Bootstrap & Topology Tests

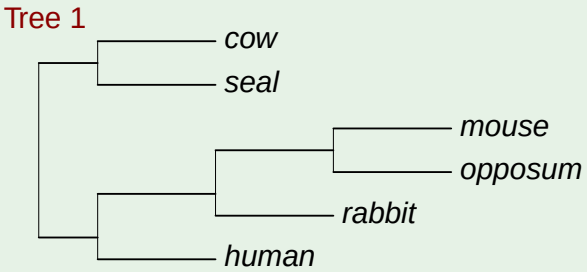
Edward Susko

Department of Mathematics and Statistics, Dalhousie University

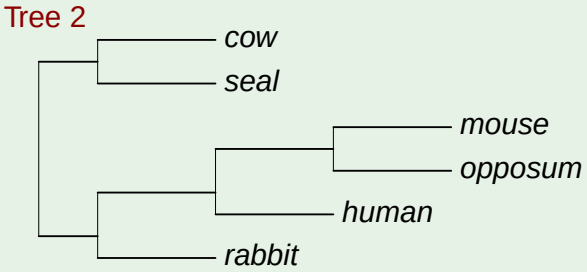
Two Main Topology Test Problems

Two trees

Tree 1



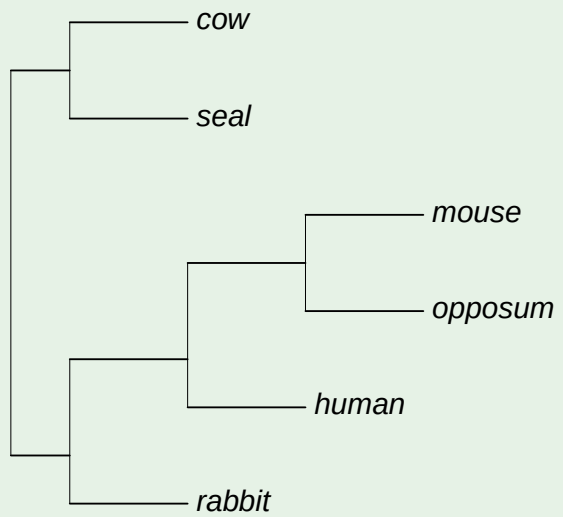
Tree 2



- Is Tree 1 significantly better than Tree 2

One Tree

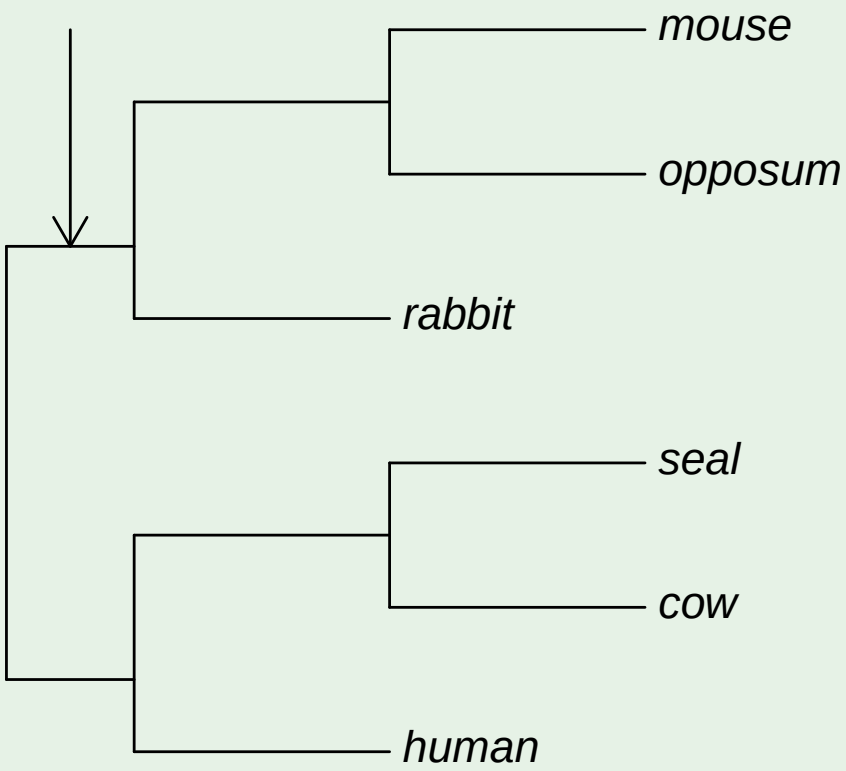
Tree 2



- Is Tree 2 plausible?

Splits

Split of Interest



- Significant evidence that the split is present?

Ingredients Needed

- Likelihood-Based Methods: Ingredient - A likelihood

$$L(T) = P(Data|T)$$

  - Note: T is usually the species tree. Might have

$$L(T) = \sum_{\text{Gene Trees } \tau} P(Data|\tau)P(\tau|T)$$

  - Can be used to test for gene tree/species tree incongruence

Tree 2 = Species Tree, Tree 1 = Tree for Gene j

- Bootstrap Methods: Ingredient - A method for estimation (sort of)

Concatenated Protein Set

	Site				
	1	2	3	...	n
Homo	S	E	S	...	-
Enceph	Y	E	K	...	S
Schizo	I	E	N	...	S
Saccha	I	D	N	...	S
...					

- Each protein has $n \approx 300$. 133 proteins
- Concatenated sets large $n = 24291$
- **Observational unit ($\boldsymbol{x}_h$)**: vector of data at a site

- Usually data at sites are treated as independent. Likelihood

$$\text{Likelihood} = L(\tau, \boldsymbol{t}, \theta) = \prod_h p(\boldsymbol{x}_h; \tau, \boldsymbol{t}, \theta)$$

τ - topology
$\boldsymbol{t}$ - edge lengths
θ - other parameters

Spanish Scores Before/After Course

Subject	1	2	$\cdots$	20
Before	30	28	$\cdots$	29
After	29	30	$\cdots$	32
d_i	-1	2	$\cdots$	3

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

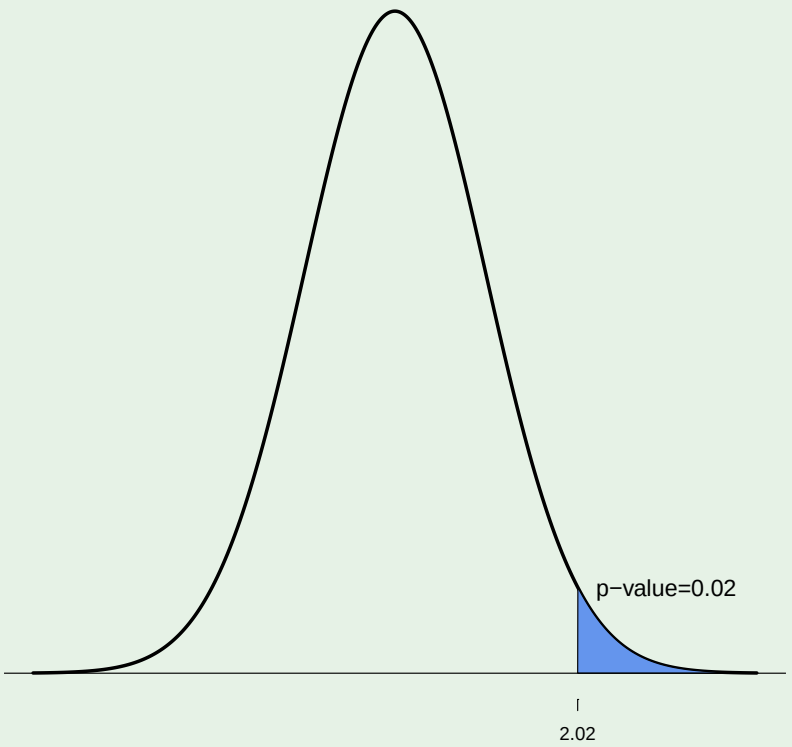
- H_0 : No significant difference Before vs After ($\mu_d = 0$)
- H_A : After scores are better ($\mu_d > 0$)
- Large $\bar{d}$ provides evidence that H_A is true.
- $\bar{d} = 1.45$. Is this large?

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
- How small is small?
  - $p < 0.01$ Strong evidence for H_A
  - $p \geq 0.1$ No evidence for H_A
  - $0.01 \leq p < 0.05$ marginal evidence for H_A.

