Phylogenetic Biology
Week 8

Biology 1425

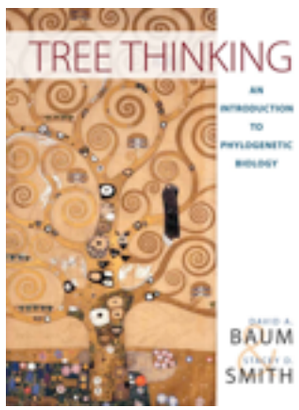
Professor: Casey Dunn, dunnlab.org
Brown University

2013

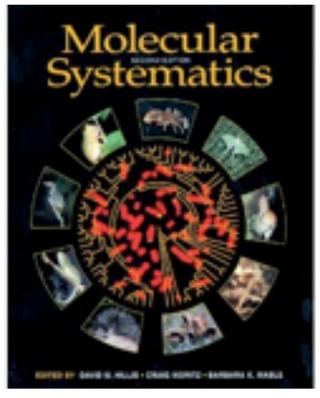
Front matter...

Sources

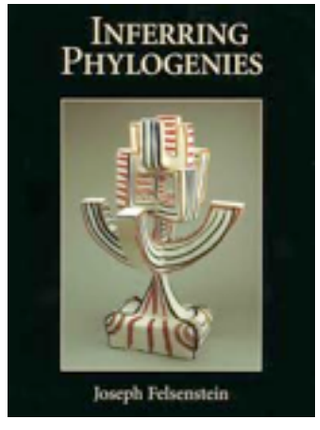
Some non-original content is drawn from:



Baum, D and S. Smith (2012) Tree Thinking: and Introduction to Phylogenetic Biology. Roberts and Company Publishers. ISBN 9781936221165



Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). Phylogenetic inference. In: Molecular Systematics, Second Edition. eds: D. M. Hillis, C Moritz, & B. K. Mable. Sinauer Associates. ISBN 9780878932825



Felsenstein, J. (2003) Inferring Phylogenies. Sinauer Associates. ISBN 978-0878931774

Other non-original content is referenced by url.

Testing for non-randomness in character data

Permutation Tail Probability Test

Permute (randomize) the order of the data within each column of the matrix

Find the shortest parsimony tree for the permuted matrix

Repeat many times (eg 1000), storing the length of the shortest trees each time

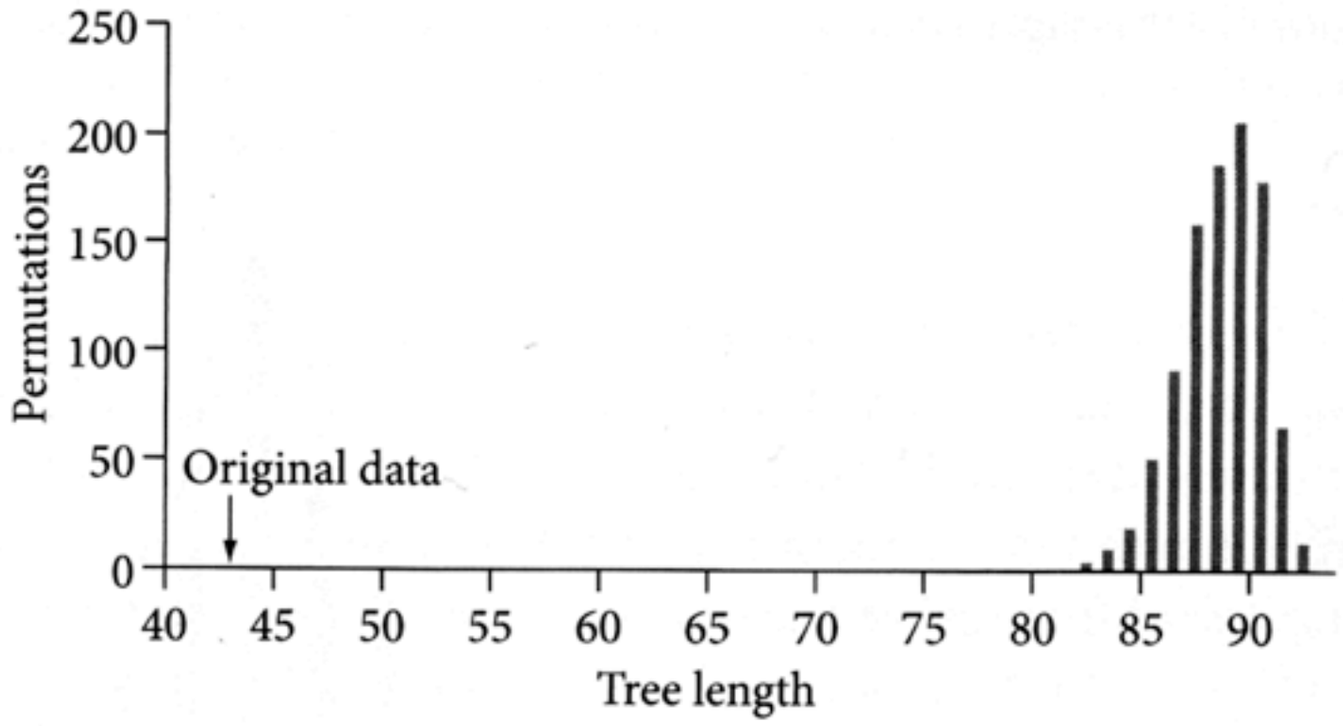
Compare the length of the shortest tree for the actual matrix to the distribution of lengths for all the permuted matrices

