"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

Integrative Biology 200a Spring 2008

University of California, Berkeley B.D. Mishler

Dating in the 21st Century: putting dates on nodes, characters, and events

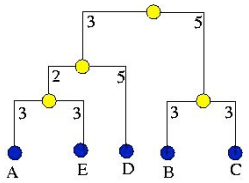
This is an area of intense recent interest, yet in need of much further thought and research. There are two fundamental steps in the process of putting time onto a node or a branch of a tree:

1. Establishing the clock

What is ultrametricity?

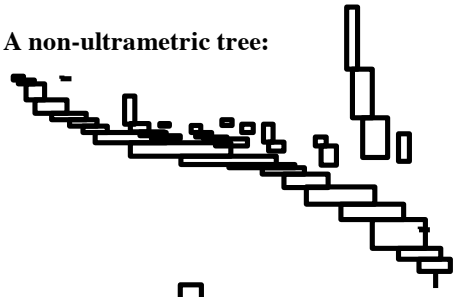
Ultrametric matrix and its tree:

	A	B	C	D	E
A		16	16	10	6
B			6	16	16
C				16	16
D					10
E					



from: http://www.diku.dk/~pawel/comp-bio/ev_trees/intro/intro/ultrametric.html

A non-ultrametric tree:



A. Determining whether your data fit a clock model

i. relative rate tests

comparing three taxa at a time, in rooted context:

ii. likelihood ratio test

Testing the Molecular Clock using a likelihood ratio test (courtesy of John Huelsenbeck)

Under the null hypothesis, the phylogeny is ultrametric (i.e., rooted and the branch lengths are constrained such that all of the tips can be drawn at a single time plane). Under the alternative hypothesis, each branch is allowed to vary independently. The alternative hypothesis invokes s - 2 additional parameters, where s is the number of sequences. The likelihood ratio test statistic is -2logL = 2(logL0 - logL1), where L0 and L1 are the likelihoods under the null and alternative hypotheses, respectively.

The significance of the likelihood ratio test statistic can be approximated using a $\chi 2$ distribution (with s - 2 degrees of freedom) or by parametric bootstrapping.

The following example shows how to perform the likelihood ratio test of the molecular clock. The data are s = 5 albumin sequences from vertebrates (a fish, frog, bird, mouse, and human). We assume the Hasegawa, Kishino, and Yano (1985) model of DNA substitution with among site rate variation described using a gamma distribution.

The maximum likelihood under the null hypothesis is logL0 = -7585.343. The best estimate of phylogeny supports the monophyly of the mammals and amniotes.

The maximum likelihood under the alternative hypothesis is logL1 = -7569.052. The likelihood under the alternative hypothesis is higher than under the null hypothesis because there are more free parameters in the substitution model (i.e., no constraints on branch lengths). The maximum likelihood estimate of phylogeny is consistent with the monophyly of mammals and amniotes (though the tree is unrooted).

The likelihood ratio test statistic is -2logL = 32.582, which is asymptotically $\chi 2$ distributed under the null hypothesis with 3 degrees of freedom. Comparing the observed value of -2logL to a $\chi 2$ with 3 df shows that the null hypothesis can be rejected at $P < 0.001$. So, we conclude the data are <u>not</u> clock-like.

B. If your data don't fit a clock model (and they usually don't), try smoothing the data to get an (at least locally) approximate clock. Two common methods (implemented in ***r8s*** by Mike Sanderson: http://loco.biosci.arizona.edu/r8s/index.html), both attempt to smooth the magnitude of changes in rate between neighboring branches, to give you something intermediate between the rigid clock assumption and completely unconstrained branch lengths:

i. non-parametric rate smoothing. This uses a least squares smoothing approach that penalizes rates that change too quickly from branch to neighboring branch.

ii. penalized likelihood. This is a "semi-parametric" approach that combines a ML approach with the above penalty function. The user can specify the relative weight of the penalty function and the ML component (in which parameters are being fitted as typical in ML). The parametric model has a different substitution rate for each branch.

2. Calibrating the clock

Some folks simply import a "known" rate from the literature into their analysis -- don't do this! You need to come up with a calibration from your analysis.

