

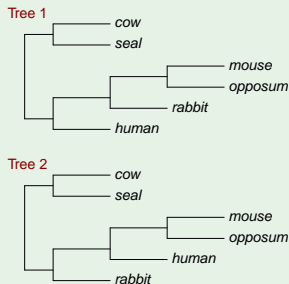
Topology Tests

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

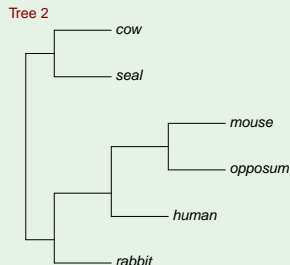
Two Main Topology Test Problems

Two trees



- Is Tree 1 significantly better than Tree 2

One Tree



- Is Tree 2 plausible?

Spanish Scores Before/After Course

Subject	1	2	...	20
Before	30	28	...	29
After	29	30	...	32
d_i	-1	2	...	3

$$\bar{d} = 1.45, s_d = 3.2, \\ n = 20$$

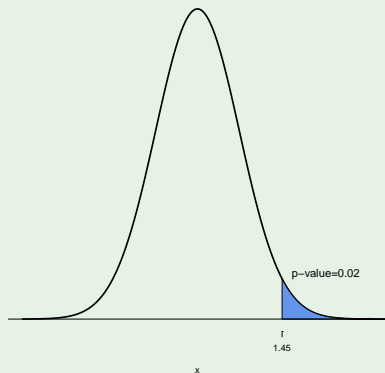
- H_0 : No significant difference Before vs After ($\mu_d = 0$)
- H_A : After scores are better ($\mu_d > 0$)
- Large \bar{d} provides evidence that H_A is true.
- $\bar{d} = 1.45$. Is this large?

How Large is Large? Expectations under Null Hypothesis

- If H_0 true, Central Limit Theorem suggests \bar{d} is approximately $N(0, \sigma_d^2/n)$
- Don't know σ_d^2 but $s_d^2 \approx \sigma_d^2$
- So H_0 true, \bar{d} is approximately $N(0, s_d^2/n)$
- Compare observed $\bar{d} = 1.45$ to $N(0, s_d^2/n)$

p-values: How Large is Large?

Distribution of \bar{d} under H_0



- Is $\bar{d} = 1.45$ abnormally large under H_0 ?
- Under H_0 expect a \bar{d} as large as this $\approx 2\%$ of the time

Types of Error

		Truth	
		H_0 True (Tree 2)	H_0 False
Decision	Reject H_0 (Tree 2)	Type I Error	
	Do not Reject		Type II Error

- α -level test: $P[\text{Type I Error}] \leq \alpha$

Reject H_0 : when $\text{p-value} < \alpha$

Issues of Concern:

p-values in phylogenetics are approximations

0.05-level test

- ▶ True $P[\text{Type I Error}] = 0.10$. High chance that we are excluding trees (Tree 2) that is reasonable. Bad
- ▶ True $P[\text{Type I Error}] < 0.05$ (Conservative test) Better. Tradeoff: $P(\text{Type II Error})$ large.

Spanish Speaking Scores Before & After Course

Subject	1	2	...	20
Before	30	28	...	29
After	29	30	...	32
d_i	-1	2	...	3

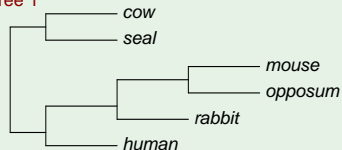
- H_0 : No difference between groups ($\mu_d = 0$)
- H_0 should be $\mu_d \leq 0$

- p-value difficulty: No longer have single distribution for \bar{d} under H_0
Under H_0 , p-value depends on choice of $\mu_d \leq 0$
- Reason for calculating p-values with $\mu_d = 0$
 - ▶ If $\bar{d} > 0$, $\mu_d = 0$ is the projection of the estimate onto H_0
 - ▶ $\mu_d = 0$ gives $P(\text{Type I error}) \leq 0.05$, all $\mu_d \leq 0$
- $\mu_d = 0$ boundary between $H_0 : \mu_d \leq 0$ and $H_A : \mu_d > 0$

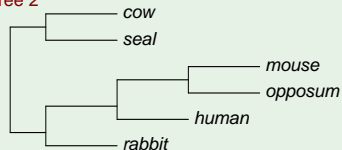
Two Tree Problem

Tree 1 vs Tree 2

Tree 1



Tree 2



- One-sided Test: Is Tree 1 significantly better than Tree 2?

