Phylogenetic Inference using `RevBayes`

Discrete Morphology

April M. Wright and Michael J. Landis

Introduction

While molecular data have become the default for building phylogenetic trees for many types of evolutionary analysis, morphological data remains important, particularly for analyses involving fossils. The use of morphological data raises special considerations for model-based methods for phylogenetic inference. Morphological data are typically collected to maximize the number of parsimony-informative characters - that is, the characters that favor one topology over another. Morphological characters also do not carry common meanings from one character in a matrix to the next; character codings are made arbitrarily. These two factors require extensions to our existing phylogenetic models. Accounting for the complexity of morphological characters remains challenging. This tutorial will provide a discussion of modeling morphological characters, and will demonstrate how to perform Bayesian phylogenetic analysis with morphology using `RevBayes`.

Contents

The Discrete Morphology guide contains several tutorials

Recommended tutorials

The Discrete Morphology tutorials assume the reader is familiar with the content covered in the following `RevBayes` tutorials

- `Rev` **Basics**
- **Molecular Models of Character Evolution**
- **Running and Diagnosing an MCMC Analysis**
- **Divergence Time Estimation and Node Calibrations**

1 Overview of Discrete Morphology Models

Molecular data forms the basis of most phylogenetic analyses today. However, morphological characters remain relevant: Fossils often provide our only direct observation of extinct biodiversity; DNA degradation can make it difficult or impossible to obtain sufficient molecular data from fragile museum specimens. Using morphological data can help researchers include specimens in their phylogeny that might be left out of a molecular tree.

To understand how morphological characters are modeled, it is important to understand how characters are collected. Unlike in molecular data, for which homology is algorithmically determined, homology in a character is typically assessed an expert. Biologists will typically decide what characters are homologous by looking across specimens at the same structure in multiple taxa; they may also look at the developmental origin of structures in making this assessment. Once homology is determined, characters are broken down into states, or different forms a single character can take. The state '0' commonly refers to absence, meaning that character is not present. In some codings, absence will mean that character has not evolved in that group. In others, absence means that that character has not evolved in that group, and/or that that character has been lost in that group. This type of coding is arbitrary, but both **non-random** and **meaningful**, and poses challenges for how we model the data.

Historically, most phylogenetic analyses using morphological characters have been performed using the maximum parsimony optimality criterion. Maximum parsimony analysis involves proposing trees from the morphological data. Each tree is evaluated according to how many changes it implied in the data, and the tree that requires the fewest changes is preferred. In this way of estimating a tree, a character that does not change, or changes only in one taxon, cannot be used to discriminate between trees (i.e., it does not favor a topology). Therefore, workers with parsimony typically do not collect characters that are parsimony uninformative.

In 2001, Paul Lewis introduced a generalization of the Jukes-Cantor model of sequence evolution for use with morphological data. This model, called the Mk (Markov model, assuming each character is in one of k states) model provided a mathemetical formulation that could be used to estimate trees from morphological data in both likelihood and Bayesian frameworks. While this model is a useful step forward, as a generalization of the Jukes-Cantor, it still makes fairly simplistic assumptions. This tutorial will guide you through estimating a phylogeny with the Mk model, and two useful extensions to the model.

1.1 The Mk Model

The Mk model is a generalization of the Jukes-Cantor model of nucleotide sequence evolution, which we discussed in **Molecular Models of Character Evolution**. The Q matrix for a two-state Mk model looks like so:

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{pmatrix},$$

This matrix can be expanded to accommodate multi-state data, as well:

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix},$$

However, the Mk model sets transitions to be equal from any state to any other state. In that sense, our multistate matrix really looks like this:

$$Q = \begin{pmatrix} -\mu_0 & \mu & \mu & \mu \\ \mu & -\mu_1 & \mu & \mu \\ \mu & \mu & -\mu_2 & \mu \\ \mu & \mu & \mu & -\mu_3 \end{pmatrix},$$

Because this is a Jukes-Cantor-like model, state frequencies do not vary as a model parameter. These assumptions may seem unrealistic. However, all models are a compromise between reality and generalizability. Because morphological characters do not carry common meaning across sites in a matrix in the way that nucleotide characters do, making assumptions that fit all characters is challenging.

We will first perform a phylogenetic analysis using the Mk model. In further sections, we will explore how to relax key assumptions of the Mk model.

1.2 Ascertainment Bias

When Lewis first introduced the Mk model, he observed that branch lengths on the trees were greatly inflated. The reason for this is that when morphological characters are collected, characters that do not vary, or vary in a non-parsimony-informative way (such as autapomorphies) are excluded. Excluding these low-rate characters causes the overall amount of evolution to be over-estimated. This causes an inflation in the branch lengths.

Therefore, when performing a morphological phylogenetic analysis, it is important to correct for this bias. Original corrections simulated invariant and non-parsimony informative characters along the proposed tree. The likelihood of these characters would then be calculated and used to normalize the total likelihood value. RevBayes implements a dynamic programming approach that calculates the same likelihood, but does so faster.

2 Example: Inferring a Phylogeny of Fossil Bears Using the Mk Model

In this example, we will use morphological character data from 18 taxa of extinct bears. The dataset contains 62 binary characters, a fairly typical dataset size for morphological characters.

2.1 Tutorial Format

This tutorial follows a specific format for issuing instructions and information.

> The boxed instructions guide you to complete tasks that are not part of the `RevBayes` syntax, but rather direct you to create directories or files or similar.

Information describing the commands and instructions will be written in paragraph-form before or after they are issued.

All command-line text, including all `Rev` syntax, are given in **`monotype font`**. Furthermore, blocks of `Rev` code that are needed to build the model, specify the analysis, or execute the run are given in separate shaded boxes. For example, we will instruct you to create a constant node called **`rho`** that is equal to `1.0` using the `<-` operator like this:

```
rho <- 1.0
```

It is important to be aware that some PDF viewers may render some characters given as `Rev command` differently. Thus, if you copy and paste text from this PDF, you may introduce some incorrect characters. Because of this, we recommend that you type the instructions in this tutorial or copy them from the scripts provided.

```
n_areas <- 3
```

2.2 Things to consider

...

3 Overview of Discrete Morphology Models

4 Overview of Discrete Morphology Models

? Alright, who did it?

‘

References

Version dated: February 20, 2017