Permutation Tail Probability Test

Original data					
	1	2	3	4	5
A	A	G	T	C	T
B	A	G	T	C	T
C	C	G	T	C	A
D	C	A	C	A	A
E	C	A	C	G	C

Permutation 1					
	1	2	3	4	5
A	C	A	C	G	T
B	A	G	C	C	A
C	C	A	T	A	C
D	A	G	T	C	A
E	C	G	T	C	T

Permutation 2					
	1	2	3	4	5
A	A	G	T	A	A
B	C	A	C	C	T
C	C	G	T	C	C
D	C	A	C	C	A
E	A	G	T	G	T



Baum and Smith 2012, Figure 9.2

A general framework for implementing tests

Simulate data under the null hypothesis (eg, data are random)

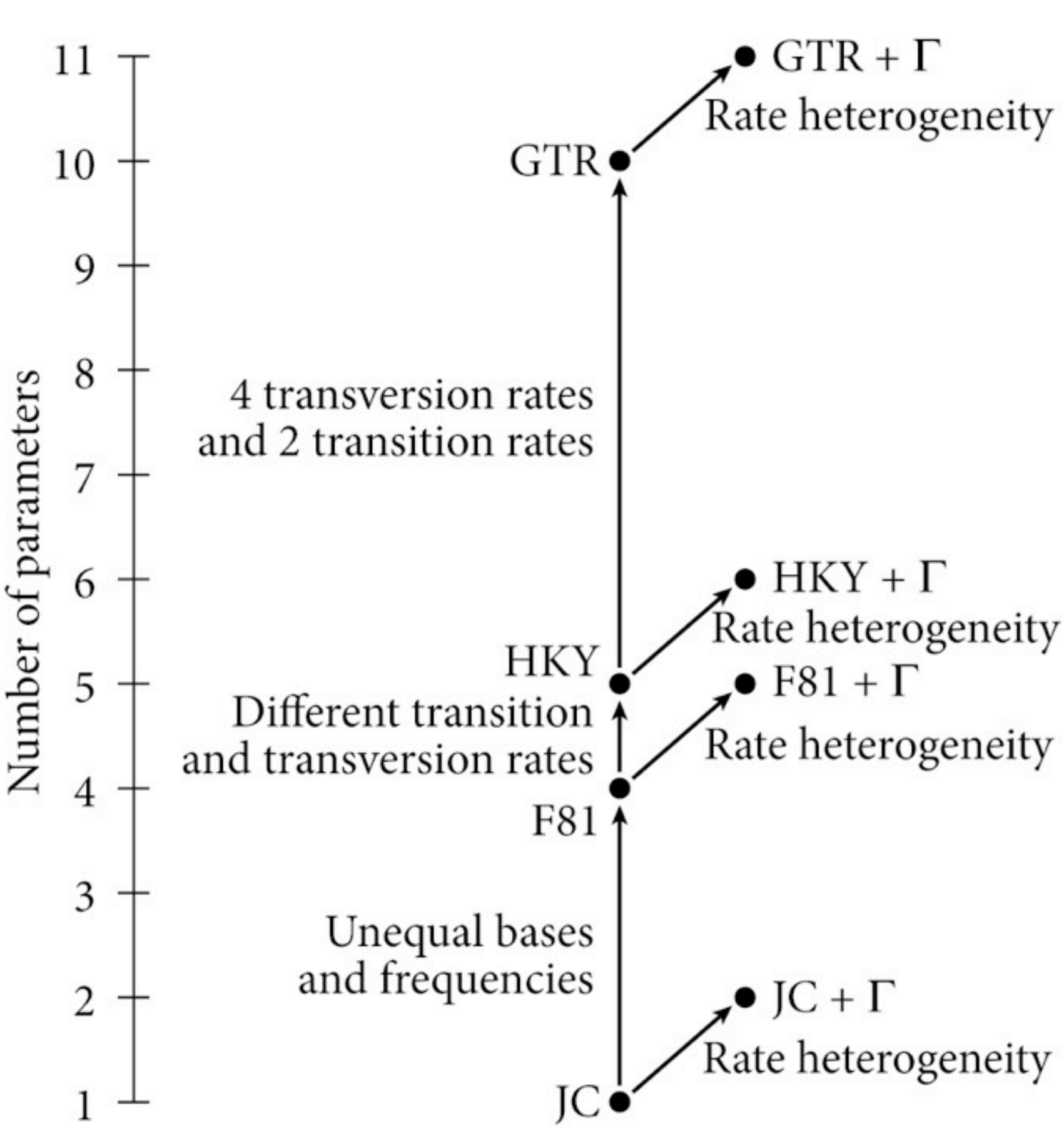
Make a measurement on the simulated data

