Phylogenetic Comparative Methods

Laura Salter Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
The Ohio State University
lkubatko@stat.ohio-state.edu

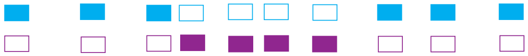
May 25, 2010

Introduction

- Thus far, we have focused on estimation of the phylogenetic tree as the primary goal of our analysis.
- Often, however, the phylogeny itself is not of interest; rather, it is a nusiance parameter.
- One setting in which this occurs is the case where we wish to study relationships among traits, either discrete or continuous, across taxa.
- We'll begin by examining some hypothetical data to motivate the method.

Example 1: Correlation between discrete traits

- Consider the following traits for 10 taxa:



Example 1: Correlation between discrete traits

- Consider the following traits for 10 taxa:

	□	🟦
□	0	6
🟪	4	0

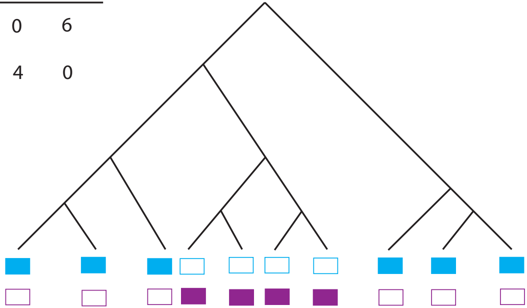
Fisher's exact test gives a p-value of 0.0048

🟦	🟦	🟦	□	□	□	□	🟦	🟦	🟦
□	□	□	🟪	🟪	🟪	🟪	□	□	□

Example 1: Correlation between discrete traits

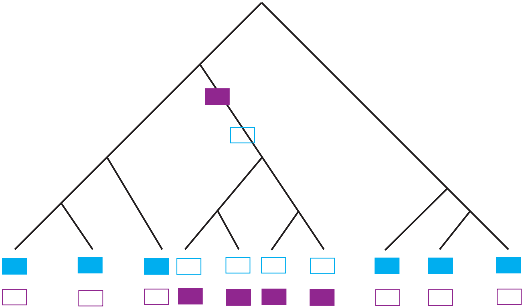
- But ....

	□	🟦
□	0	6
🟪	4	0



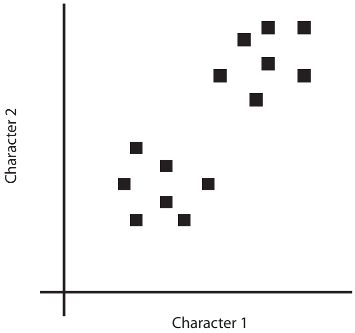
Example 1: Correlation between discrete traits

- Correlation is completely explained by phylogeny: the tree has 18 branches – the probability of two changes on the same branch is then $\frac{1}{18} = 0.056$.



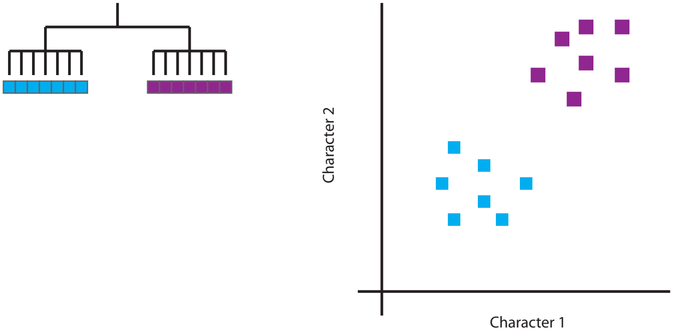
Example 2: Correlation between continuous traits

- Consider the following traits for 14 taxa:



Example 2: Correlation between continuous traits

- But .....



Brownian Motion

- The problem is that the taxa are not independent outcomes of the evolutionary process; they're all related by the phylogeny.
- We need to model the evolution of traits along the phylogeny and adjust appropriately for correlation due to shared evolutionary history.
- The simplest model of trait evolutionary along a phylogeny is the Brownian motion (BM) model.
- BM was proposed by Robert Brown (1773-1858) based on observation that pollen grains suspended in solutions "jiggled" continually. Later theoretical work is due to Einstein and Weiner, among others.
- BM was first applied to model trait evolution along phylogenies by Edwards and Cavalli-Sforza in 1964.

Brownian Motion

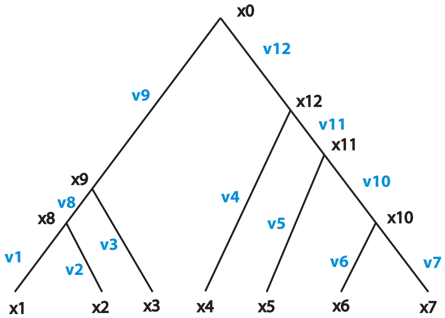
- Consider a particle moving in a single dimension (say along the x-axis).
- Measure position of particle at small intervals of time.
- The movement of the particle in each interval is assumed to:
  - be independent of movement in other intervals of time
  - have mean 0
  - have constant variance, regardless of position of the particle
- Consider *n* distinct intervals of time.
- After *n* steps, the net displacement is the sum of the displacements at each step.