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[Type I Error] $\leq \alpha$

  Reject H_0 : when p-value$< \alpha$

- Best Test P[Type I Error] $\leq \alpha$ and

$$\begin{aligned} \text{Maximal Power} &= 1 - P(\text{Type II error}) \\ &= P(\text{Correctly Reject } H_0) \end{aligned}$$

Phylogenetics Concerns

- Ideal Case: P[Type I Error] $= \alpha$, all n (sequence length)
- Theoretical Results:
  - P[Type I Error] $\approx \alpha$, n large
  - P[Type I Error] $< \alpha$
- True P[Type I Error] $= 0.10$. Bad
- True P[Type I Error] < 0.05 (Conservative test) Better. Tradeoff: Lower Power than if P[Type I Error] $= 0.05$
  - Large phylogenomic data sets. Power expected to be large

Composite Null Hypothesis

Spanish Speaking Scores Before & After Course

Subject	1	2	$\cdots$	20
Before	30	28	$\cdots$	29
After	29	30	$\cdots$	32
d_i	-1	2	$\cdots$	3

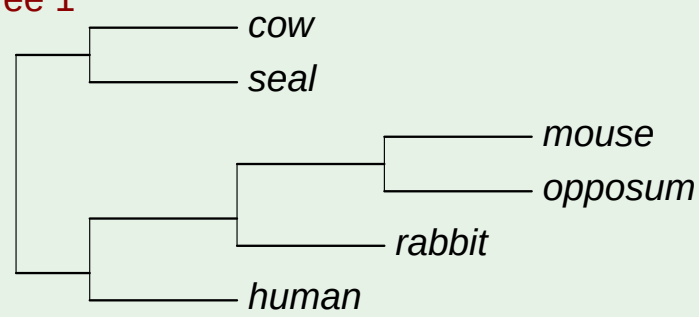
- H_0 : No difference between groups ($\mu_d = 0$)
- H_0 should be $\mu_d \leq 0$

- 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 $\mu_d = 0$
  - $\mu_d = 0$ gives $P(\text{Type I error}) \leq 0.05, \quad \text{all } \mu_d \leq 0$
- $\mu_d = 0$ boundary between $H_0 : \mu_d \leq 0$ and $H_A : \mu_d > 0$

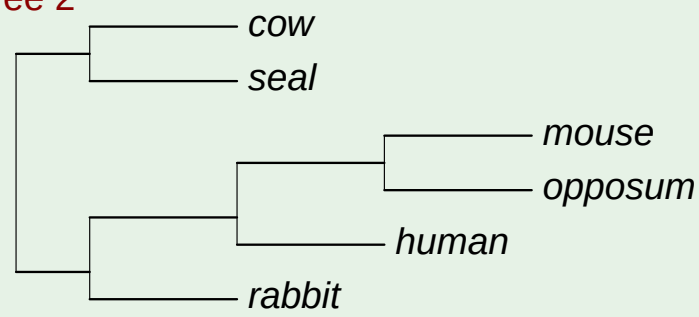
Two Tree Problem

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

- d_h: difference in maximized site log-likelihoods (site h)

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

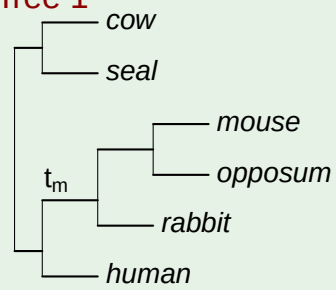
- Paired z-test of $H_0 : E[d_h] = 0$, $H_A : E[d_h] > 0$.
- Compare $\bar{d}$ to $N(0, s_d^2/n)$
- Equivalently, compare $\sum_h d_h$ to $N(0, ns_d^2)$
-

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

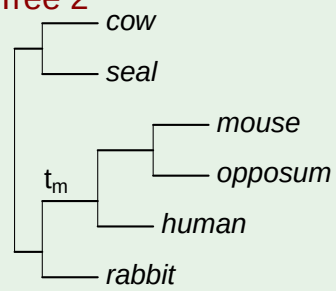
Null Hypothesis: Tree 2 correct

Tree 1 vs Tree 2

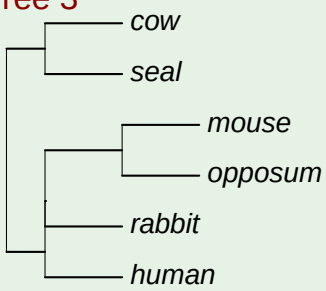
Tree 1



Tree 2



Tree 3



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

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

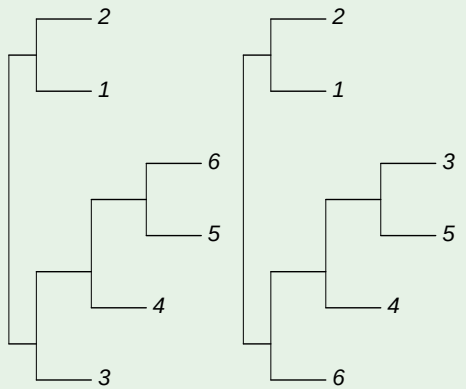
all T_2

- 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

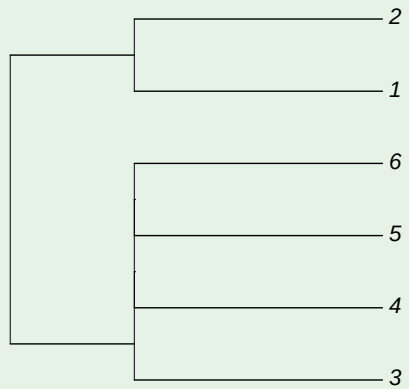
General Null Hypothesis: Consensus Tree

- 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

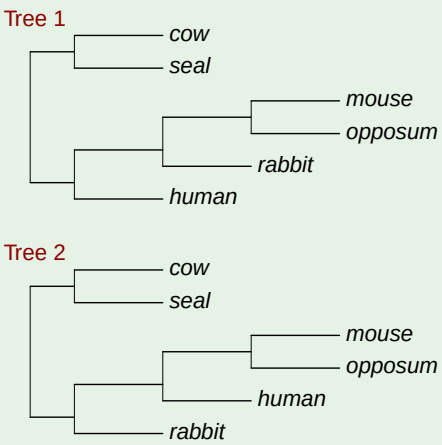


Null Tree



KH Example - Mammalian Mitochondrial Data

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
$\vdots$		$\vdots$	
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

KH Test Motivation in More Detail

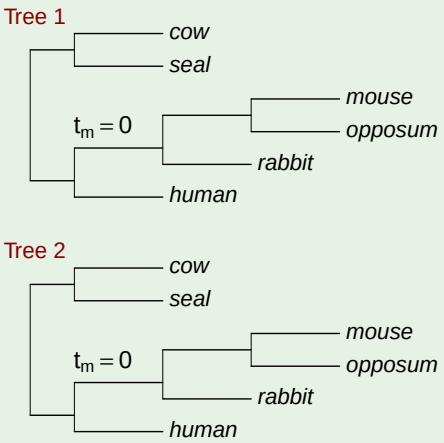
- If $\log p(x_h; \tau_j, \hat{\boldsymbol{t}}_j)$ are i.i.d. then
  Central Limit Theorem $\Longrightarrow \bar{d}$ approximately normal
- Usual models: sites evolve independently
- But sites $1, \ldots, n$ all contribute to $\hat{\boldsymbol{t}}_j$
- So $\log p(x_h; \tau_j, \hat{\boldsymbol{t}}_j)$ are not independent
  whereas $\log p(x_h; \tau_j, \boldsymbol{t}_j)$ are independent
- Argument by approximation: $\hat{\boldsymbol{t}}_j \approx \boldsymbol{t}_j$,

$$\sum_{h=1}^{n} \log p(x_h; \tau_j, \hat{\boldsymbol{t}}_j) \approx \sum_{h=1}^{n} \log p(x_h; \tau_j, \boldsymbol{t}_j) + r_{jn}(\boldsymbol{t}_j)$$

$r_{1n}(\boldsymbol{t})$ is relatively small.

Properties under Null Hypothesis

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(Type I Error) $<< \alpha$

KH motivation difficulty

$$\sum_{h=1}^{n} \log p(x_h; \tau, \hat{\boldsymbol{t}}_1) \approx \sum_{h=1}^{n} \log p(x_h; \tau_1, \boldsymbol{t}_1) + r_{jn}(\boldsymbol{t}_1)$$

and

$$\sum_{h=1}^{n} \log p(x_h; \tau, \hat{\boldsymbol{t}}_2) \approx \sum_{h=1}^{n} \log p(x_h; \tau_2, \boldsymbol{t}_2) + r_{jn}(\boldsymbol{t}_2)$$

but under H_0 : true tree is $T_3 = (\tau_1, \boldsymbol{t}_1) = (\tau_2, \boldsymbol{t}_2)$
So first terms cancel

$$\Lambda_2 \approx r_{1n}(\boldsymbol{t}_1) - r_{2n}(\boldsymbol{t}_2)$$

Bootstrapping (Paired z-test)

- Setting: $d_1, \ldots, d_n$ iid.
- If $H_0 : \mu_d = 0$ true, $\bar{d}$ approximately $N(0, \sigma_d^2/n)$
- Can we approximate $P_{H_0}(\bar{d} > x)$ without probability calculations?