See if the same measurement made on the real data is greater than or less than a large fraction of the measurements made on simulations.

If so, reject the null

Testing alternative models

Many hypotheses are nested, i.e. they can be formed from alternative hypotheses by adding parameters



Baum and Smith 2012, Figure 8.10

Adding parameters always increases the likelihood, so a change in likelihood alone can’t be taken as an indication that one model is a better fit than another.

For nested models, one can use the likelihood ration test to evaluate the magnitude of the difference between likelihoods.

The test statistic LR:

$$LR = 2 * (ln(L_1) - ln(L_2))$$

Where L_1 is the likelihood of the more complex model, and L_2 is the likelihood of the simpler model.

LR is distributed roughly as a chi-squared distribution, where the number of degrees of freedom is the difference in number of parameters between the models.

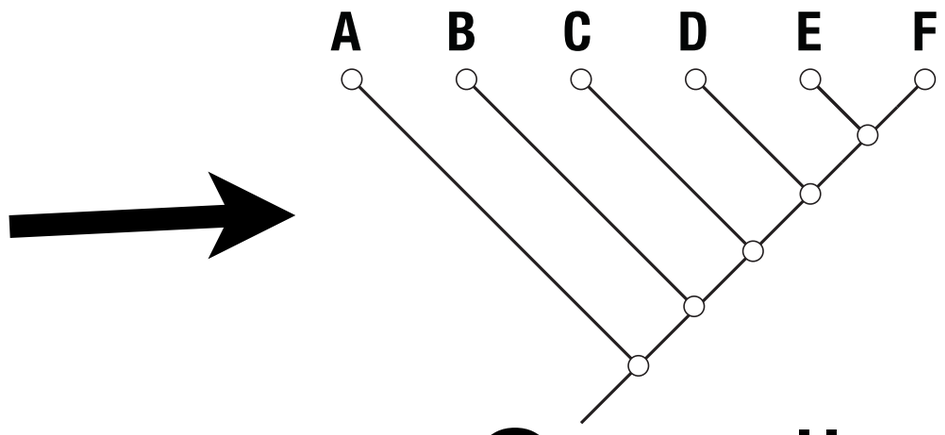
Forms the foundation of the program modeltest, http://code.google.com/p/jmodeltest2/

Testing alternative hypotheses about tree topology

The likelihood ratio test is sometimes used to test alternative topologies.

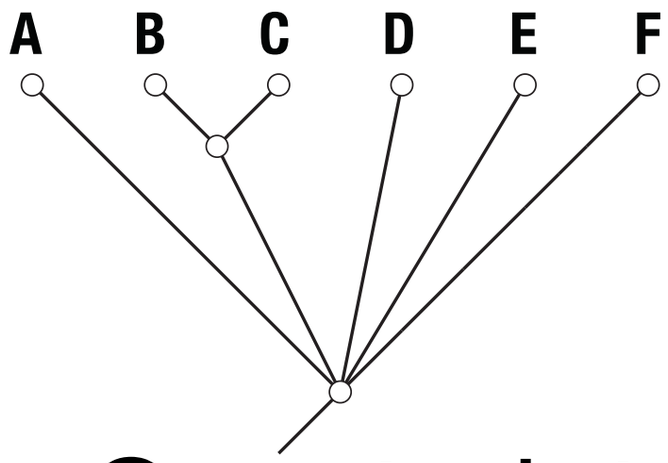
Can I reject a particular relationship that isn’t recovered in the maximum likelihood tree?