A. Three ways that have been used to estimate the age of a node

i. a fossil (see below for details) -- gives a *minimum* age for a node

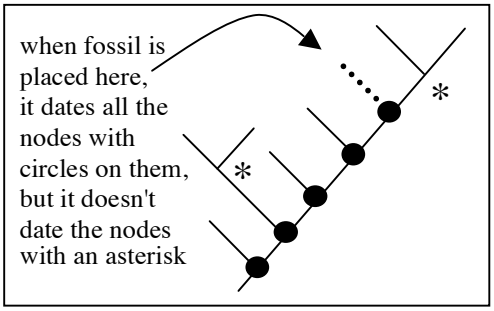
ii. availability of necessary habitat -- gives a *maximum* age for a node (maybe)

iii. geographic vicariance event -- neither a *maximum* or *minimum* age for a node

B. How to use a fossil to date a node? Some principles:

i. you never find a taxon in the fossil record, or a lineage; you find remains of an organism displaying some *characters*. These characters need to be analyzed using the principles talked about earlier in class, in relation to other fossils and extant organisms in the group.

ii. therefore, a fossil can never be compared to a strictly molecular phylogeny (unless it has preserved molecular data!); all relevant morphological characters need to have been analyzed and incorporated in the phylogenetic reconstruction.



iii. When a fossil can be placed using synapomorphies as sister to some other lineage, that other lineage (and the node connecting them) must be at least as old as the fossil. Nodes deeper must also have been in existence by that time. This is the important principle of ***equal age of sister groups***.

C. Rules of thumb:

i. For many questions in evolutionary biology you don't need absolute time anyway; relative time will do (i.e., ordering of nodes in time). So, don't bother with clocks unless you need them.

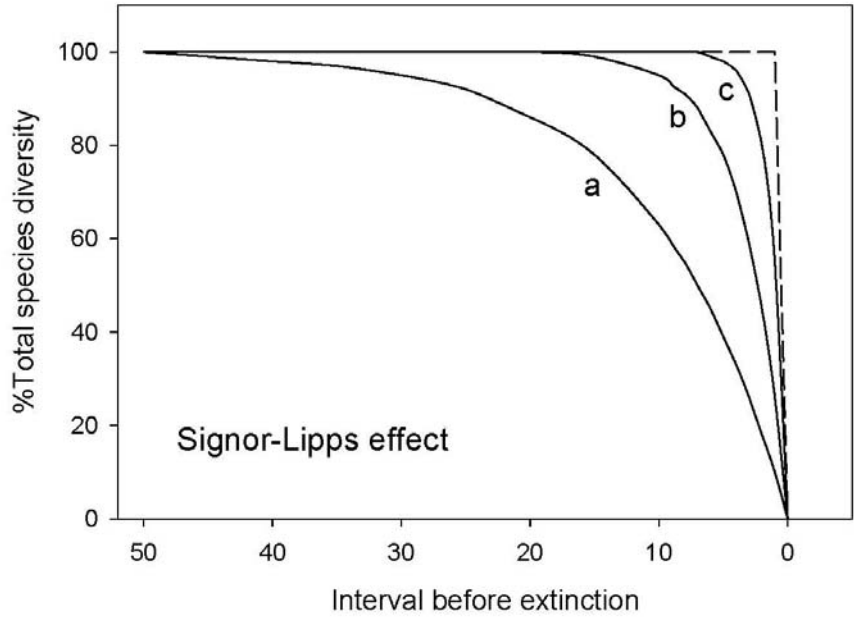
ii. If you do need to calibrate a clock, you want to have as many calibration points (preferably fossils), as local to your questions, as possible.

ii. If you have enough calibration points, you don't need an actual molecular clock (or even a manufactured one) to answer many questions.

iii. As always, carefully consider what questions you want to address first, then select your approach; for every positive hypothesis, be sure you have a null hypothesis.