Calculation using P_t, true distribution

- Whether H_0 true or not, $\bar{d}$ approximately $N(\mu_d, \sigma_d^2/n)$
- **Centering**: Distribution of $\bar{d} - \mu_d$ is $N(0, \sigma_d^2/n)$

$$\mu_d = \sum_d d \times P_t(D = d) =: \mu(P_t)$$

So $P_t(\bar{d} - \mu(P_t) > x) \approx P_{H_0}(\bar{d} > x)$

Difficulty: Don't know P_t.

Bootstrapping - Empirical Distribution

$$P_t(\bar{d} - \mu(P_t) > x) \approx P_{H_0}(\bar{d} > x)$$

- Empirical distribution: Assign mass $1/n$ to each d_i. $\hat{P}(D = d_i) = 1/n$.
- Empirical distribution gives probabilities as proportions. eg

$$\hat{P}(D > 5) = \sum_{d_i > 5} \hat{P}(D = d_i) = \sum_{d_i > 5} (1/n) = \text{Proportion of } d_i > 5$$

  Generally, $\hat{P}(A) =$ Proportion of d_i satisfying A.
- Law of Large Numbers: Proportions approximate true probabilities. So

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

Bootstrapping - Approximation with Empirical Distribution

$$P_t(\bar{d} - \mu(P_t) > x) \approx P_{H_0}(\bar{d} > x)$$

- Since $\hat{P} \approx P_t$,

$$\hat{P}(\bar{d}^* - \mu(\hat{P}) > x) \approx P_t(\bar{d} - \mu(P_t) > x) \approx P_{H_0}(\bar{d} > x)$$

- **Centering**:

$$\mu(\hat{P}) = \sum_{d_i} d_i \hat{P}(D = d_i) = \sum_{i=1}^{n} d_i/n = \bar{d}$$

So

$$\hat{P}(\bar{d}^* - \bar{d} > x) \approx P_{H_0}(\bar{d} > x)$$

Difficulty: Need to do probability calculations with $\hat{P}$

Bootstrapping in Practice

$$\hat{P}(\bar{d}^* - \bar{d} > x) \approx P_{H_0}(\bar{d} > x)$$

- Law of Large Numbers: Proportions approximate true probabilities
- Repeatedly generate data from $\hat{P}$

$$\begin{aligned} d_1^{(1)}, \ldots, d_n^{(1)} \sim \hat{P} &\Rightarrow \bar{d}^{(1)} \\ d_1^{(2)}, \ldots, d_n^{(2)} \sim \hat{P} &\Rightarrow \bar{d}^{(2)} \\ &\vdots \\ d_1^{(B)}, \ldots, d_n^{(B)} \sim \hat{P} &\Rightarrow \bar{d}^{(B)} \end{aligned}$$

Proportion of $\bar{d}^{(b)} - \bar{d} > x \approx \hat{P}(\bar{d}^* - \bar{d} > x) \approx P_{H_0}(\bar{d} > x)$.

$$\text{Proportion of } \bar{d}^{(b)} - \bar{d} > x \approx \hat{P}(\bar{d}^* - \bar{d} > x) \approx P_{H_0}(\bar{d} > x).$$

- Approximate p-value:

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

- Note: Generating $d_1^*, \ldots, d_n^*$ from $\hat{P}$ equivalent to sampling from $d_1, \ldots, d_n$ with replacement.
- 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 and Bootstrap Principle

- RELL: Resampling estimated log likelihoods:

  $d_1^*, \ldots, d_n^*$ sampled with replacement from $d_1, \ldots, d_n$
- Bootstrap principle: Mimic what is done with original data
- $d_1, \ldots, d_n$ were not the original data

Nonparametric and Parametric Bootstrap

- **Nonparametric** bootstrap
  1. Site columns $x_1^*, \ldots, x_n^*$ sampled with replacement from $x_1, \ldots, x_n$
  2. Estimate $\hat{\boldsymbol{t}}_j^*, \hat{\boldsymbol{\theta}}_j^*, \ldots$

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

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

- **Parametric** bootstrap:
  Replace Step 1 with: Generate $x_1^*, \ldots, x_n^*$ from $p(x; \tau_2, \hat{t}_2, \hat{\theta}_2)$
  (eg. seq-gen)
- Pros and Cons:
  - Better Type I error rate
  - Nonparametric robust to model misspecification
  - Parametric less variable
  - Both much more computationally expensive than RELL

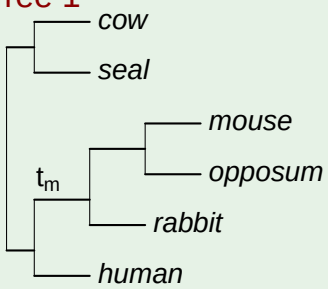
Parametric Bootstrap

Swofford et al (1996) Molecular Systematics

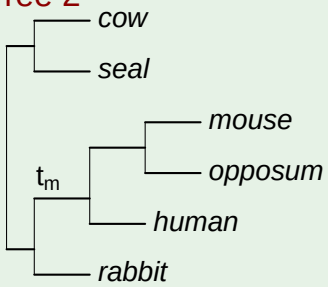
- **Nonparametric**: $x_1^*, \ldots, x_n^*$ from $\hat{P} \approx P$
  We use $l_1^* - l_2^* - \text{ave}(l_1^* - l_2^*)$ instead of $l_1^* - l_2^*$
- H_A might be true $\Longrightarrow$ mean $l_1^* - l_2^*$ not ≈ 0 as under H_0

Tree 1 vs Tree 2

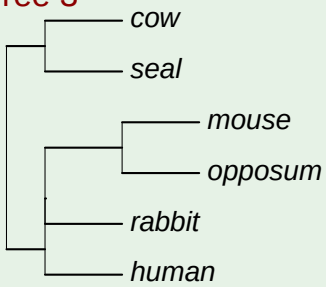
Tree 1



Tree 2



Tree 3

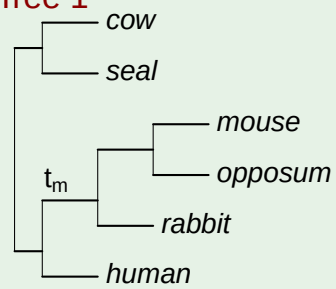


- Parametric bootstrap: Can generate from H_0 :Tree 3
- Centering not needed
- KHns approximates parametric bootstrap Tree 3 using simple normal simulation

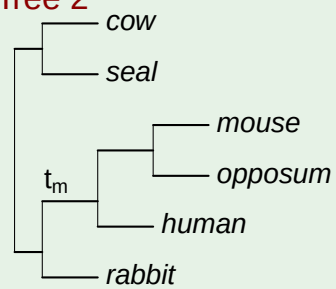
Comparison of P-values Mammal Data

Tree 1 vs Tree 2

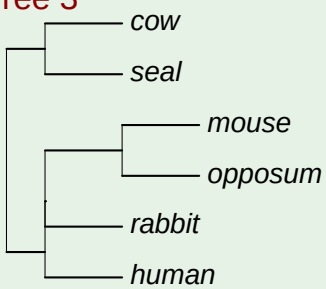
Tree 1



Tree 2



Tree 3

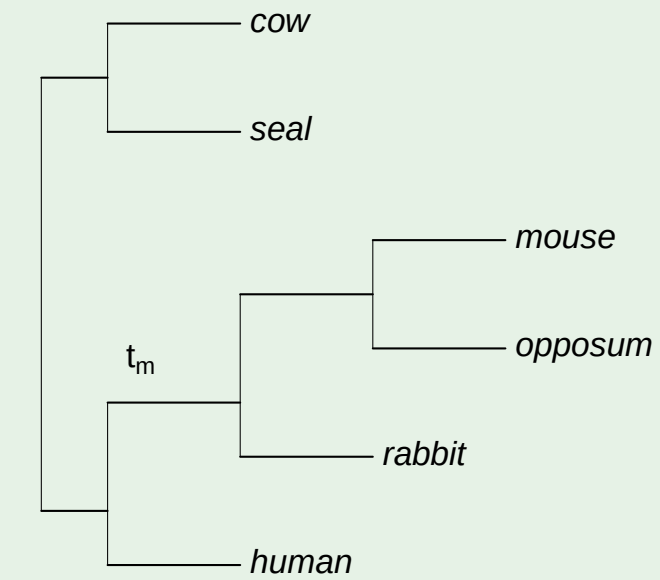


Test	p-value
KH	0.45
Nonpar (center)	0.45
Par, Tree 3 (uncenter)	0.05
Par, Tree 3 (center)	0.05
KHns	0.05