KH Test

Kishino & Hasegawa (1989) J. Mol. Evol. 31:151

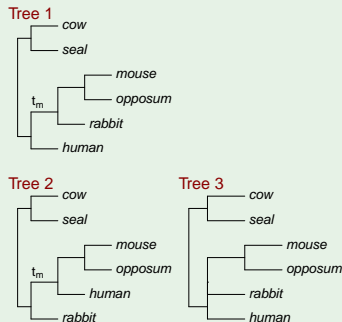
- Paired z-test of $H_0 : E[d_h] = 0$, $H_A : E[d_h] > 0$.
- d_h : difference in maximized site log-likelihoods (site h)

$$d_h = \log p(x_h; \tau_1, \hat{\mathbf{t}}_1, \hat{\boldsymbol{\theta}}_1) - \log p(x_h; \tau_2, \hat{\mathbf{t}}_2, \hat{\boldsymbol{\theta}}_2)$$

- Compare \bar{d} to $N(0, s_d^2/n)$
- Equivalently, compare $\sum_h d_h$ to $N(0, ns_d^2)$
-

$$\sum_{h=1}^n d_h = \Delta \text{ in max log likelihoods (Tree 1 - Tree 2) } = l_1 - l_2 = \Lambda_2$$

Tree 1 vs Tree 2



- Composite Null: $H_0 : T_2$ (many versions of T_2)
- $P(\text{Type I Error}) \leq \alpha$ requires

$$P_{T_2}(\text{reject } H_0) \leq \alpha$$

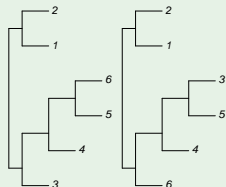
all T_2

- T_3 is on boundary between T_1 and T_2 .
- Approximate KH null: $E[d_i] \approx 0$ (large n)
 - ▶ T_3 is only version of T_2 satisfying

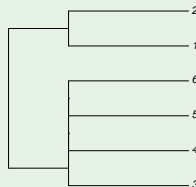
General Null Hypothesis: Consensus Tree

- Collapse as many branches as needed to make the trees equivalent.
- Don't collapse more.
- Consensus Tree of Tree 1 and Tree 2

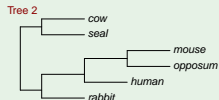
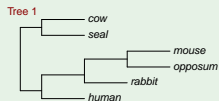
Tree 1 vs Tree 2



Null Tree



Mammal Trees



- mtREV24, 8 Gamma rate categories

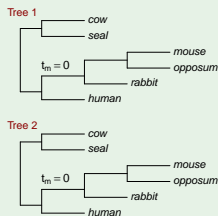
site i	l_{1i}	l_{2i}	d_i
1	-8.533	-8.556	0.023
2	-3.775	-3.776	-0.001
\vdots			
3414	-14.053	-14.158	0.105
	-21765.04	-21766.23	1.190

$$l_1 - l_2 = \sum_i d_i = 1.190 \quad \sqrt{3414} s_d = 9.012$$

One sided p-value = $P(Z > 1.190) = 0.45$

Z normal with mean 0 and standard deviation 9.012

Mammal Trees



- Simulate 5000 data sets under mtREV24 model
- $\alpha = 0.44$, 8 Gamma categories
- Tree 2, with $t_m = 0$

	α	
	0.05	0.10
Number of Rejections	0	5
Expected ($5000 \times \alpha$)	250	500

- Very Conservative Test: $P(\text{Type I Error}) \ll \alpha$

Bootstrapping (Paired z-test)