Brownian Motion

- Let s^2 be the variance of the displacement at each interval. The variance in the net displacement is then ns^2 (since the displacement in intervals is independent).
- Now let $s^2 \rightarrow 0$ as $n \rightarrow \infty$ such that their product is constant. This is the BM or Weiner process.
- What is important for us is the distribution of the net displacement after t units of time.
- Let σ^2 be the variance per unit time.Then after t units of time, the variance of the net displacement is $\sigma^2 t$.
- In addition, the net displacement across an interval of t units of time is $N(0, \sigma^2 t)$.

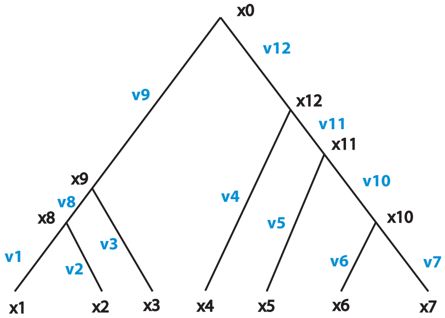
BM Along a Phylogeny

- How do we apply this to trait evolution along a phylogeny?
- Assume that displacements on different branches of a tree are independent; traits evolve over branches of tree according to BM model.



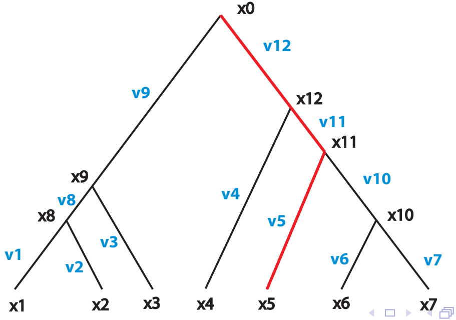
BM Along a Phylogeny: An Example

- Let *xi* denote the phenotype at node *i*, and let *vi* denote the branch length.
- As an example, look at value of trait at external node 5, *x*5.



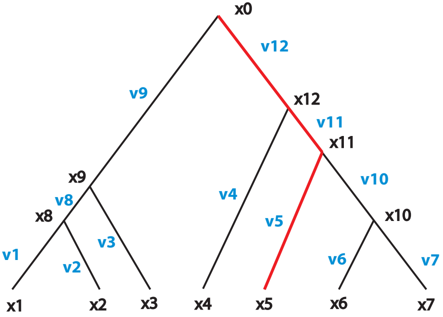
BM Along a Phylogeny: An Example

- 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.



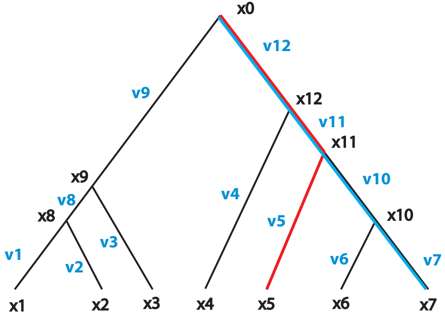
BM Along a Phylogeny: An Example

- We have
  - $E(x_5) = x_0$
  - $Var(x_5) = \sigma^2 v_{12} + \sigma^2 v_{11} + \sigma^2 v_5$



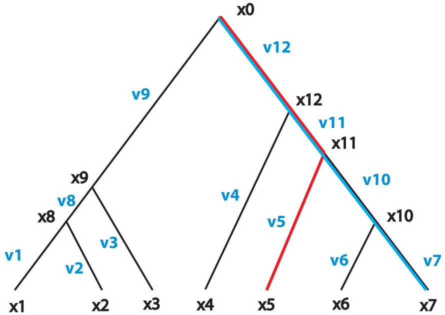
BM Along a Phylogeny: An Example

- 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}
- $cov(x_5, x_7) = \sigma^2 v_{12} + \sigma^2 v_{11}$



BM Along a Phylogeny

- 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:

$$\begin{pmatrix}
v_1+v_8+v_9 & v_8+v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_8+v_9 & v_2+v_8+v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_9 & v_9 & v_3+v_9 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & v_4+v_{12} & v_{12} & v_{12} & v_{12} \\
0 & 0 & 0 & v_{12} & v_5+v_{11}+v_{12} & v_{11}+v_{12} & v_{11}+v_{12} \\
0 & 0 & 0 & v_{12} & v_{11}+v_{12} & v_6+v_{10}+v_{11}+v_{12} & v_{10}+v_{11}+v_{12} \\
0 & 0 & 0 & v_{12} & v_{11}+v_{12} & v_{10}+v_{11}+v_{12} & v_7+v_{10}+v_{11}+v_{12}
\end{pmatrix}$$

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., $x_{0i}, i = 1, \cdots p$ and $\sigma_i^2, i = 1 \cdots p$
- However, we run into a problem – likelihood goes to ∞ as branch length goes to 0 (see Ch. 23 in text for a worked example).

Inference Under the BM Model

- Some possible solutions:
  - Assume a molecular clock
  - Recode data as differences between character states at the tips – contrasts
- Look more at the second case – let **C** be an $(n-1) \times n$ matrix of contrasts.
- Let **V** be the original variance-covariance matrix (multiplied by σ^2).
- Then for the data re-coded as differences in trait values, which we denote by the vector **y**, we have

$$\mathbf{y} = \mathbf{Cx} \sim N(\mathbf{0}, \mathbf{CVC^T})$$

Inference Under the BM Model

- Next, we'll discuss computing the likelihood under this model.
- Need a notion of independent contrasts.