Correlations of Discrete Characters, and Ancestral State Reconstruction

April 9, 2009

The Mk Model

- Simple model for the evolution of discrete variables

- Constant instantaneous probability of changing between any two character states out of k possibilities

- Described by **rate matrix** Q:

$$Q = \alpha \begin{bmatrix} 1-k & 1 & \dots & 1 \\ 1 & 1-k & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1-k \end{bmatrix}$$

The Mk Model

- Generalized JC model (Mk with k=4)

- At equilibrium, stationary distribution of characters is [1/k, 1/k, ... , 1/k]

- Described by transition matrix

- m = exp(Qt)

Alternatives to Mk
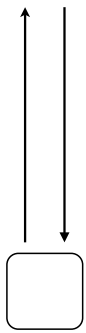
$$Q = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$$
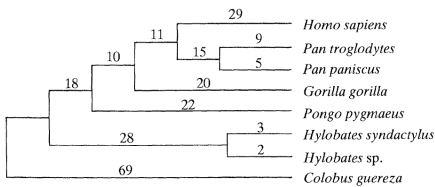
Example

Figure 2. A phylogeny of the Hominoidea including one Old World Monkey. Two characters, social system and oestrous advertisement are scored for each species, where s = single male social system, m = multi-male, and ' + ' = the presence of oestrous swellings, ' − ' absence. *Homo sapiens* (s, −); *Pan* (m, +); *Gorilla* (s, −); *Pongo* (m, −); *Hylobates* (s, −); *Colobus guereza* (s, −).

Tree tips: 29 *Homo sapiens*; 11, 9 *Pan troglodytes*; 15, 5 *Pan paniscus*; 10 *Gorilla gorilla*; 20 *Pongo pygmaeus*; 18, 22 ; 28, 3 *Hylobates syndactylus*; 2 *Hylobates* sp.; 69 *Colobus guereza*.

Consider two discrete characters

- Are they evolving in a correlated fashion?

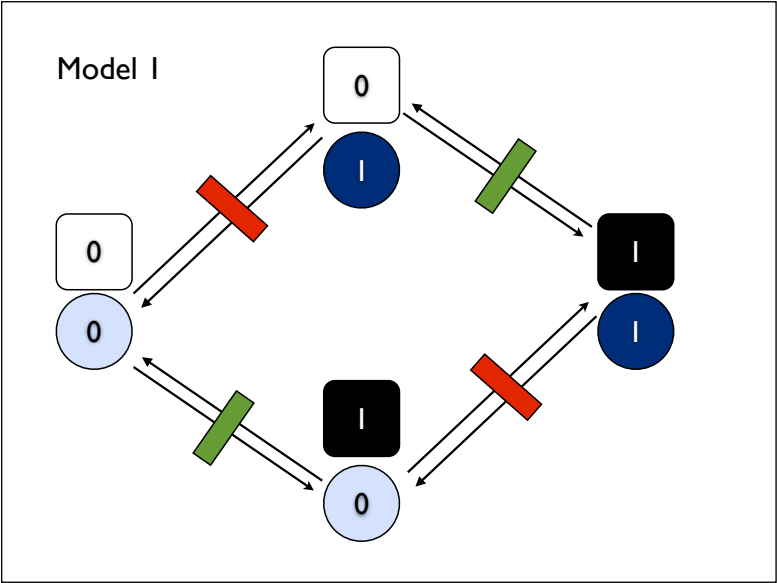- When character A changes to state 1, is character B likely to change to 1 as well?