- Setting: d_1, \dots, d_n iid. True distribution (P)
- What is the distribution of \bar{d} under H_0 ? Use P
- Answer: Approximately $N(0, \sigma_d^2/n)$
- True distribution P might correspond to H_A
- Generally \bar{d} approximately $N(\mu_d, \sigma_d^2/n)$
- **Centering**: Distribution of $\bar{d} - \mu$ is $N(0, \sigma_d^2/n)$
Want to figure out distribution of $\bar{d} - \mu$ under P

Difficulties

- Don't know P
- Centering won't always give the right distribution
(eg. σ_d^2 is a function of μ_d)

- Want to figure out distribution of $\bar{d} - \mu$. Don't know P
- Empirical distribution: \hat{P} assigns mass $1/n$ to each observed d_i .

$$\hat{P}(A) \approx P(A)$$

$$E_{\hat{P}}[d] = \sum_{d_i} p(d_i) d_i = \bar{d}$$

- Suggests: If d_1^*, \dots, d_n^* are generated from \hat{P}

$$\text{distribution of } \bar{d}^* - \bar{d} \approx \text{distribution of } \bar{d} - \mu$$

- Difficulty: Probability calculations to get $\bar{d}^* - \bar{d}$ distribution is intractable

- Long run proportions match probabilities
- Select sites from $\hat{P} \iff$ Select sites with replacement
eg. $d_1^* = d_{32}, d_2^* = d_{32} \dots, d_n^* = d_3$ give a sample from \hat{P} .

$\Rightarrow \bar{d}^* - \bar{d}$ gives a realization from \hat{P}

- Repeat a large number (B) of times

Proportion of $\bar{d}^* - \bar{d} > x \approx \hat{P}(\bar{d}^* - \bar{d} > x) \approx P(\bar{d} - \mu > x)$

- For observed test statistic \bar{d} , approximate p-value

Proportion of $\bar{d}^* - \bar{d} > \text{obs}(\bar{d})$

- Nonparametric bootstrap: \hat{P} does not involve a model

RELL KH Test

Kishino et al. (1990). J. Mol. Evol. 31:151

- Test statistic: \bar{d}
- Original KH p-value = $P[D > \text{obs}(\bar{d})]$, $D \sim N(0, s_d^2/n)$.
- RELL version: $P[D > \text{obs}(\bar{d})]$, calculated using bootstrap distribution of $\bar{d}^* - \bar{d}$

$$\text{p-value} = \text{Proportion of } \bar{d}^* - \bar{d} > \text{obs}(\bar{d})$$

- Minor adjustment: \bar{d} replaced by $\text{ave}_b \bar{d}^*$
- Results are almost always identical to paired z-test version

- RELL: Resampling estimated log likelihoods:

d_1^*, \dots, d_n^* sampled with replacement from d_1, \dots, d_n

- Bootstrap principle: Mimic what is done with original data
- d_1, \dots, d_n were not the original data

- **Nonparametric** bootstrap

- ① Site columns x_1^*, \dots, x_n^* sampled with replacement from x_1, \dots, x_n
- ② Estimate $\hat{t}_j^*, \hat{\theta}_j^*, \dots$

$$\Lambda_2^* = \sum_h \log[p(x_h; \tau_1, \hat{t}_1^*, \hat{\theta}_1^*)] - \log[p(x_h; \tau_1, \hat{t}_2^*, \hat{\theta}_2^*)]$$

③

p-value = Proportion of $\Lambda_2^* - \text{ave}(\Lambda_2^*) \geq l_1 - l_2 = \Lambda_2$

- **Parametric** bootstrap:

Replace Step 1 with: Generate x_1^*, \dots, x_n^* from $p(x; \tau_j, \hat{t}_j, \hat{\theta}_j)$
(eg. seq-gen)

- Pros and Cons:

- ▶ Both much more computationally expensive than RELL
- ▶ Nonparametric robust to model misspecification (sort of)
- ▶ Parametric less variable