Two Tree Test as One Tree Test

Tree 1 vs Tree 2

Tree 1

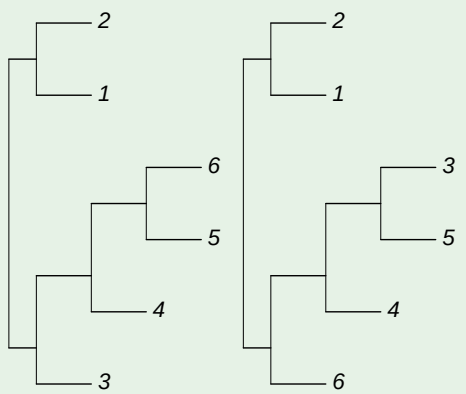


- Null hypothesis Tree 2 is a special case of Tree 1
  $\Rightarrow$ Conventional parametric model test
- $H_0 : t_m = 0$ vs $H_A : t_m > 0$

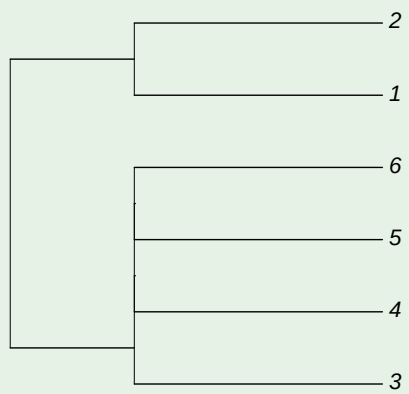
General Null Hypothesis: Consensus Tree

- General Case
- Tree 2 is special case of Tree 1
- But $\boldsymbol{t}_m = 0$ where $\boldsymbol{t}_m$ is p-dimensional

Tree 1 vs Tree 2



Null Tree



Likelihood Ratio Test

- Null Hypothesis $H_0 : \boldsymbol{t}_m = 0, \quad H_A : \boldsymbol{t}_m > 0$
  $\boldsymbol{t}_m$ 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 $\boldsymbol{t}_m = 0$ fixed
  Equivalently, maximized log likelihood for T_3
- Standard statistical theory indicates p-values should be calculated as

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

  Chi-square test

- Standard theory condition: $\boldsymbol{t}_m$ be in the interior of the parameter space
- Null Hypothesis $H_0 : \boldsymbol{t}_m = 0, \quad H_A : \boldsymbol{t}_m > 0$
- $\boldsymbol{t}_m$ 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

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.

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

$$P(\chi^2_p > 2\Lambda_3) \leq \sum_{j=0}^{p} w_j P(\chi^2_j > 2\Lambda_3) = \text{correct p-value} \qquad (1)$$

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

Using KH test statistic

- 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.

Naiive (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

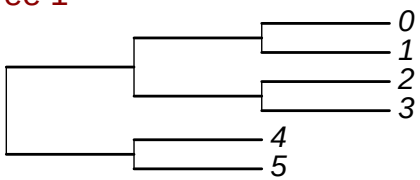
Table of Tests

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

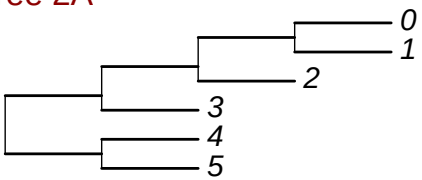
Trees in Simulations

- 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.

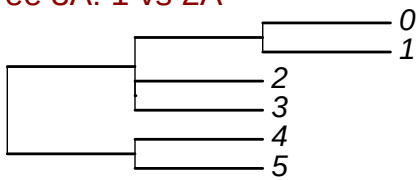
Tree 1



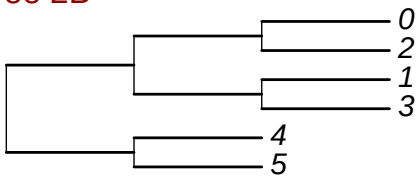
Tree 2A



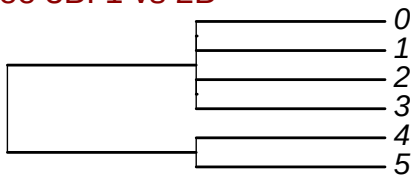
Tree 3A: 1 vs 2A



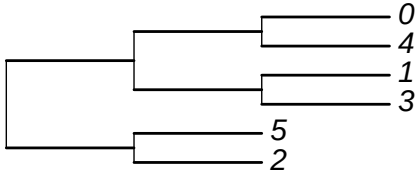
Tree 2B



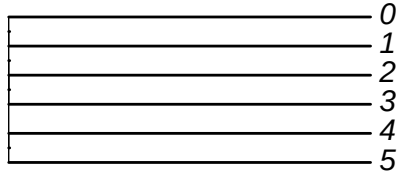
Tree 3B: 1 vs 2B



Tree 2C

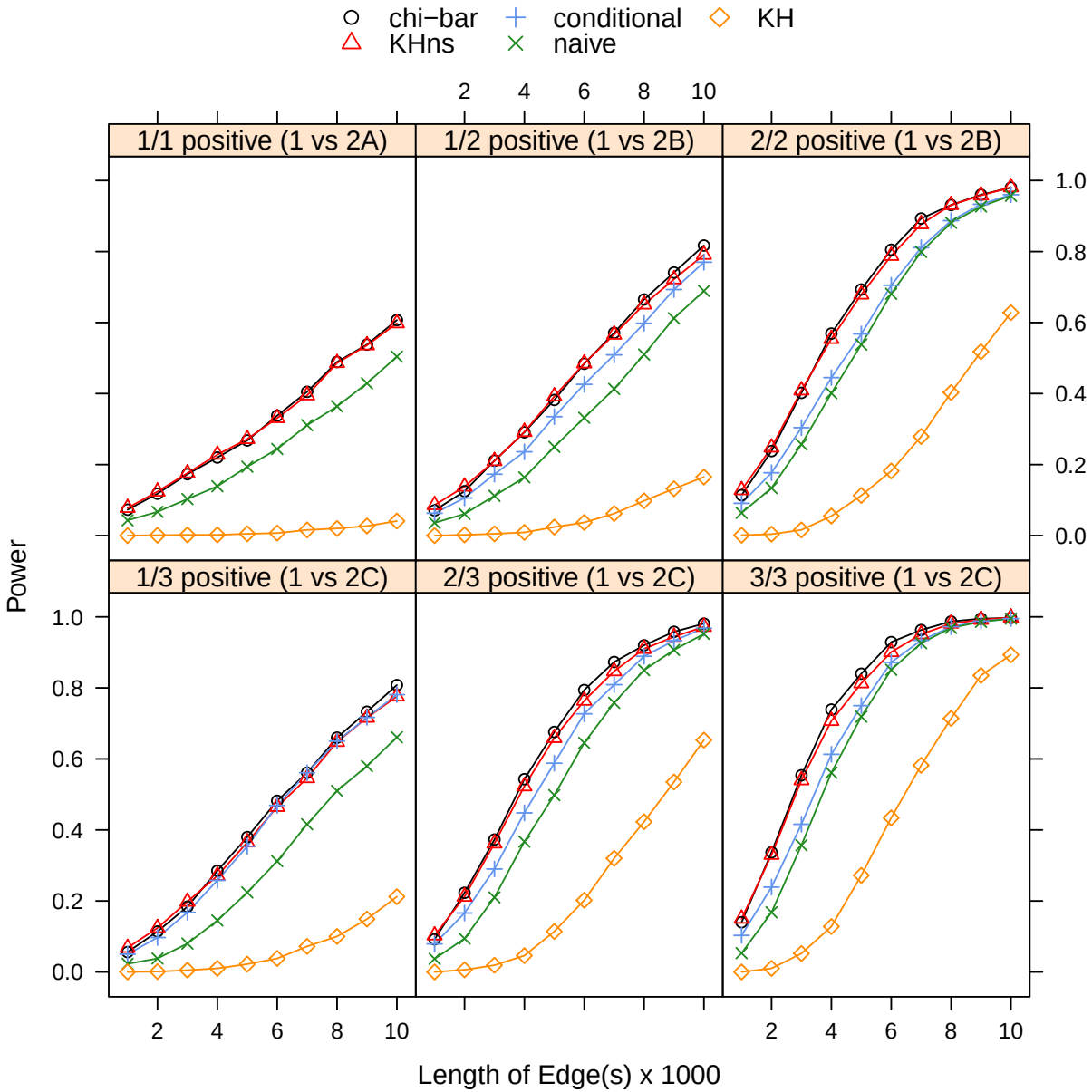


Tree 3C: 1 vs 2C



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
KHns
conditional
naive
KH
1/1 positive (1 vs 2A)
1/2 positive (1 vs 2B)
2/2 positive (1 vs 2B)
1/3 positive (1 vs 2C)
2/3 positive (1 vs 2C)
3/3 positive (1 vs 2C)
Power
Length of Edge(s) x 1000