Character 1

Character 2

Model I

Model II
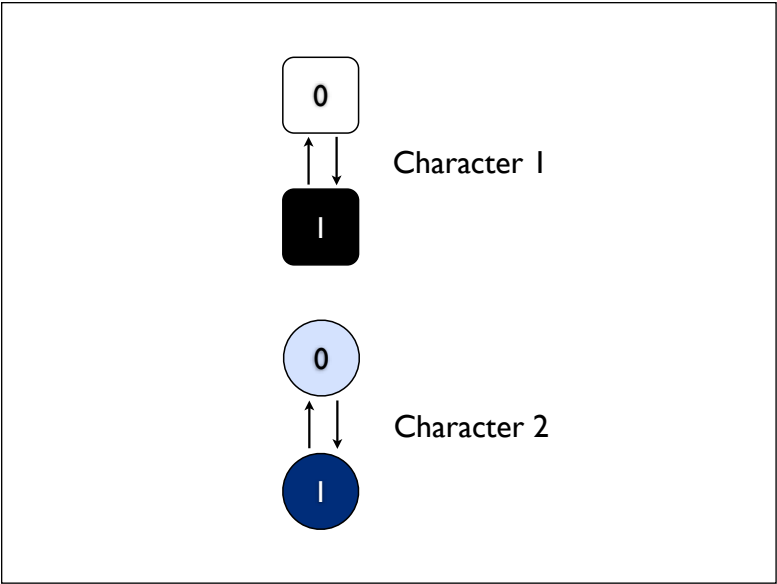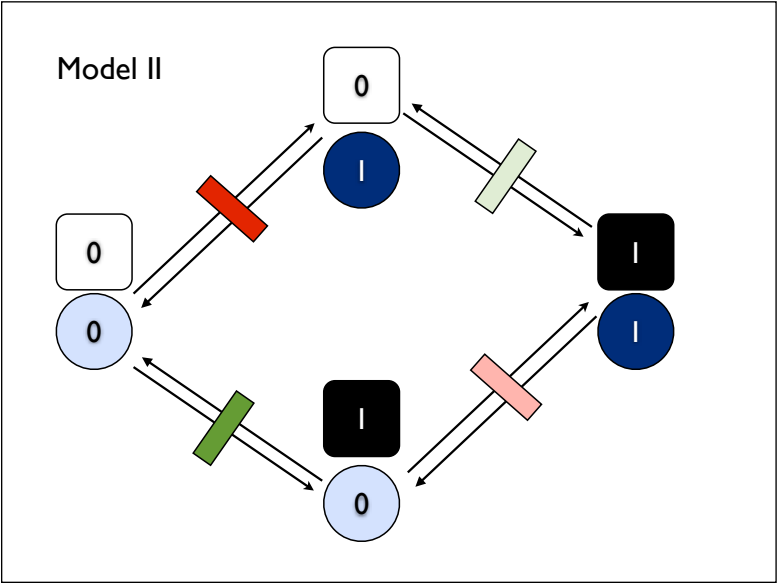
Correlation of Discrete Characters

- Does the rate of change of character 2 depend on the state of character 1?
- Compare models using maximum likelihood

X,Y	X 0 Y 0	0 1	1 0	1 1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	$1-(q_{21}+q_{24})$	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

X,Y	X 0 Y 0	0 1	1 0	1 1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	$1-(q_{21}+q_{24})$	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

Model A: constant rates characters independent

X,Y	X: 0, Y: 0	X: 0, Y: 1	X: 1, Y: 0	X: 1, Y: 1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	$1-(q_{21}+q_{24})$	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

Model B: characters have different rates independent evolution

X,Y	X: 0, Y: 0	X: 0, Y: 1	X: 1, Y: 0	X: 1, Y: 1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	$1-(q_{21}+q_{24})$	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

X,Y	X: 0, Y: 0	X: 0, Y: 1	X: 1, Y: 0	X: 1, Y: 1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	$1-(q_{21}+q_{24})$	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

X,Y	X:0 Y:0	X:0 Y:1	X:1 Y:0	X:1 Y:1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	**Model C: symmetric characters dependent**	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

X,Y	X:0 Y:0	X:0 Y:1	X:1 Y:0	X:1 Y:1
0,0	$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1	q_{21}	**Model D: asymmetric characters dependent**	0	q_{24}
1,0	q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1	0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

Example

- Data on evolution of dewlap color (red, yellow) and perch behavior (high, low) of anoles

- Fit the four models to these data

- Obtain the following results:

Example

Model	lnL	k	AIC
A	-132.2	1	
B	-130.1	2	
C	-125.3	4	
D	-124.2	8	

Example

29 Homo sapiens
11
9 Pan troglodytes
10 15
5 Pan paniscus
18
20 Gorilla gorilla
22 Pongo pygmaeus
28
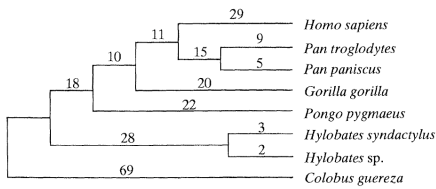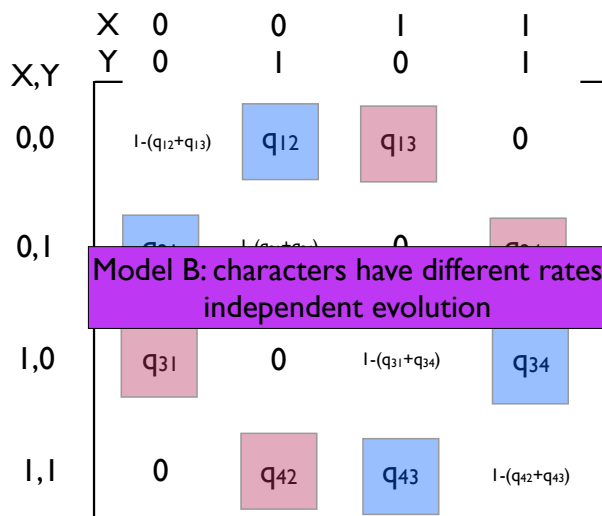3 Hylobates syndactylus
2 Hylobates sp.
69 Colobus guereza

Figure 2. A phylogeny of the Hominoidea including one Old World Monkey. Two characters, social system and oestrous advertisement are scored for each species, where s = single male social system, M = multi-male, and ' + ' = the presence of oestrous swellings, ' − ' absence. *Homo sapiens* (s, −); *Pan* (M, +); *Gorilla* (s, −); *Pongo* (M, −); *Hylobates* (s, −); *Colobus guereza* (s, −).

X	0	0	1	1
X,Y Y	0	1	0	1
0,0	1-(q_{12}+q_{13})	q_{12}	q_{13}	0
0,1	q_{21}	1-(q_{21}+q_{24})	0	q_{24}
1,0	q_{31}	0	1-(q_{31}+q_{34})	q_{34}
1,1	0	q_{42}	q_{43}	1-(q_{42}+q_{43})

Model B: characters have different rates independent evolution

Slide 1

	X	0	0	1	1
X,Y	Y	0	1	0	1
0,0		$1-(q_{12}+q_{13})$	q_{12}	q_{13}	0
0,1		q_{21}		0	q_{24}
1,0		q_{31}	0	$1-(q_{31}+q_{34})$	q_{34}
1,1		0	q_{42}	q_{43}	$1-(q_{42}+q_{43})$

Model C: symmetric characters dependent

Slide 2

Model B: characters have different rates independent evolution

$\ln L = -11.91$

Model C: symmetric characters dependent

$\ln L = -8.43$

$$\Delta = 2*(-8.43 - -11.91) = 6.96$$
$$P = 0.12$$

Slide 3

Reconstructing Ancestral Character States

- Goal: to make inferences about the states of characters in ancestral lineages
- Usually make these inferences at nodes in the tree
- These inferences can be made using maximum likelihood
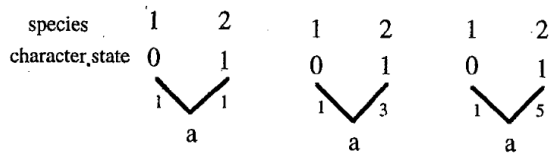
Discrete Ancestral State Reconstruction