Parametric Bootstrap

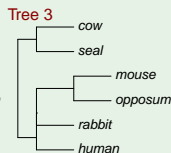
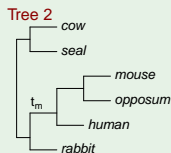
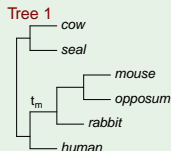
Swofford et al (1996) Molecular Systematics

- **Centering:** x_1^*, \dots, x_n^* from $\hat{P} \approx P$

We use $l_1^* - l_2^* - \text{ave}(l_1^* - l_2^*)$ instead of $l_1^* - l_2^*$

- H_A might be true \implies mean $l_1^* - l_2^*$ not ≈ 0 as under H_0
- Difficulty: If H_A true, other parts of distribution might differ (eg Standard Deviation)

Tree 1 vs Tree 2

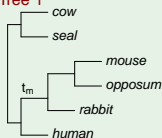


- Parametric bootstrap: Can generate from H_0 : Tree 3
- Centering not needed
- KHns approximates parametric bootstrap Tree 3 using simple normal simulation

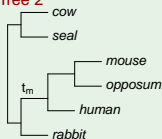
Comparison of P-values Mammal Data

Tree 1 vs Tree 2

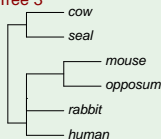
Tree 1



Tree 2



Tree 3

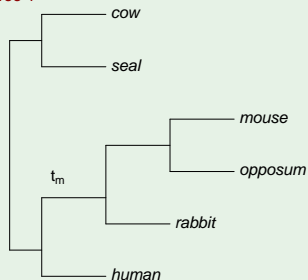


Test	p-value
KH	0.45
Nonpar (center)	0.45
Par, Tree 3 (uncenter)	0.05
Par, Tree 3 (center)	0.05
KHns	0.05

Two Tree Test as One Tree Test

Tree 1 vs Tree 2

Tree 1

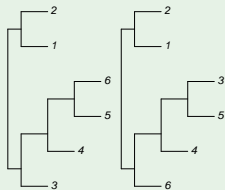


- Null hypothesis Tree 2 is a special case of Tree 1
⇒ Conventional parametric model test
- $H_0 : t_m = 0$ vs $H_A : t_m > 0$

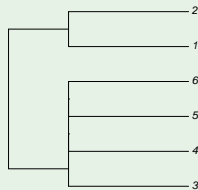
General Null Hypothesis: Consensus Tree

- General Case
- Tree 2 is special case of Tree 1
- But $t_m = 0$ where t_m is p-dimensional

Tree 1 vs Tree 2



Null Tree



- Null Hypothesis $H_0 : \mathbf{t}_m = 0$, $H_A : \mathbf{t}_m > 0$
 \mathbf{t}_m p-dimensional. Edges set to 0 in consensus tree
- LR Statistic

$$2\Lambda_3 = 2\{l_1 - l_3\}$$

l_3 maximized log likelihood for T_1 holding $\mathbf{t}_m = 0$ fixed
Equivalently, maximized log likelihood for T_3

- Standard statistical theory indicates p-values should be calculated as

$$P(\chi_p^2 > 2\{l_1 - l_3\})$$

Chi-square test

- Standard theory condition: \mathbf{t}_m be in the interior of the parameter space
- Null Hypothesis $H_0 : \mathbf{t}_m = 0$, $H_A : \mathbf{t}_m > 0$
- \mathbf{t}_m is on the boundary of the parameter space
- Shapiro (1985)

$$P(2\{l_1 - l_3\} > y) = \sum_{j=0}^p w_j P(\chi_j^2 > y)$$

Chi-bar test

- Consensus Tree with p zero-length edges. Shapiro (1985) \Rightarrow

$$\text{p-value} = \sum_{j=0}^p w_j P(\chi_j^2 > 2\Lambda_3)$$

- $p = 1$, Ota et al. (2000) Mol Biol Evol 17:793: $w_1 = w_2 = 1/2$
- $p = 2, 3$: w_j can be approximated using second derivatives of log likelihood.
- $p \geq 4$: no expression for w_j .

