Project 1: Modeling Biological Trait Evolution Across Species

Laura Kubatko and Radu Herbei

The Ohio State University
The Mathematical Biosciences Institute

July 7, 2015
References:


Background: phylogenetic trees

- Graphs that display ancestry-descent relationships among a collection of organisms
- Nodes = hypothetical ancestors
- Branches = evolutionary time
- Common goals:
  - Estimate species history given genetic data
  - Understand trait evolution along phylogeny
Motivating example 1: correlation between discrete traits

Consider two traits for a collection of organisms – the “blue” trait and the “purple” trait – both with two possible values:

The example above shows data for 10 individuals (columns) with a perfect association between the observed values for the two traits.

Fisher’s exact test gives a p-value of 0.0048
Motivating example 1: correlation between discrete traits
Motivating example 1: correlation between discrete traits

The association is completely explained by two mutations phylogeny
Motivating example 2: correlation between continuous traits

Consider the following traits for 14 species:
Motivating example 2: correlation between continuous traits
Recall that our goal is to study the evolution of traits along the phylogeny.

Data are observed traits at the present time, but our examples show that the traits are not independent observations – they share some portion of their evolutionary history.

Need to come up with a method to incorporate this shared portion of their history – remove the “phylogenetic” correlation in order to study trait evolution.

First model introduced by Edwards and Cavalli-Sforza in 1964: BM along a phylogeny (continuous traits).
BM along a phylogeny

- Assume that changes in trait value on different branches of a tree are independent; traits evolve over branches of the tree according to the BM model.

- Over a branch of length $t$, the change in trait value is $N(0, \sigma^2 t)$. 
BM along a phylogeny: an example

Let $x_i$ denote the trait value at node $i$, and let $v_i$ denote the branch length.
BM along a phylogeny: an example

- Assume that the state at the root $x_0$ is fixed.
- Note that $x_5 = x_0 + (x_{12} - x_0) + (x_{11} - x_{12}) + (x_5 - x_{11})$
- The last three terms above are independent draws from Normal distributions with mean 0 and variances depending on the $v_i$. 
BM along a phylogeny: an example

We have

\[ E(x_5) = x_0 \]
\[ \text{Var}(x_5) = \sigma^2 v_{12} + \sigma^2 v_{11} + \sigma^2 v_5 \]
BM along a phylogeny: an example

Similarly, for $x_7$, we have

\[
E(x_7) = x_0
\]

\[
Var(x_7) = \sigma^2 v_{12} + \sigma^2 v_{11} + \sigma^2 v_{10} + \sigma^2 v_7
\]
BM along a phylogeny: an example

Note: $x_5$ and $x_7$ are not independent – they share a history up to node $x_{11}$

$$\text{cov}(x_5, x_7) = \sigma^2 v_{12} + \sigma^2 v_{11}$$
BM along a phylogeny: an example

- Now do this for all pairs of tips on a tree.
- The characters on the tips are jointly multivariate normal with expectation $x_0$ and covariances based on their shared history.
- For our example, the variance-covariance matrix is:

$$
\begin{pmatrix}
  v_1 + v_8 + v_9 & v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 & 0 \\
  v_8 + v_9 & v_2 + v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 & 0 \\
  v_9 & v_9 & v_3 + v_9 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & v_4 + v_{12} & v_{12} & v_{12} & v_{12} & v_{12} \\
  0 & 0 & 0 & v_{12} & v_{5 + v_{11} + v_{12}} & v_{11 + v_{12}} & v_{11 + v_{12}} & v_{11 + v_{12}} \\
  0 & 0 & 0 & v_{12} & v_{11 + v_{12}} & v_{6 + v_{10} + v_{11} + v_{12}} & v_{10 + v_{11} + v_{12}} & v_{10 + v_{11} + v_{12}} \\
  0 & 0 & 0 & v_{12} & v_{11 + v_{12}} & v_{10 + v_{11} + v_{12}} & v_{7 + v_{10} + v_{11} + v_{12}} & v_{10 + v_{11} + v_{12}} \\
\end{pmatrix}
$$
Inference under the BM model

- Suppose that $p$ characters are observed – each follows the multivariate normal distribution
- We can write a likelihood function as the product over characters (assuming characters evolve independently along the phylogeny)
- Want to estimate parameters – e.g., $x_0, i = 1, \ldots, p$ and $\sigma^2_i, i = 1, \ldots, p$
Problem with BM model: biological realism

- Recall that
  - Expected value of the trait at each tip is $x_0$ for all tips
  - Variance increases as a function of time

- BM model omits two important evolutionary concepts
  - Selection: we expect traits that are more “fit” (≡ better adapted to the environment) to be maintained in the population, while traits that are less fit die out
  - “Fitness” varies over
    - Time - e.g., ice age vs. present time
    - Lineages (branches in the tree) - e.g., speciation might happen due to available habitat

- Still want to include the process of genetic drift: random fluctuation in which traits are maintained or die out
The OU model

- **Goal:** develop a phylogenetic model to incorporate both selection and genetic drift

- **OU Process:**
  \[ dX_t = \alpha(\theta - X_t) + \sigma dW_t \]

- **Deterministic term:** \( \alpha(\theta - X_t) \)
  - Used to model selection
  - \( \alpha = \) strength of selection
  - \( \theta = \) optimum trait value
  - Force of selection is proportional to the distance of current trait value from the optimum \( (\theta - X_t) \)

- **Diffusion term:** \( \sigma dW_t \)
  - Used to model random genetic drift
  - \( \sigma = \) intensity of drift
The OU model in phylogenetics

Notes:

- If $\alpha=0$, the model reduces to BM
- Often called the Hansen model in phylogenetics
- Assume the molecular clock (distance from tip to root is the same for all species)

To apply the model, we need:

- Data for quantitative characters across species
- A phylogeny with branch lengths relating the species of interest
- Hypotheses regarding selective forces operating on each branch (or a method to estimate these)
We’ll look at variation in body size on the Lesser Antilles Islands

- Islands with two species: one large and one small
- Islands with one species: intermediate body size
- Hypothesis of interest is character displacement: did two initially intermediate-sized species come into contact and then evolve in opposite directions?
Models to be compared:

- BM or OU(1)
- OU(3)
- OU(4)
- OU(LP)

**Figure 5:** Alternative adaptive regime models for character displacement. Motion; -Uhlenbeck with one (1), three BM, one OU, or four optimal online parsimony (LP) reconstruction of colonization events (see text). Color codes indicate body size regimes.

Timescale of the phylogeny is standardized to 1.0 from most basal node to terminal species. Data and phylogeny reproduced from Losos (1990) and Butler and Losos (1997). All species belong to the Anolis bimaculatus series. Species name, body size (mean values for jaw length of upper one-third of male specimens in millimeters), and island name are given.

**Figure 6:** Full adaptive regime model for size dimorphism associated with habitat type. Color codes indicate separate adaptive regimes for the different habitat types (ecomorphs). Diameter of circles are proportional to the degree of sexual size dimorphism (log [male body length/female]). Data from Butler et al. (2000). Phylogeny (based on body length mtDNA) compiled from Jackman et al. (1997); Jackman et al. (1999; see Butler et al. 2000 for details). All species are within the genus Anolis.

Timescale of the phylogeny is standardized to 1.0 from most basal node to terminal species.

Results and Discussion.

The overall fit of each of the five models is summarized in table 1. Each OU model was tested against BM using the likelihood ratio test, and all...