0   1

1.0   1.0
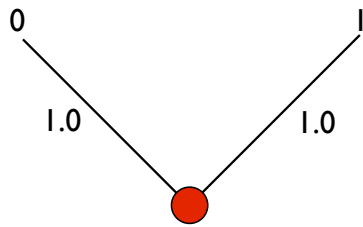
$L = P(d \mid m) = 0.250$

$L(a=0 \mid d, m) = 0.125$

$L(a=1 \mid d, m) = 0.125$
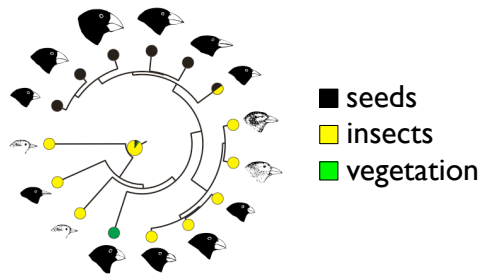
Model: Mk with $q_{01} = q_{10} = 8$

species	1 2	1 2	1 2
character state	0 1	0 1	0 1
	1 1	1 3	1 5
	a	a	a
$\hat{L}(d)$	0.250	0.256	0.296
$\hat{L}(a = 0 / d)$	0.125	0.236	0.291
$\hat{L}(a = 1 / d)$	0.125	0.125	0.125

$w(a=0)=0.291/(0.291+0.125) = 0.70$

$w(a=1)=0.125/(0.291+0.125) = 0.30$

■ seeds
□ insects
■ vegetation

Local and global

- There is a difference between:

  - the set of characters that, together, maximize the likelihood

  - the character that represents the maximum-likelihood estimate for a single node (marginal likelihood)

- Pagel (1999) even recommends reestimating model parameters (q_{01}, etc.) independently for each ancestral state (the "local" estimate)

Continuous Characters

- Can also reconstruct ancestral characters using likelihood

- Assume characters are evolving under a Brownian motion model

$$\Pr[y = x] = \frac{1}{\sigma\sqrt{2\pi t}}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2 t}\right)$$

Continuous characters

- ML ancestral states for continuous characters under BM are equivalent to ancestral states estimated under weighted squared change parsimony

- Choose the ancestral states that minimize the amount of squared change on the tree