Naive (Chi-square) Test

- Ignore boundary issue: Under H_A , p additional parameters.

$$\text{naive p-value} = P(\chi_p^2 > 2\Lambda_3)$$

$$P(\chi_p^2 > 2\Lambda_3) \leq \sum_{j=0}^p w_j P(\chi_j^2 > 2\Lambda_3) = \text{correct p-value} \quad (1)$$

$$P(\text{Type 1 Error}) \leq \alpha$$

- Λ_2 available from any software.
- Since consensus tree T_3 is special case of Tree 2, $l_2 > l_3$. So

$$\Lambda_2 = l_1 - l_2 < l_1 - l_3 = \Lambda_3$$

- So using Λ_2 with Λ_3 -based thresholds gives a conservative test.

Naive (Chi-square) Test using KH Test Statistic

<https://www.mathstat.dal.ca/~tsusko/software.html>

Susko (2014)

```
$ cat mammal-2trees
(human, (seal, cow), (rabbit, (opposum, mouse)));
(human, (rabbit, (seal, cow)), (opposum, mouse));
$ iqtree -s mtprot.phy -z mammal-2trees \
    -m mtREV+F+G8 -n 0
$ cat mtprot.phy.trees
[ tree 1 lh=-21765.1 ] (human:0.2731, ...
[ tree 2 lh=-21766.2341157180 ] (human:0.26588, ...
$ trees2df -n 6 < mtprot.phy.trees
1 1
...
```

In R

```
> 2*(-21765.1-(-21766.2341157180))
[1] 2.268231
> 1-pchisq(2.268231,1)
[1] 0.1320506
```

p-value comparison

Test	p-value
KH	0.45
Nonpar (center)	0.45
Par, Tree 3 (center)	0.05
Par, Tree 3 (uncenter)	0.05
KHns	0.05
Naive (KH)	0.13
Chi-bar (KH)	0.07
Naive (Λ_3)	0.00
Chi-bar (Λ_3)	0.00

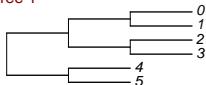
Table of Tests

Test	Comments
KH	Highly Conservative Type I
KHns	Approximate Type I
chi-bar	Approximate Type I
chi-bar(KH)	Conservative Type I NA $p \geq 4$
naive	Conservative Simple Calculation
naive(KH)	Simplest calculation

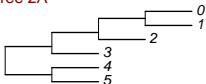
Trees in Simulations

- 1000 simulated data sets, 1000 sites each.
- HKY, $\kappa = 2$, frequency of A, C, G and T 0.1, 0.2, 0.3 and 0.4.