Data, model

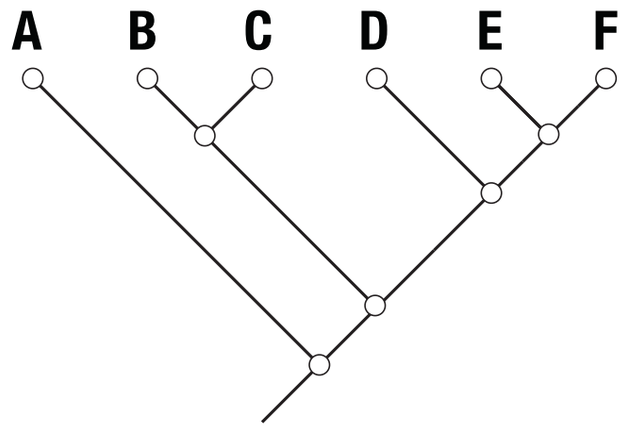


Overall
ML Tree

Data, model,

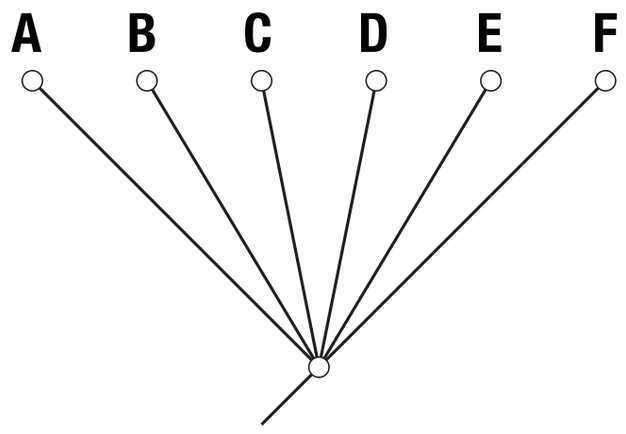


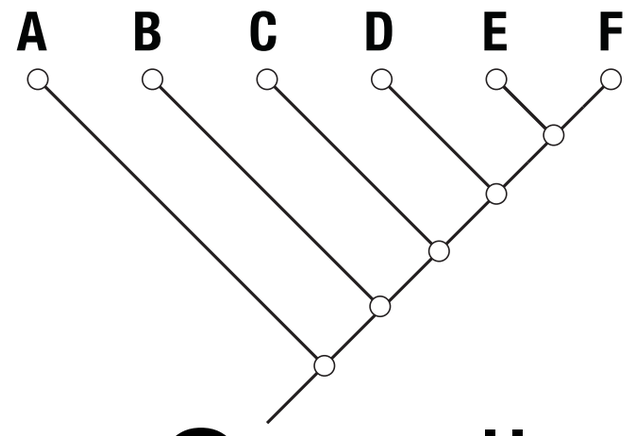
Constraint



Constrained
ML Tree

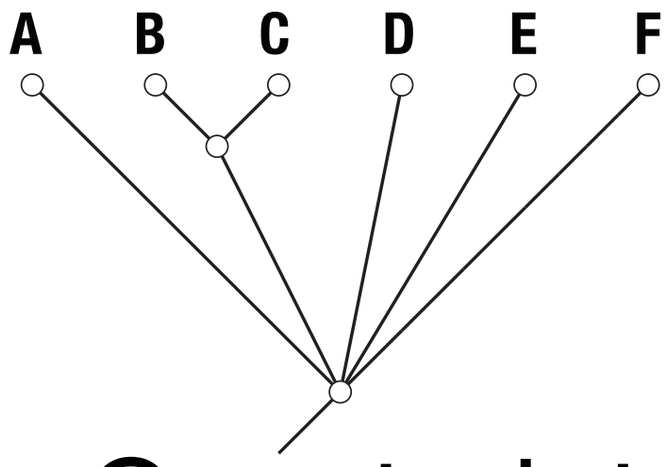
Data, model,



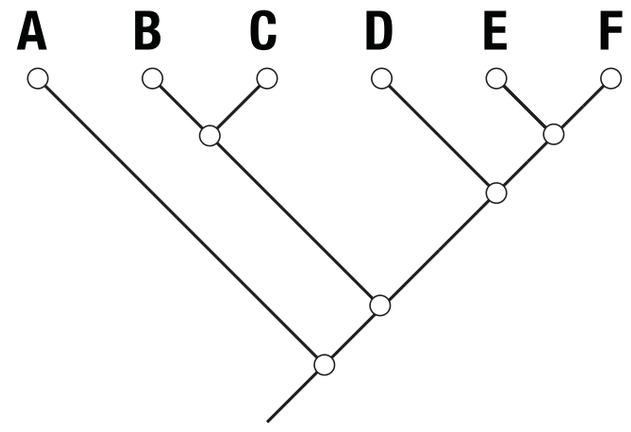


Overall
ML Tree

Data, model,



Constraint



Constrained
ML Tree

Will the overall ML tree or the constrained ML tree have a higher likelihood?

Syst. Biol. 49(4):652–670, 2000

Likelihood-Based Tests of Topologies in Phylogenetics

NICK GOLDMAN,[1] JON P. ANDERSON,[2] AND ALLEN G. RODRIGO[3]

[1] *University Museum of Zoology, Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK; E-mail: N.Goldman@zoo.cam.ac.uk*

[2] *Department of Molecular Biotechnology, University of Washington, Seattle, Washington USA*

[3] *School of Biological Sciences, University of Auckland, Auckland, New Zealand*

Kishino–Hasegawa (KH) Test

Two hypotheses:

Tree 1 (T_1) and Tree 2 (T_2)

Each with likelihood L_1 and L_2

Are L_1 and L_2 significantly different?

Kishino–Hasegawa (KH) Test

$$\delta \equiv L_1 - L_2$$

$$H_0 : E[\delta] = 0$$

$$H_A : E[\delta] \neq 0$$

Kishino–Hasegawa (KH) Test

Need to know the distribution of δ to use the test statistic

Non-parametric bootstrapping:

- Resample the data to create pseudoreplicate i
- Calculate $L_1^{(i)}$ and $L_2^{(i)}$, then $\delta^{(i)}$
- Repeat many times
- Recenter the distribution of $\delta^{(i)}$
- See of the obtained δ falls within the distribution of $\delta^{(i)}$

Kishino–Hasegawa (KH) Test

Is this an OK use of KH?

I have two alternative hypotheses about the relationships of my species. I go collect new data, calculate δ, and then perform my test.

Kishino–Hasegawa (KH) Test

Is this an OK use of KH?

I have two alternative hypotheses about the relationships of my species. I go collect new data, calculate δ, and then perform my test.