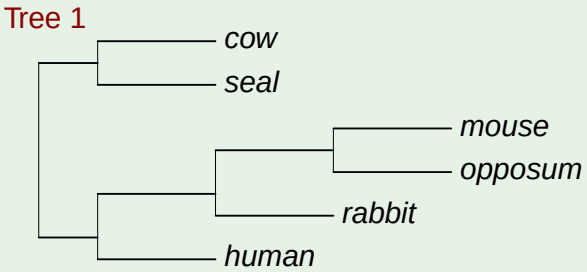
Summary (Two Tree Tests)

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

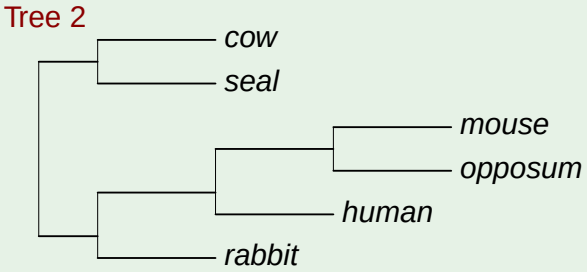
Two Main Topology Test Problems

Two trees

Tree 1



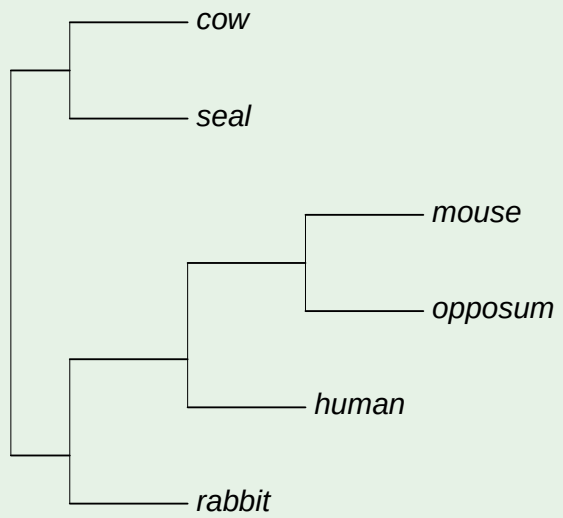
Tree 2



- Is Tree 1 significantly better than Tree 2

One Tree

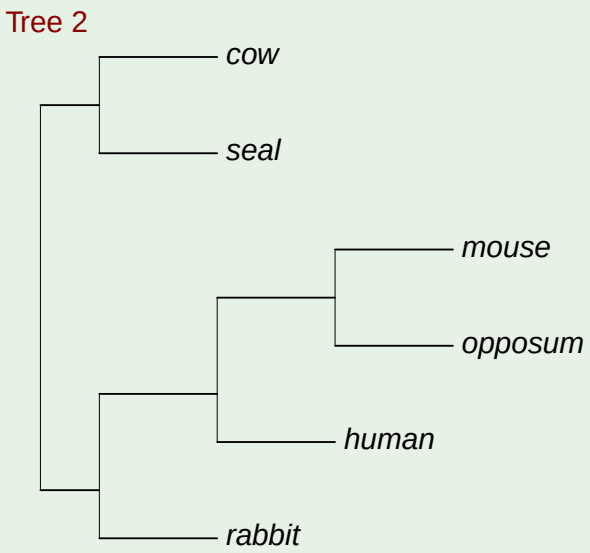
Tree 2



- Is Tree 2 plausible?

Confidence Sets of Trees are Equivalent to One Tree Tests

One Tree



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

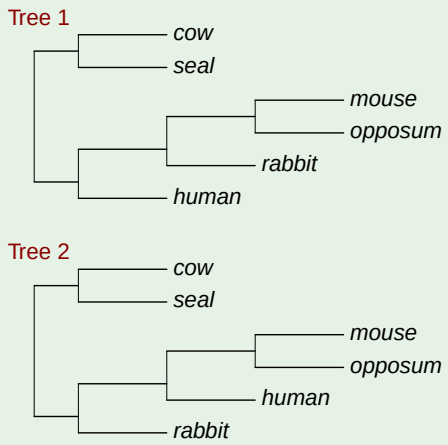
  $P[\text{True Tree in } C] \geq 0.95$

- 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

SH Test

Shimodaira & Hasegawa (1999) Mol Biol Evol 16:1114

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?

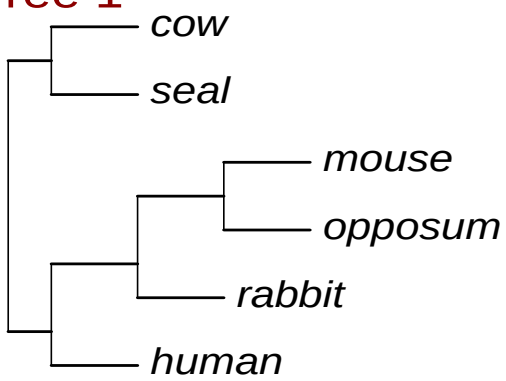
  Selection Bias: Tree 1 is selected based on the data

- 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

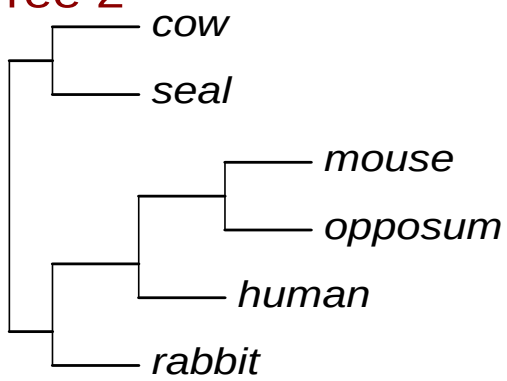
- **Setting**: T_1 and T_2 become $T_1, \ldots, T_M \Rightarrow l_1, \ldots, l_M$
  - Mammal data. 6 taxa $\Rightarrow M = 105$ trees.
  - Possibly less due to constraints
    eg. (Cow,Harbour Seal) $\Rightarrow M = 15$
  - Possibly less for pragmatic reasons.
    10 taxa $\Rightarrow M = 2,027,025$
- Test statistic $l_1 - l_2$ replaced by $l_m - l_1$
  m indice of MLE.
- **RELL Bootstrapping**
  - Replace $l_1^*, \ldots, l_M^*$ by
    $l_1^* - \text{ave}_b l_1^*, \ldots, l_M^* - \text{ave}_b l_M^*$
  - Use observed $l_{m^*}^* - l_2^*$ from bootstrapping for null distribution.
    m^*: indice of MLE for bootstrap sample.

Mammal Data Example - Three trees

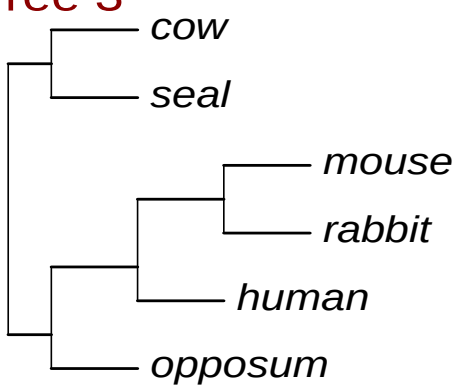
Tree 1



Tree 2



Tree 3



- 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

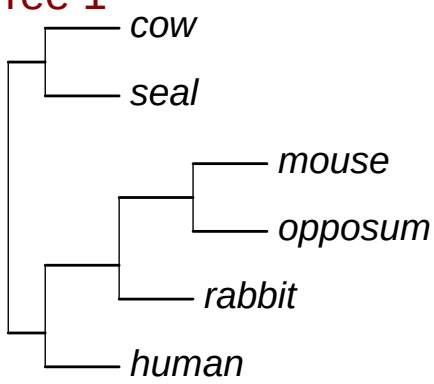
l_1^*	l_2^*	l_3^*	m^*	$l_1^* - l_2^*$	$l_{m^*} - l_2^*$
-359.78	-360.62	-352.52	3	0.84	8.10
-84.45	-94.44	-95.87	1	9.99	9.99
-65.93	-58.62	-62.19	2	-7.31	0.00

$$pKH = \text{proportion of } l_1^* - l_2^* > 1.19 = 0.45$$

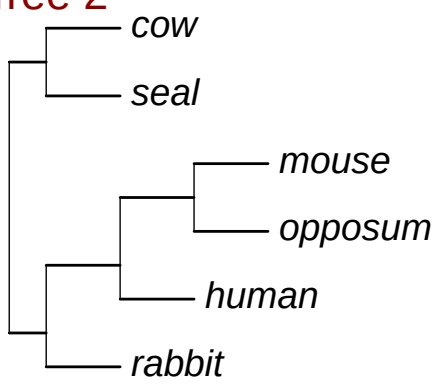
$$pSH = \text{proportion of } l_{m^*}^* - l_2^* > 1.19 = 0.59$$

