Site-Heterogeneous Character Change Models for Morphology

Wright, Pett, Students, Heath

Today

1 Introduction

1.1 Bayesian Modeling of Morphology

Recent interest in using Bayesian methods to model morphological evolution to estimate phylogenetic trees has spurred many studies into how well the existing toolkit performs, particularly in relation to parsimony methods. However, there has been little work expanding that toolkit. The most common way to perform likelihood or Bayesian phylogenetic analysis using morphological data is by applying a model of morphological evolution called the Mk model. Many workers using morphological data have raised questions about the realism of this model. In this paper, we will discuss methodological advancements aimed at improving the realism of the model.

The Mk model was introduced by Lewis in 2001. As a generalization of the Jukes-Cantor model, it makes the same set of assumptions: that change between any two states is equally likely, that the stationary character frequencies of every character are the same, and that each character is always in one of k states. The model also makes the Markovian assumption that a character can change instantaneously, regardless of previous states. Any one of these assumptions may strike a reader as problematic for their particular dataset.

Heterogeneity in the evolutionary process is difficult to accommodate in morphological data. Making concrete *a priori* statements about the relative probabilities of transitions between character states in a morphological character matrix is challenging. Unlike a matrix of nucleotide characters, a '0' character at one state may mean that the trait never evolved in one lineage, but was lost in another. This lack of common meaning complicates the interpretation of how change occurs over time. Likewise, the magnitude of change from one state to another may be different between characters, even if the characters mean the same thing (i.e. '0' representing loss and '1' representing gain). For example, in one character, changing from a '0' state to a '1' state could represent a change at a single locus, but at another could represent a much larger change coordinated across many genes. This has complicated the ability to specify α parameters to the Q-matrix *a priori*. This is a stark contrast to molecular data, in which base

pairs or amino acid residues are generally assumed to have similar properties across an alignment.

Prior work extending the Mk model has focused on relaxing the assumption of equal character frequencies at stationarity. MrBayes implemented a parameter called the Symmetric Dirichlet Hyperprior, which allowed users to place a prior or hyperprior on state frequencies. The probability of observing a change in a character is dependent not only on the probability of change from this character to another, but on the frequency of the starting character. Even if a character has a low probability of changing, change in that character may still be observed many times if the stationary frequency of that character is high (i.e., the character is observed many times). Likewise, a highly-probable change may be seen relatively rarely if the starting character is observed rarely. Therefore, relaxing the assumption of equal stationary frequencies has been a way of changing the probabilities of observing different changes without making strong *a priori* statements about transition probabilities.

In the case of binary data, the prior used was a Beta prior (the Dirichlet is a generalization of the Beta distribution), which operated in a fashion similar to gamma-distributed rate variation: the distribution is discretized into a user-specified number of categories, the median forward and backward transition rate are calculated for each category, and the likelihood of the character is calculated over each category and summed to form the total likelihood. For binary data, the state frequencies are integrated out of the likelihood, making the symmetric Dirichlet a prior. For multistate data, the state frequencies are drawn from a Dirichlet distribution. The parameter was referred to as a hyperprior because users can place a distribution on the parameter to the Beta distribution, and because the state frequencies are not integrated out of the likelihood function for multistate data. As implemented in MrBayes, the Beta distribution was assumed to be symmetric, meaning that if there were characters for which the stationary frequency of 0 is high, there would also be corresponding characters for which the stationary frequency of 1 is high.

This is a very useful model extension for morphological evolution. There are clear contexts to improve this concept by borrowing from the molecular literature. The CAT model of Lartillot represents one way forward. This model, implemented in PhyloBayes, uses a Dirichlet process to assign individual sites to categories, which differ in their stationary frequencies. In this model, the number of such categories that describe the data is a free parameter. The use of a Dirichlet prior allows for flexibility in the number of states, as opposed to a Beta prior which is limited to binary data. However, many datasets that have been analyzed under the CAT model are phylogenomic datasets of thousands of loci, as compared to small morphology datasets.

In RevBayes, we have implemented a number of useful extensions to the Dirichlet prior of MrBayes, and a constrained version of the CAT model. Our extension to the symmetric Beta prior allows for the Beta distribution to be asymmetrical. This removes the need for there to be equal numbers of characters with high stationary frequencies for a character state as there are characters with low stationary frequencies for that same character state. Our CAT-like

model (hereafter Site-Heterogeneous Discrete Morphology (SDHM) model) is a finite mixture model that allows characters to fit into a user-defined number of character categories that differ in their stationary frequencies. The bins draw their stationary frequencies from a Dirichlet to accommodate multistate data.

1.2 Morphological Phylogeny of the Formicidae

To test the efficacy of our analytical techniques, we use an ant dataset created from datasets from Barden and Grimaldi and one from Keller. Ants are ecologically crucial organisms as both interacting partners for a variety of plants and animals, and as shapers of the ecosystem via soil cycling and nest building. As such, they have attracted much work in the world of molecular systematics. Ants have also have a rich fossil record, with most subfamilies being represented. This fossil record has been used for systematic work, as well as a variety of other ecological and evolutionary questions.

The monophyly of major subfamilies has been consistently supported by most molecular and morphological studies, with the exception of the Cerapachyinae. Molecular systematics had originally shaken up the ant tree of life considerably, breaking up clades considered to be monophyletic based on morphological work. For example, based on morphological work, six current subfamilies had been previously considered to be a single subfamily, the Ponerinae, until molecular evidence indicated that two of the clades (Ectatomminae and Heteroponerinae) within that subfamily were demonstrated by molecular evidence to more closely related to other ants on the tree. The Ponerinae was then broken into six distinct subfamilies, one of which is called Ponerinae. Recent molecular work has continued to support these six subfamilies as monophyletic, though their relationships to one another remain poorly supported.

Excellent morphological matrices on the ingroup of the Formicidae have been available since the early '90s. Recent morphological matrices have expanded the sampling of ants from in the large Ponerine family, which was previously undersampled. Sampling of fossil ants was also expanded to include specimens from the Cretaceous. These ants were concluded to be stem lineages, and some did not demonstrate any particular taxonomic affinity. Even with expanded taxon sampling in the Ponerinae, there is still conflict between the molecular and morphological estimates of phylogenetic relationships.

In this study, we combine the extant matrices of Keller, and the extinct-extant matrix of Barden and Grimaldi to expand the taxon and character sampling. This dataset has interesting properties. Because there are stem ant lineages represented, there are taxa that have characteristics that are lost after the divergence of the stem lineages. This extremely one-sided loss structure violates the Mk model assumption of equal transition rates. We would expect for these characters to be better modeled by a model that can accommodate asymmetrical transition rates. Some apomorphies of the ant group are also gained after the divergence of the stem lineages, which also violates the assumption of equal change probabilities. Because of these model violations, this dataset is an excellent test case for models that relax assumptions of the Mk model.

We use Bayes Factor model selection to assess the fit of these relaxed models to the data. Using this dataset, we strongly support that the use of models that relax key assumptions of the Mk model can greatly improve the fit of the model to the data. Some stuff about trees and parameters!

2 Methods

a) Bayes Factor model fitting b) TreeSet viz showing how use of the asymetric beta changes our pattern of exploration of treespace

Results

a) Best fit model? b) Tree comparisons, esp. branch lengths. No asc. bias correction + misspecified model, no correction + misspecified model, correction + correct model, correction + misspecified model c) Tree set visualization comparisons

Discussion

a) How has the modeling of morphology changed since the Mk was proposed by Lewis? What aspects are comparable to how we model DNA? Which are not? What are potentially interesting future directions? b) General discussion of usage of treeset visualization c) something else.