D. Why does the molecular clock often indicate earlier divergence dates?

i. The Signor-Lipps effect



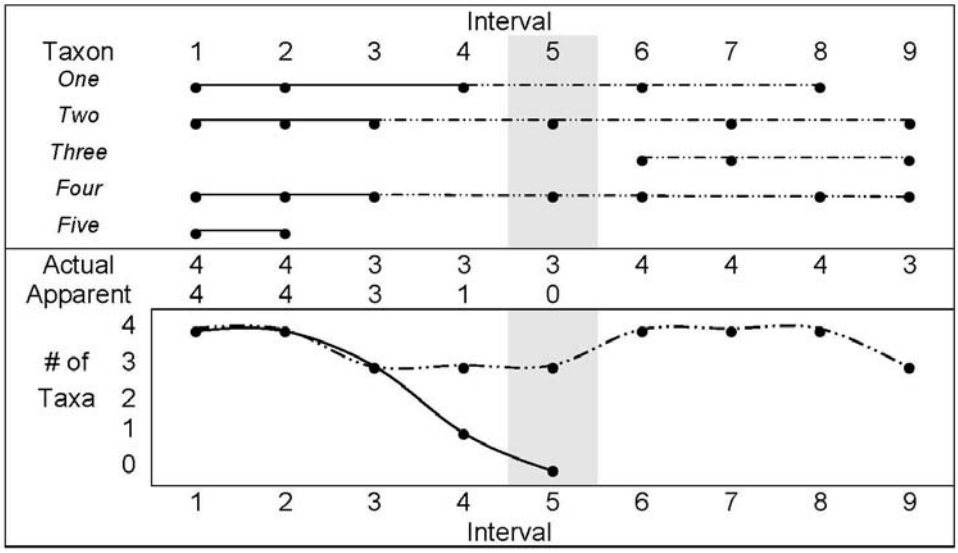
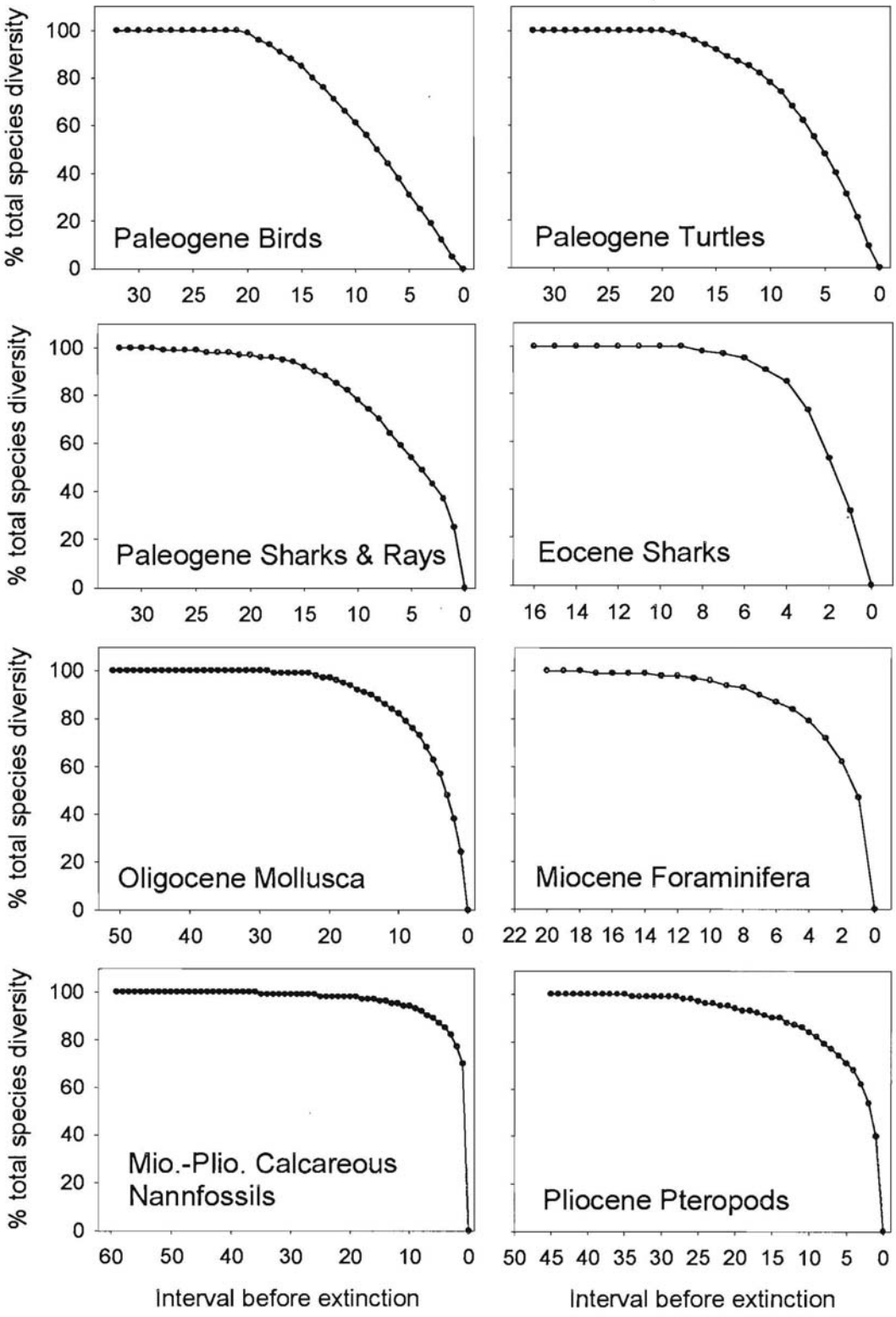


Fig 2. Simulated data set with a mass extinction event at interval 5. Rows = taxa, columns = intervals. Solid circles denote taxon occurrences. Solid lines define stratigraphic ranges prior to extinction event, dashed lines indicated range "after" extinction event. Lower half of figure graphically presents apparent vs. actual diversity curves. Although actual taxon diversity fluctuated between 3-4 in this example, an distinct Signor-Lipps effect is apparent in the data.


% total species diversity
Paleogene Birds
Paleogene Turtles
Paleogene Sharks & Rays
Eocene Sharks
Oligocene Mollusca
Miocene Foraminifera
Mio.-Plio. Calcareous Nannfossils
Pliocene Pteropods
Interval before extinction
Interval before extinction