Mammal Data Example - Four trees

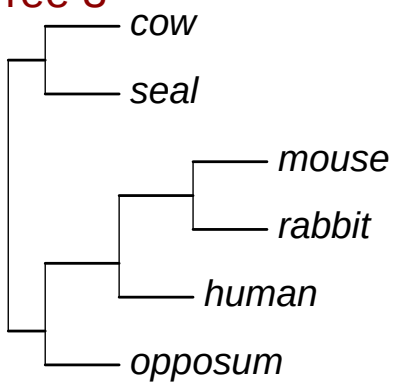
Tree 1



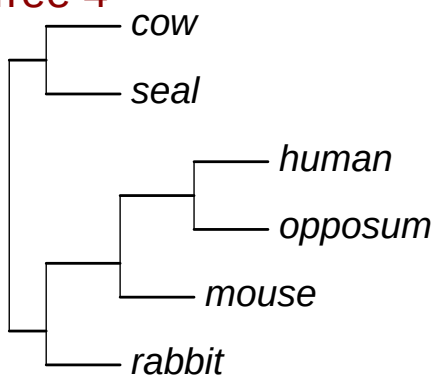
Tree 2



Tree 3



Tree 4



Mammal Data Example - Four trees

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

$$pKH = \text{proportion of } l_1^* - l_2^* > 1.19 = 0.45$$

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

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

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

SH test - Choice of Trees

- 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} = Proportion of $l_m^* - l_2^* > l_m - l_2$
- $p_{SOWH} << p_{SH}$
- Sometimes, SOWH will generate from a fully-resolved Tree 2 Under fully-resolved Tree 2, less likely to see large $l_m - l_2$

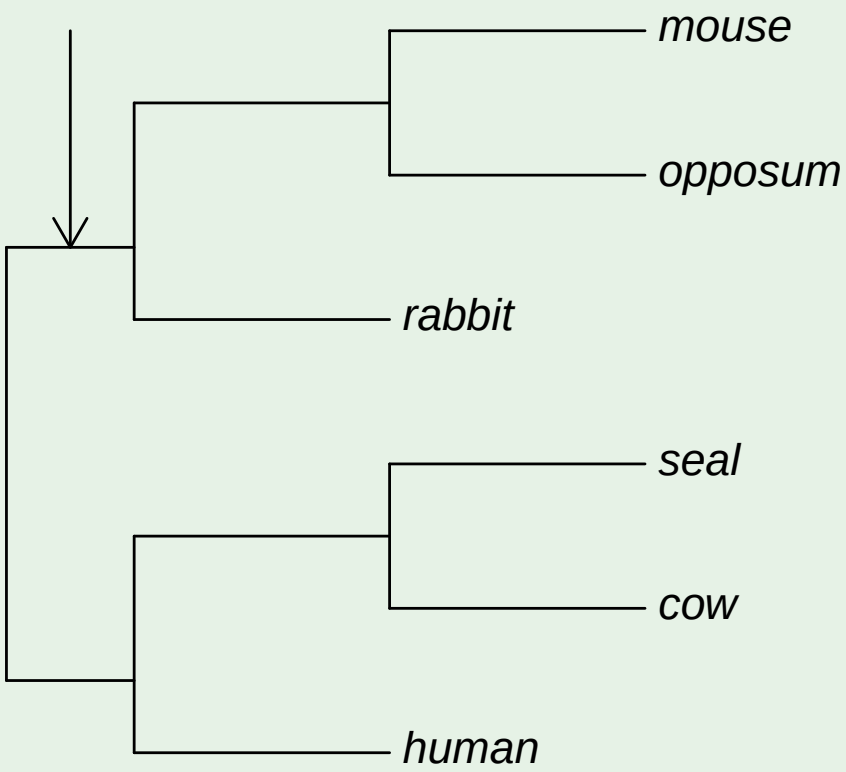
p-value comparison

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

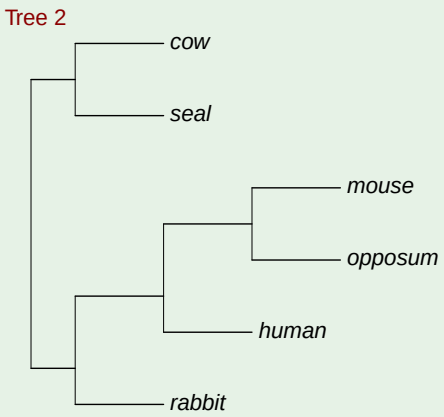
Split of Interest



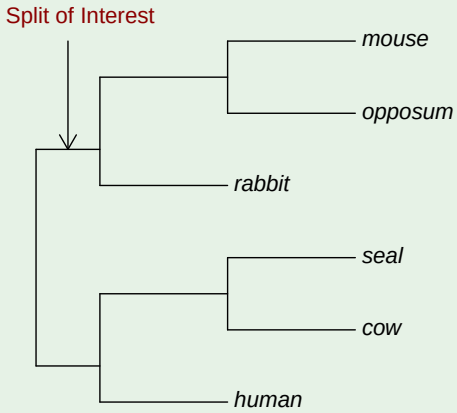
- Significant evidence that the split is present?

Support for Splits

One Tree



Splits



- Support value $=$ 1-p-value for test of

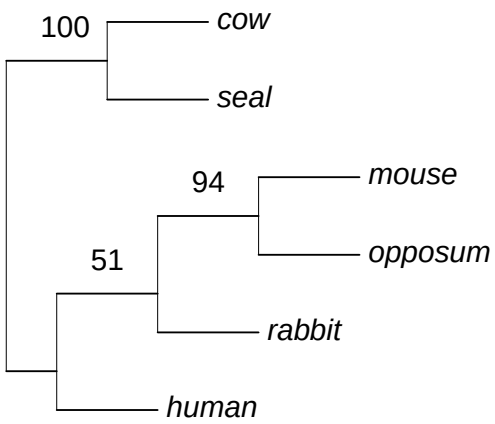
$$H_0 : \text{split } S \text{ is not present \& } H_A \text{ split is present}$$

- $$P[\text{Type I error}] = P[\text{Reject } H_0; H_0 \text{ true}]$$

No unique probability. H_0 true for any $(\tau, \boldsymbol{t})$ with a split S_c that is incompatible with S.

Bootstrap Support

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

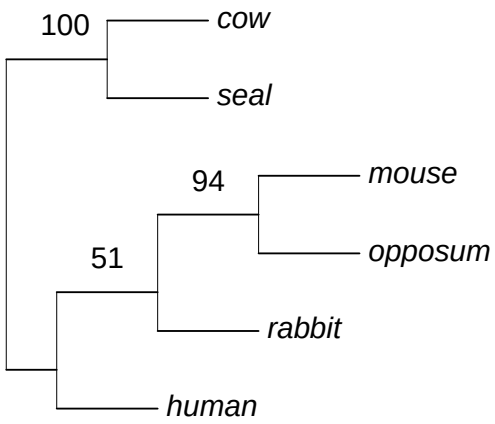


- By far the most frequent measure of uncertainty
- How large of BP is large?

History

- 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 if H_0 true

Bootstrap Support for Splits



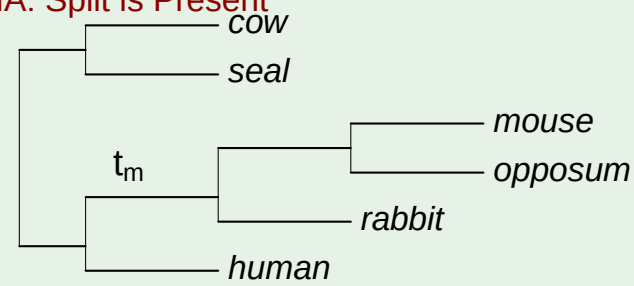
- Fixed Tree: BP is conservative:
  - Expect $>$95% BP less than 5% of time if H_0 true
  - Expect $<$ 40% BP more than 40% of time
- Selection Bias: ML tree is not fixed a priori. Very unlikely BP$<$ 40%

aLRT Support Values

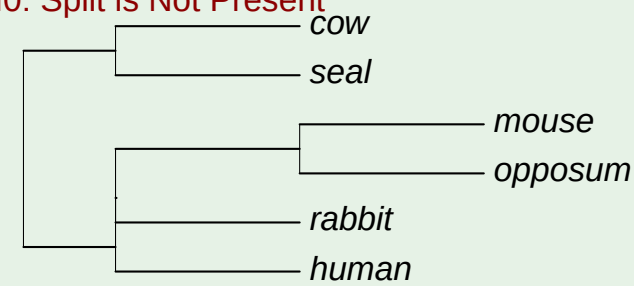
Anisimova & Gascuel (2006) Syst Biol 55:539