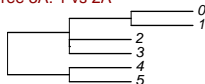
Tree 1



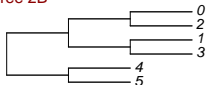
Tree 2A



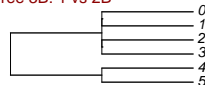
Tree 3A: 1 vs 2A



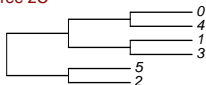
Tree 2B



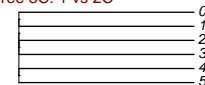
Tree 3B: 1 vs 2B



Tree 2C



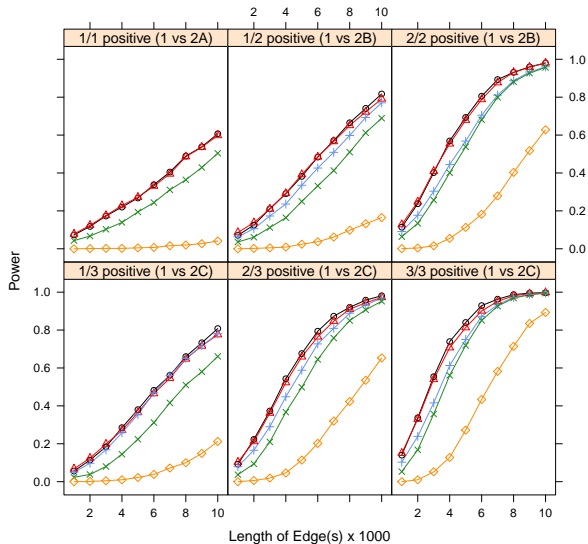
Tree 3C: 1 vs 2C



Null Simulations - Number of False Positives

Tree	KHns	chi-bar		naive		KH
		Λ_3	KH	Λ_3	KH	
3A	48	44	40	24	23	0
3B	50	39	31	18	15	0
3C	40	31	23	10	6	0

○ chi-bar + conditional ◇ KH
 △ KHns × naïve

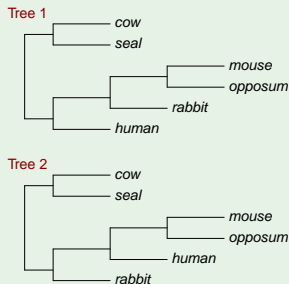


Summary (Two Tree Tests)

- KHns & chi-bar best performers
 - ▶ Implementation is complicated
- KH simple but very conservative
- Parametric Bootstrapping should be applied with consensus tree

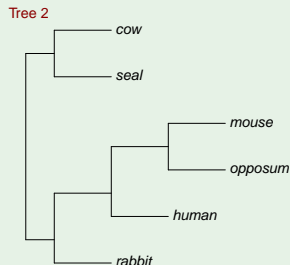
Two Main Topology Test Problems

Two trees



- Is Tree 1 significantly better than Tree 2

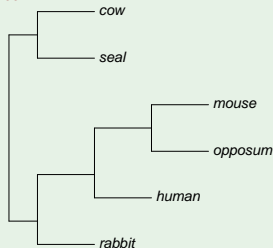
One Tree



- Is Tree 2 plausible?

One Tree

Tree 2

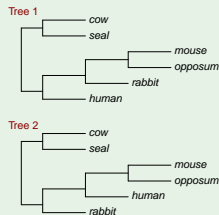


- Given a test, 95% confidence set of trees C : All trees giving $p \geq 0.05$

$$P[\text{True Tree in } C] \geq 0.95$$

- Coverage: proportion of times true tree is in confidence set
- If a test has $P(\text{Type I error}) = 0.02$ then the coverage of the confidence set is 0.98

Mammal Trees



- T_1 and T_2 fixed a priori:

Q_{KH} : T_1 significantly better than T_2 ?

- If instead, only T_2 is fixed a priori,

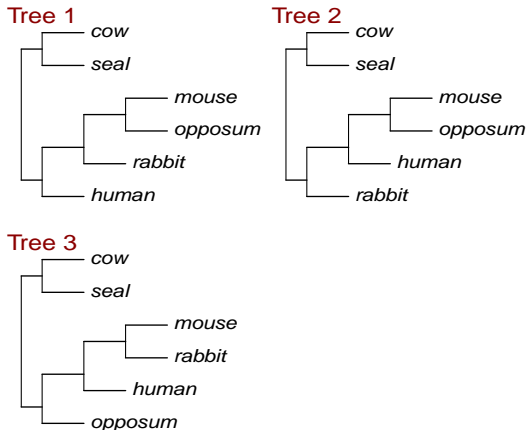
Q_{SH} : ML tree significantly better than T_2 ?

Selection Bias: Tree 1 is selected based on the data

- If Tree 1 fixed and H_0 true $l_1 - l_2 < 0$ approximately 50% of time
- $l_1 - l_2$ never < 0 if Tree 1 is ML tree

- **Setting:** T_1 and T_2 become $T_1, \dots, T_M \Rightarrow l_1, \dots, l_M$
 - ▶ Mammal data. 6 taxa $\Rightarrow M = 105$ trees.
- Test statistic $l_1 - l_2$ replaced by $l_m - l_1$
 m indice of MLE.
- **RELL Bootstrapping**
 - ▶ Replace l_1^*, \dots, l_M^* by
 $l_1^* - \text{ave}_b l_1^*, \dots, l_M^* - \text{ave}_b l_M^*$
 - ▶ Use observed $l_{m^*}^* - l_2^*$ from bootstrapping for null distribution.
 m^* : indice of MLE for bootstrap sample.