Maximum likelihood

- Find model parameters that maximize the likehood of obtaining the data

- Calculate $P(a=x_i \mid model, tree, data)$

- Can generate confidence intervals based on t-distribution (Schluter 1997)
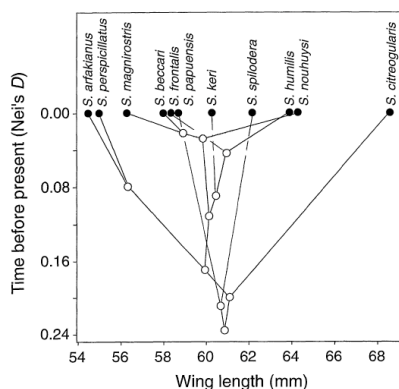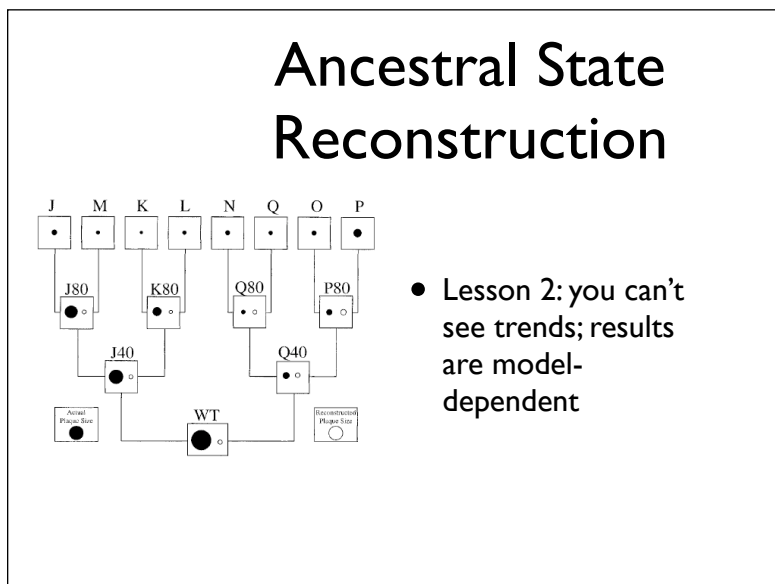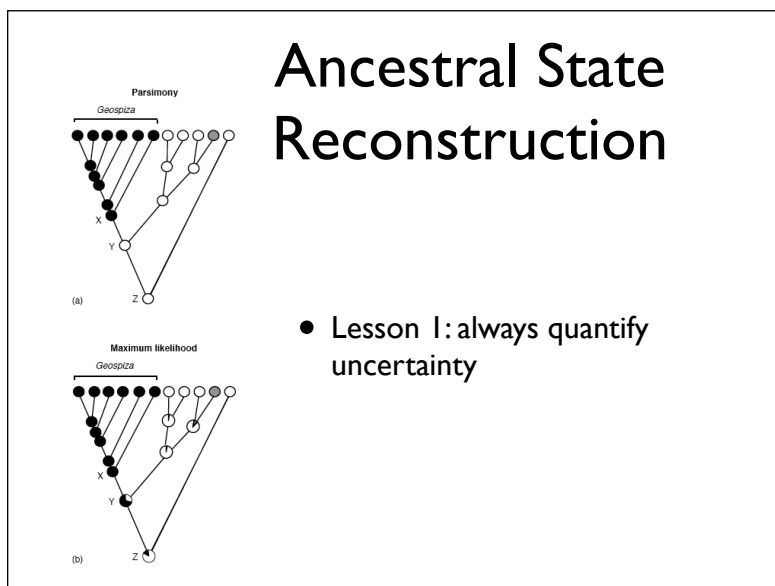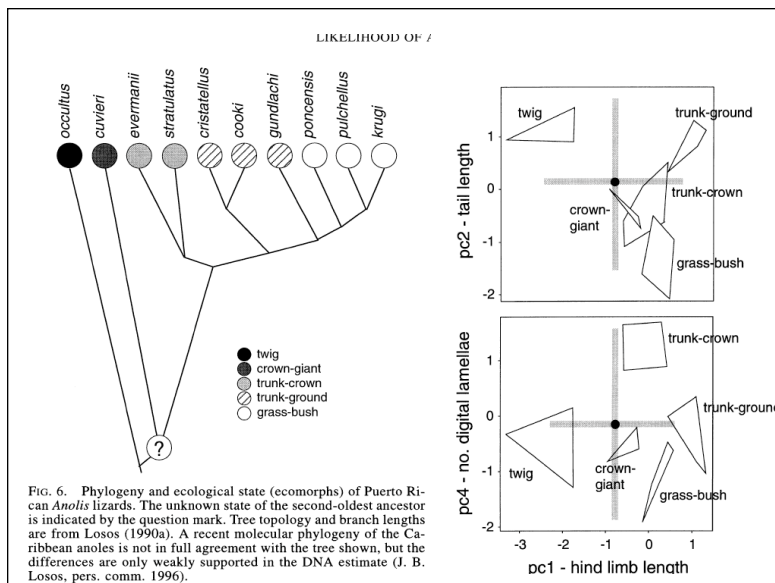

FIG. 8. Wing lengths of Australian and New Guinea scrubwrens (*Sericornis*) and of their ancestors. The tree is the same as in Figure 3. Shaded lines span the support limits for each ancestor state. Marginal distributions were computed separately for each ancestor one at a time.

FIG. 6. Phylogeny and ecological state (ecomorphs) of Puerto Rican *Anolis* lizards. The unknown state of the second-oldest ancestor is indicated by the question mark. Tree topology and branch lengths are from Losos (1990a). A recent molecular phylogeny of the Caribbean anoles is not in full agreement with the tree shown, but the differences are only weakly supported in the DNA estimate (J. B. Losos, pers. comm. 1996).

Ancestral State Reconstruction

- Lesson 1: always quantify uncertainty

Ancestral State Reconstruction

- Lesson 2: you can't see trends; results are model-dependent

Summary

- We can test for character correlations, or other patterns of interest, with discrete characters

- It is possible to estimate ancestral character states and incorporate uncertainty