Split Support

HA: Split is Present

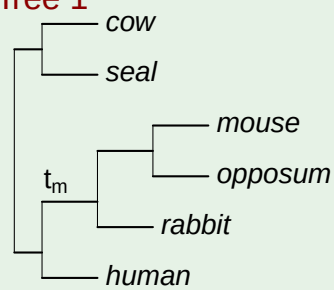


H0: Split is Not Present

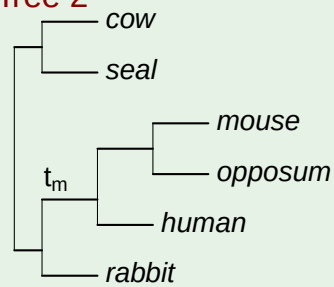


Tree 1 vs Tree 2

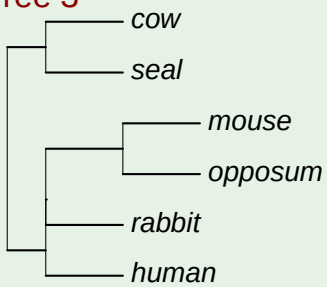
Tree 1



Tree 2

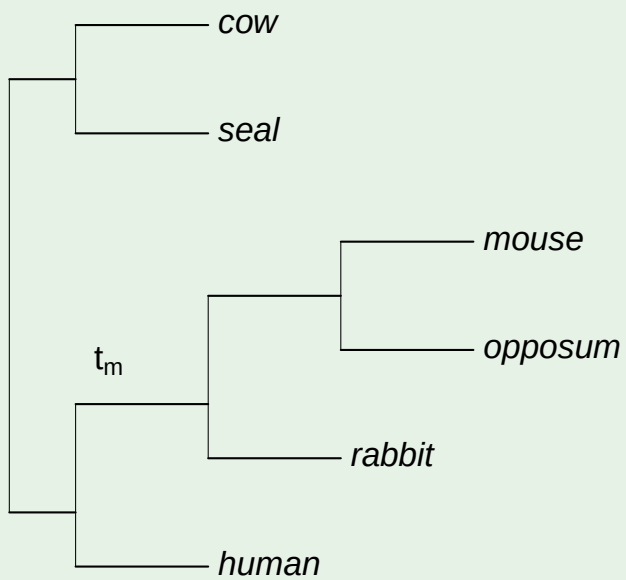


Tree 3



Split Not Present vs Present

Tree 1



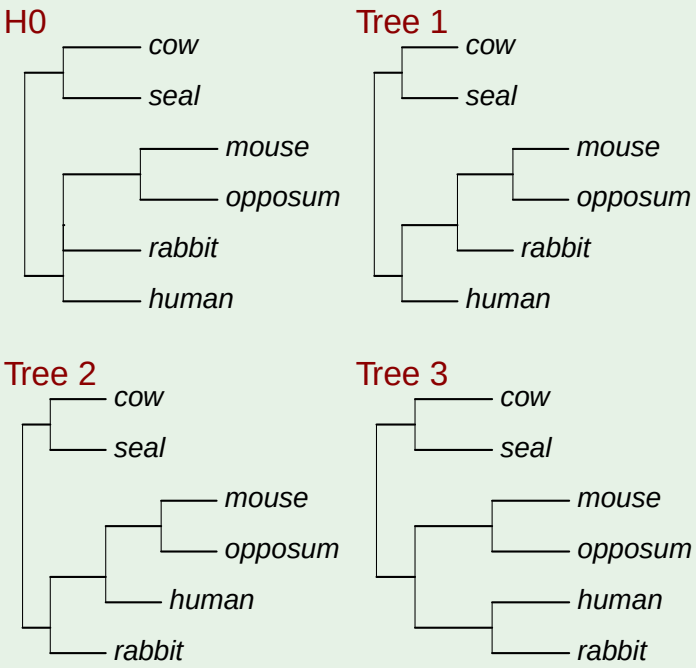
- Null hypothesis Tree 2 is a special case of Tree 1
  $\Rightarrow$ Conventional parametric model test
- $H_0 : t_m = 0$ vs $H_A : t_m > 0$

- Chi-bar Test with $p = 1$ (Ota et al. 2000):

$$\text{p-value} = P(\chi^2_0 > 2\Lambda_3)/2 + P(\chi^2_1 > 2\Lambda_3)/2$$

aLRT Adjustment for Selection Bias

Split Not Present vs Present



- Selection Bias: Tree 1 is not fixed (ML Tree)
- Tree 1 gives largest test statistic among 1,2,3

- Anisimova & Gascuel: Use $2\Lambda_2 = 2\{l_1 - l_2\}$ in place of $2\Lambda_3$ where l_2 - second best LnL and

$$p(\hat{\tau}) = 3\{P(\chi_0^2 > 2\Lambda_2)/2 + P(\chi_1^2 > 2\Lambda_2)/2\}$$

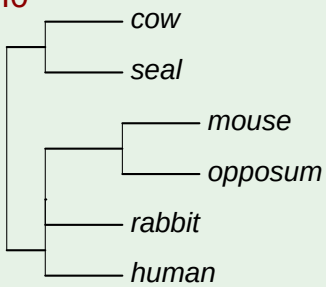
- Then $p(\hat{\tau}) < \alpha$ less than $\alpha \times 100\%$ of time

aLRT Support Value

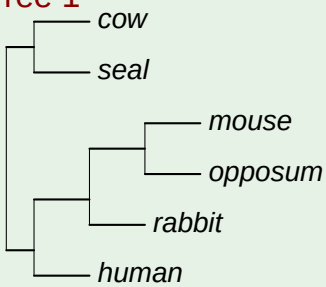
$$p(\hat{\tau}) = 3\{P(\chi_0^2 > 2\Lambda_2)/2 + P(\chi_1^2 > 2\Lambda_2)/2\}$$

Split Not Present vs Present

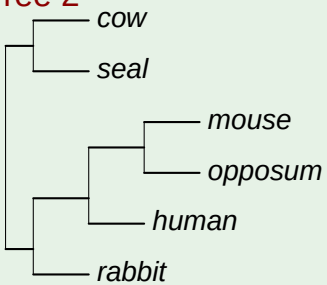
H0



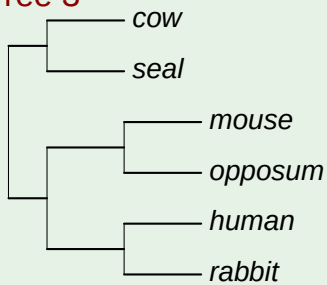
Tree 1



Tree 2



Tree 3



- aLRT Support Value

$$100 \times [1 - p(\hat{\tau})]\%$$

- Key Assumption: Only Trees 1-3 competing for ML Status

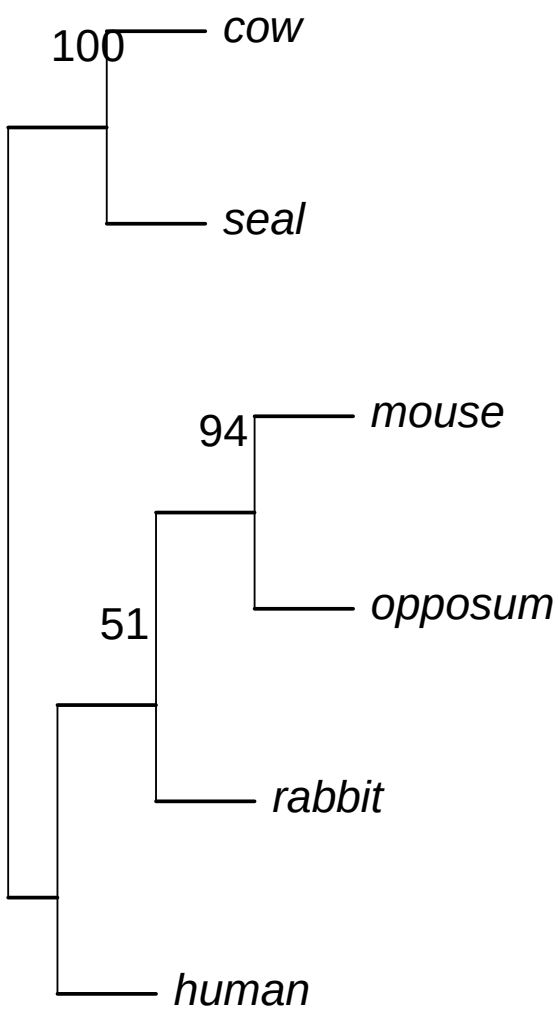
IQ-TREE aLRT Support Value

```
$ iqtree -s mtprot.phy -m mtREV+F+G8 -alrt 0
$ cat mtprot.phy.treefile
```