Mammal Data Example - Three trees



- Tree 1 was the ML tree for this data

Mammal Data Example - Three trees

- $B = 5000, M = 3$
- $l_1 - l_2 = 1.19$
- l_j^* (after centering), first three bootstrap samples

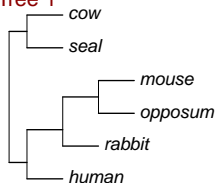
l_1^*	l_2^*	l_3^*	m^*	$l_1^* - l_2^*$	$l_{m^*}^* - l_2^*$
-359.78	-360.62	-352.52	3	0.84	8.10
-84.45	-94.44	-95.87	1	9.99	9.99
-65.93	-58.62	-62.19	2	-7.31	0.00

pKH = proportion of $l_1^* - l_2^* > 1.19 = 0.44$

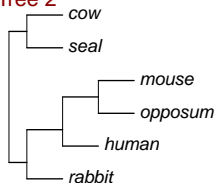
pSH = proportion of $l_{m^*}^* - l_2^* > 1.19 = 0.59$

Mammal Data Example - Four trees

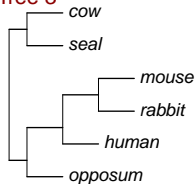
Tree 1



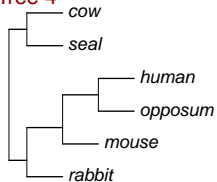
Tree 2



Tree 3



Tree 4



Mammal Data Example - Four trees

- $B = 5000, M = 4$
- $l_1 - l_2 = 1.19$

$$pKH = \text{proportion of } l_1^* - l_2^* > 1.19 = 0.44$$

$$pSH_3 = \text{proportion of } l_{m_3}^* - l_2^* > 1.19 = 0.59$$

$$pSH_4 = \text{proportion of } l_{m_4}^* - l_2^* > 1.19 = 0.74$$

$$pSH_{105} = \text{proportion of } l_{m_{105}}^* - l_2^* > 1.19 = 0.95$$

- Bootstrap Principle: Bootstrapping should mimic what is being done with original data.
- If exhaustive search for ML tree, $M = 105$
- Impossible with large number of taxa
 - ▶ Best 100 or 1000 trees found in tree searching
 - ▶ Collection of bootstrap trees
 - ▶ ...

Selection bias corrected parametric bootstrap

- Full parametric bootstrapping from current tree (Tree 2) considered for confidence set
- Calculate l_m^* ML Tree for each bootstrap sample
- $p_{SOWH} = \text{Proportion of } l_m^* - l_2^* > l_m - l_2$
- $p_{SOWH} \ll p_{SH}$
- Sometimes, SOWH will generate from a fully-resolved Tree 2
Under fully-resolved Tree 2, less likely to see large $l_m - l_2$

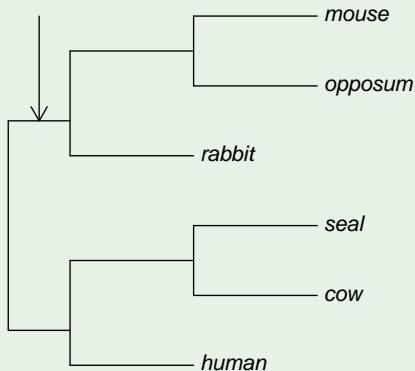
p-value comparison

Test	p-value
KH	0.45
Par, Tree 3	0.05
Chi-bar (KH)	0.07
SOWH	0.00

- SOWH gets smaller p-value even though it adjusts for selection bias
- Should apply SOWH with Tree 3 as simulating tree

Splits

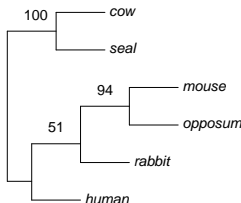
Split of Interest



- Significant evidence that the split is present?