Yes.

Kishino–Hasegawa (KH) Test

Is this an OK use of KH?

I have a hypothesis about the relationships of my species. I go collect new data, and find that the ML tree differs from my original hypothesis. I calculate δ for the ML tree and my tree, and then perform my test.

Kishino–Hasegawa (KH) Test

Is this an OK use of KH?

I have a hypothesis about the relationships of my species. I go collect new data, and find that the ML tree differs from my original hypothesis. I calculate δ for the ML tree and my tree, and then perform my test.

No - there is no expectation that E[δ] = 0

Non-parametric bootstrapping

Make new datasets by resampling the original dataset

Parametric bootstrapping

Make new datasets by simulating them under a given model

SOWH Test

H_0: T_1 is the true topology

H_A: some other topology is true

- Calculate the test statistic $\delta \equiv L_{\mathrm{ML}} - L_1$.
- Simulate data sets i by parametric bootstrapping, based on the null hypothesis topology T_1 and the ML estimates of any free parameters, $\hat{\theta}_1$, derived for T_1 from the original data set.
- Use T_1 and reestimate free parameters θ_1 to get maximized log-likelihoods $L_1^{(i)}$ under H_0.
- Maximize likelihood over all topologies T_x and their respective parameters θ_x to get log-likelihoods $L_{\mathrm{ML}}^{(i)}$.
- Calculate values of $\delta^{(i)} \equiv L_{\mathrm{ML}}^{(i)} - L_1^{(i)}$, the set of these giving an estimate of the distribution (under H_0) of δ.
- Test whether the attained value of δ(from the original data) is a plausible sample from the estimated distribution of δ given by the set of the $\delta^{(i)}$ by seeing if it falls below the 95% point (for example) of the ranked list of the $\delta^{(i)}$. Such a one-sided test is appropriate because we know that δ must be >0; in this example, a 5% significance level is being used.

As described by Goldman et al. 2000