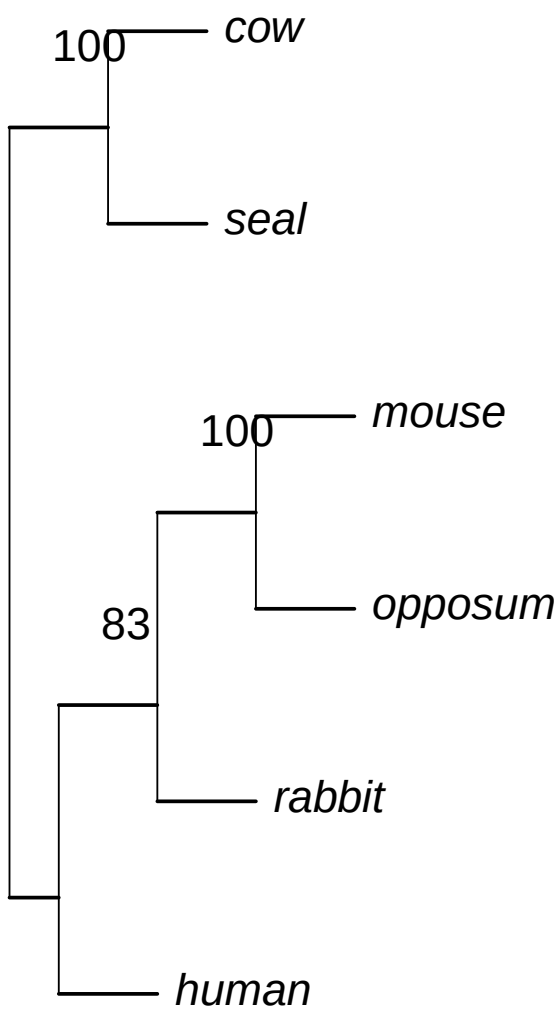
(human:0.27,(seal:0.08,cow:0.07)/1:0.04,
(rabbit:0.11,(mouse:0.19,opposum:0.29)/1:0.046) /0.833:0.02);

Support Value Comparison

Bootstrap Support



aLRT Support



Bootstrap Support for Trees

- For each bootstrap sample $x_1^*, \ldots, x_n^*$ 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
  $\Longrightarrow$ conservative. Expect 95% BP less than 5% of time
- AU Test (Shimoadiara (2002) Syst Biol 5:492). Bootstrap correction
  - Bootstraps with differing fractions of original sample size
    $\Longrightarrow BP(r_1) \ldots, BP(r_m)$.
  - AU p-value is transformation of $BP(r_1) \ldots, BP(r_m)$
  - 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);

Problem of Selection Bias

- Khns gives accurate p-values for $l_1 - l_2$. So

  p-value < 0.05 approximately 5% of time

- Selection bias

$$l_m - l_2 = \max_{j \in \{1,\ldots,M\}} \{l_j - l_2\}$$

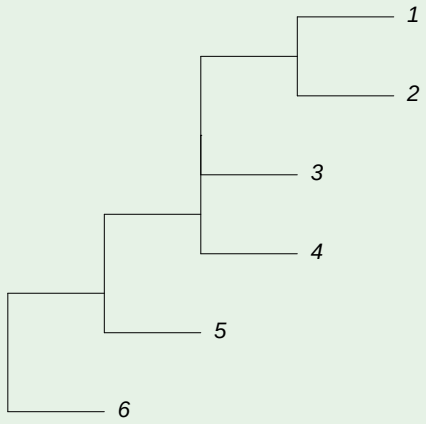
- Suppose that $l_j - l_2$ independent. Then chance p-value < 0.05

Taxa	Trees	$1 - [1 - 0.05]^M$
4	3	14%
5	15	54%
6	105	99.5%

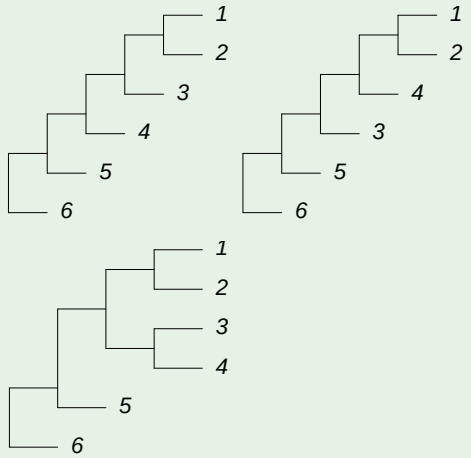
- In phylogenetics
  - $l_j - l_2$ are not independent
  - KH and χ^2 give p-value < 0.05, (much) less than 5% of time

Selection Bias Effect Depends on True Tree

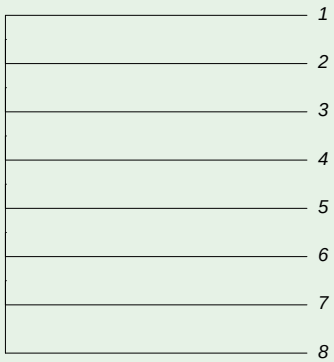
One poorly-resolved edge



Plausible ML Trees



Star Tree



- Each of the 10,395 trees equally likely to be ML tree ($M = 10,395$)

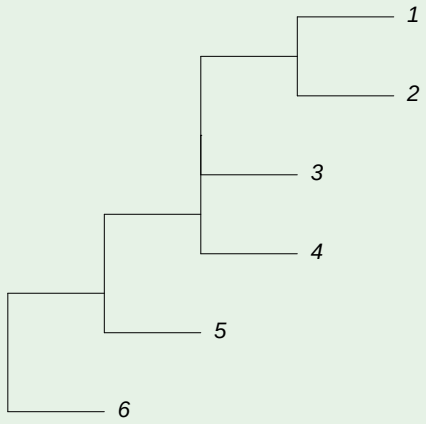
Selection Bias - Extreme Cases

			One Zero-length Edge		
t_l	SH	KH	AU	χ^2	Bo
0.1	100/37	100/3	100/3	95/3	98/3
0.01	100/2905	100/615	100/297	96/47	98/1112
		Two Adjacent Zero-length Edges			
0.1	100/76	100/15	99/15	91/14	98/15
0.01	100/5843	100/2097	91/512	89/254	100/4112
			Star Tree		
Star	100/10395	99/10247	11/1109	71/7335	100/10394

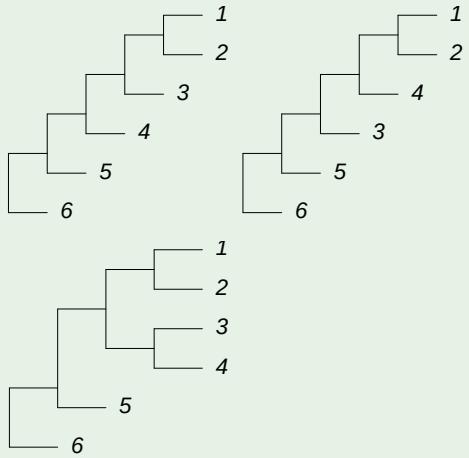
- Entries are Coverage/Mean Set Size

Selection Bias Effect Depends on True Tree - aLRT Support Values

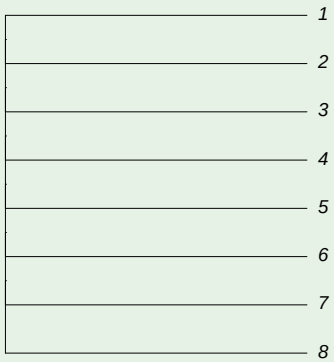
One poorly-resolved edge



Plausible ML Trees



Star Tree



- Each of the 10,395 trees equally likely to be ML tree ($M = 10,395$)

Selection Bias - Extreme Cases

One Zero-length Edge		
t_l	Type I Error Rate	Power
0.1	4%	100%
0.01	6%	73%
Two Adjacent Zero-length Edges		
t_l	Type I Error Rate	Power
0.1	5%	100%
0.01	8%	69%
Star Tree		
t_l	Type I Error Rate	Power
Star	7%	NA%

- Type I: Proportion of incorrect splits in ML tree that were supported
- Power: Proportion of correct splits in ML tree that were supported

Concluding Comments

- Trees
  - SH too conservative. Values depends on number of input trees M
  - 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. Hardly affected by selection bias
  - Chi-square (Naive) reasonable but somewhat affected by selection bias
- Splits
  - BP is conservative
  - aLRT works well even with extreme selection bias