Bootstrap Support

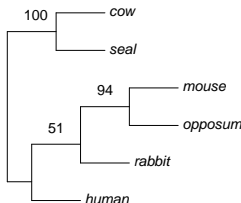
- For each bootstrap sample x_1^*, \dots, x_n^* obtain \hat{T}^*
- BP for opossum, mouse and rabbit = proportion of T^* with that split.



- By far the most frequent measure of uncertainty
- How large of BP is large?

- Felsenstein (1985): Bootstrap Support (BP) introduced
- Hillis and Bull (1993): BP is probability split is correct. 70% is large.
- Felsenstein and Kishino (1993): 1-BP is p-value for hypothesis that split is not present. 95% is large.
- Efron, Halloran and Holmes (1996) 1-BP is first order correct.
 - ▶ Efron and Tibshirani (1998) proof for analogous problem of regions
- Susko (2009) E & T result correct for problem of regions but not phylogenetics
 - ▶ Fixed Tree: 1-BP is conservative: Expect 95% BP less than 5% of time

Bootstrap Support for Splits



- Fixed Tree: BP is conservative:
 - ▶ Expect $>95\%$ BP less than 5% of time
 - ▶ Expect $< 40\%$ BP more than 40% of time
- Selection Bias: ML tree is not fixed a priori. Very unlikely $BP < 40\%$

- For each bootstrap sample x_1^*, \dots, x_n^* obtain \hat{T}^*
- BP for tree T is proportion of times $\hat{T}^* = T$
- BP for Tree always \leq BP for each of its splits
 \implies conservative. Expect 95% BP less than 5% of time
- AU Test (Shimoaladiara (2002) Syst Biol 5:492). Bootstrap correction
 - ▶ Bootstraps with differing fractions of original sample size
 $\implies BP(r_1) \dots, BP(r_m)$.
 - ▶ AU p-value is transformation of $BP(r_1) \dots, BP(r_m)$
 - ▶ The correction relies on 1-BP being first-order correct
 - ★ Works for the analogous problem of regions
 - ★ Not justified in phylogenetics

IQ-TREE

```
$iqtree -s mtprot.phy -z mammal-trees -m mtREV+F+G8 \  
-n 0 -zb 1000 -au
```

```
$cat mtprot.phy.iqtree
```

Tree	bp-RELL	p-KH	p-SH	c-ELW	p-AU
1	0 -	0 -	0 -	9.64e-38 -	0.00118 -
2	0 -	0 -	0 -	4.17e-33 -	1.02e-74 -
..					
91	0.422 +	0.442 +	0.945 +	0.421 +	0.571 +

95% Confidence Sets of Trees

AU Test: 6 Trees

(human,(seal,cow),(rabbit,(opposum,mouse)));
(opposum,(human,(rabbit,(seal,cow))),mouse);
(opposum,((rabbit,human),(seal,cow)),mouse);
((opposum,human),(rabbit,(seal,cow)),mouse);
(opposum,human,((rabbit,mouse),(seal,cow)));
(seal,cow,(opposum,(human,(rabbit,mouse))));

KH Test: Same as AU Test

SH Test: 15 Trees: All Trees with (cow,seal)

Naive (Chi-square) Test: 2 Trees

(human,(seal,cow),(rabbit,(opposum,mouse)));
(opposum,(human,(rabbit,(seal,cow))),mouse);

- SH is fast but extremely conservative. Values depends on number of input trees M
- BP is conservative
- AU is not justified. Properties unclear.
- SOWH more accurate but intensive and should use consensus tree (Tree 3)
- KH test is very conservative as two tree test.
- Chi-square (Naive) reasonable but does not adjust for selection bias
- Working on alternatives