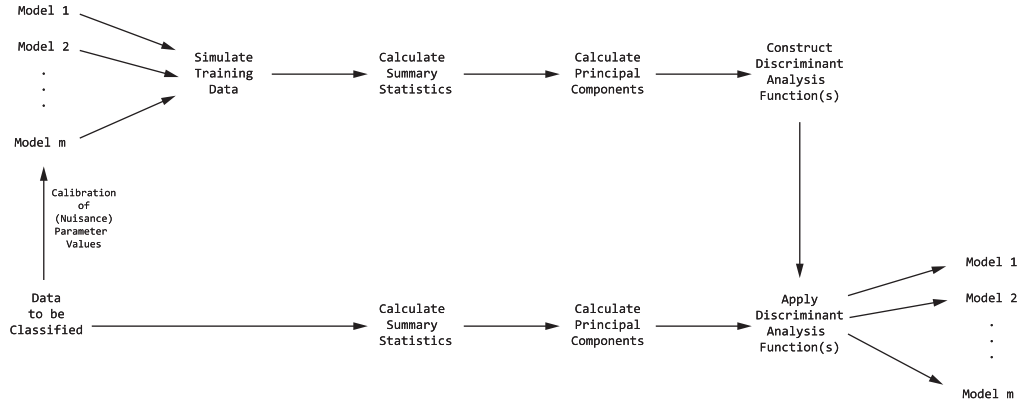
  - adegenet
  - picante
  - BioGeoBears
  - GEIGER

2.2 Installation from Source

Run the following in the top-level directory of the project:

```
python setup.py install
```

3 Usage

4 Overview

    Let D be the *target* data, consisting of an ultrametric phylogeny with each of the tips associated with a geographic range (presence/absence over one or more areas) as well as with set of trait states. Let $\mathbf{M} = \{M_1, M_2, ...M_k\}$ be a set of k models that we are interested in studying. Our analytical objective is, for each $M_i, M_i \in \mathbf{M}$, estimate the probability that it generated the target data D, relative to all the other models in $\mathbf{M}$. We achieve this by generating data under each of the models in $\mathbf{M}$, and then using this

training data to construct a machine learning classifier, i.e., a discriminant analysis function, that classifies the original data D with respect to each of the models in $\mathbf{M}$.