

ARCHIPELAGO

Jeet Sukumaran

November 2, 2015

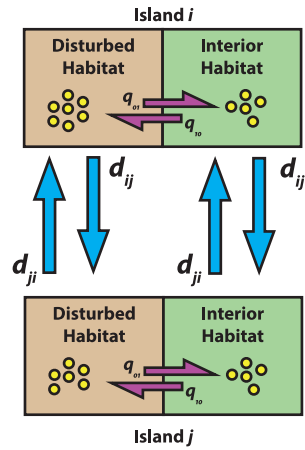
1 Introduction

“Archipelago” is the name of a generative phylogeny-based model that simultaneously incorporates the diversification processes of speciation and extinction and 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:

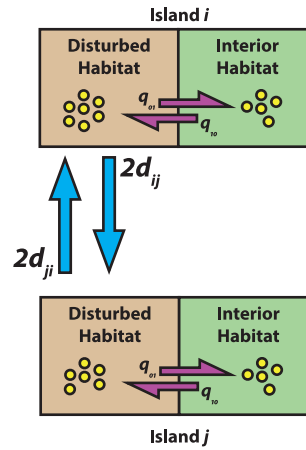
This software project, “archipelago” presents a suite of programs to generate and analyze data under the Archipelago model. The primary objective of the analysis is *model selection*: i.e., statistically identifying the model that generated a particular set of data. “archipelago” is, thus, a computational 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).

As an example of a very basic analysis, let us consider testing whether or not dispersal is constrained by habitat association of a lineage (Figure). We propose two different models, each corresponding to a different dispersal regime. In the “unconstrained” model, lineages are free to disperse between islands regardless of their habitat affinities. In the “constrained” model, only lineages associated with the disturbed or open habitats can disperse between islands, while lineages associated with interior habitats are not able to disperse.

UNCONSTRAINED



CONSTRAINED



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

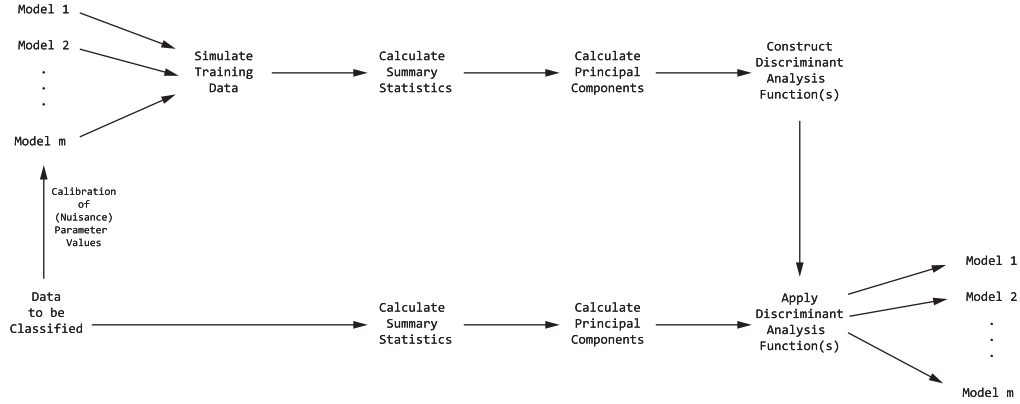
2.2 Installation from Source

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

```
python setup.py install
```

3 Usage

3.1 Overview of Workflow



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 $\mathbf{M} = \{M_1, M_2, \dots, M_m\}$ be a set of k models that we are interested in studying. Let $S(\cdot)$ be a function that takes a set of data returns a set of summary statistics. 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} . The operational procedure consists of the following steps:

1. **Simulation of Training Data:** For each $M_i, M_i \in \mathbf{M}$, simulate n replicates of data, D^{*i} . We label each replicate of this data with the name of the model that generated it (e.g., “Model i ”). The set of all data so generated constitutes the training data, \mathbf{D}^* .
2. **Calculation of Summary Statistics:** Calculate summary statistics on the training data to yield, $S(\mathbf{D}^*)$, and summary statistics on the target data to yield $S(D)$.

3. **Classification of Target Data:** Construct a Discriminant Analysis (DA) function on principal components (PC) calculated on the training data set summary statistics, $S(\mathbf{D}^*)$; apply the discriminant analysis function to principal components calculated on the summary statistics of the target data, $S(D)$.

3.2 Simulation of Training Data

We use the program, “`archipelago-simulate.py`”, to simulate the training data.