Phylogenetic Support

Statistical Testing of Trees

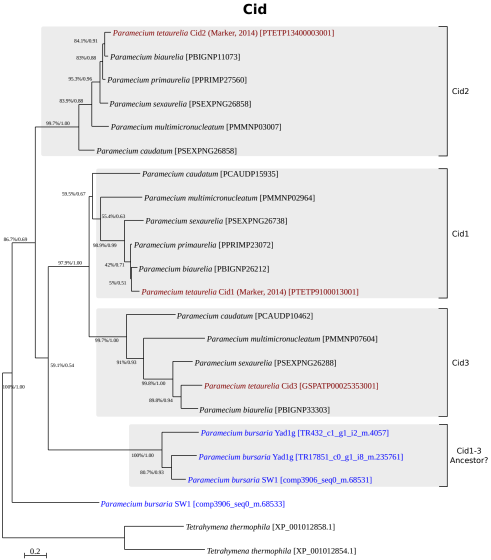
Finlay Maguire

April 1, 2020

FCS, Dalhousie

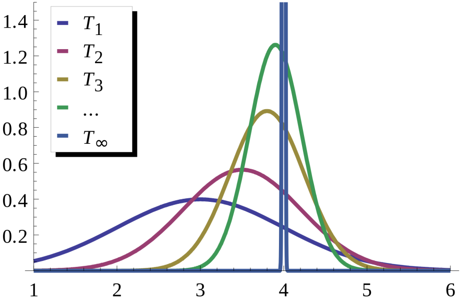
Introduction

Phylogenies are hypotheses



Hypotheses can be wrong

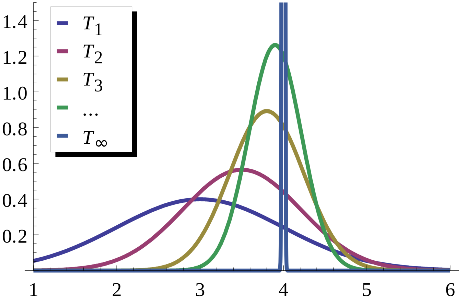
Assessing phylogenetic accuracy



`bit.ly/3dHBiPT`

- Consistency

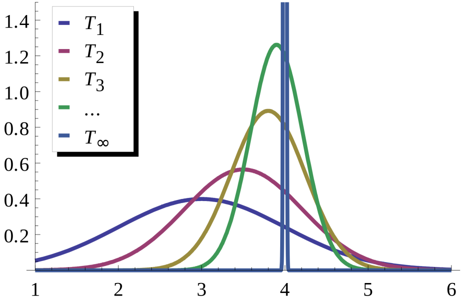
Assessing phylogenetic accuracy



`bit.ly/3dHBiPT`

- Consistency
- Efficiency

Assessing phylogenetic accuracy



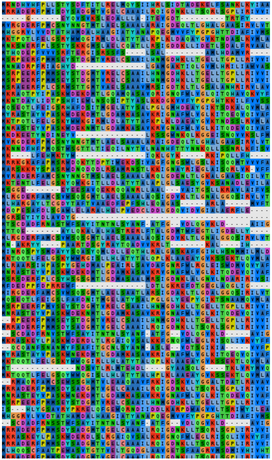
`bit.ly/3dHBiPT`

- Consistency
- Efficiency
- Robustness

Sources of Error: Bad Data

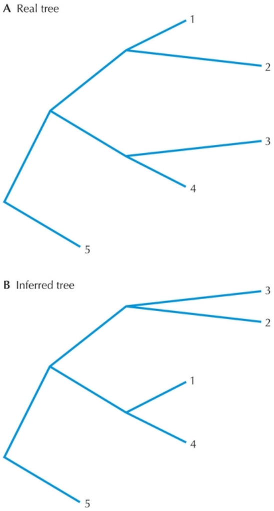
Throw unrelated sequences at an aligner

Sources of Error: Bad Data



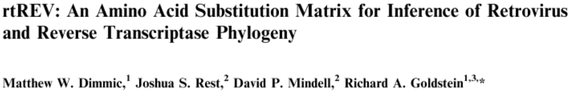
Throw alignment at a trimmer

Sources of Error: Sampling



[Barton, 2007]

Choosing the wrong model



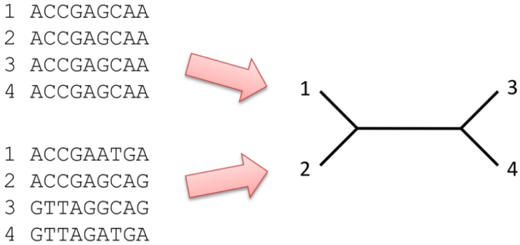
rtREV: An Amino Acid Substitution Matrix for Inference of Retrovirus and Reverse Transcriptase Phylogeny

Matthew W. Dimmic,[1] Joshua S. Rest,[2] David P. Mindell,[2] Richard A. Goldstein[1,3,*]

[Dimmic et al., 2002]

Choosing the wrong paradigm

Ask for a tree get a tree.

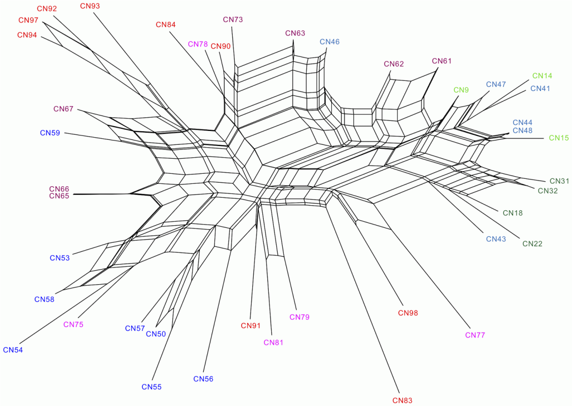


1 ACCGAGCAA
2 ACCGAGCAA
3 ACCGAGCAA
4 ACCGAGCAA

1 ACCGAATGA
2 ACCGAGCAG
3 GTTAGGCAG
4 GTTAGATGA

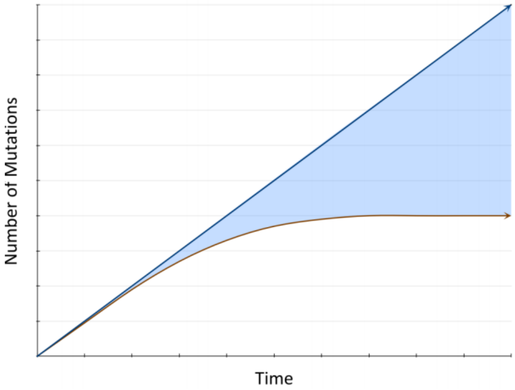
Tree not always correct paradigm

Ask for a tree get a tree.



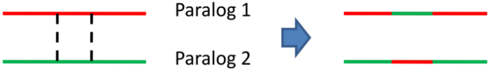
Reanalysis of [Marwick, 2012] from `http://phylonetworks.blogspot.ca/2013/02/`

Misleading Evolutionary Signals: Saturation

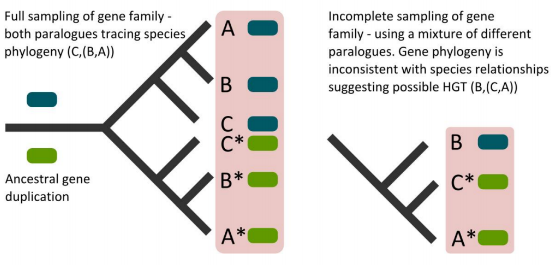


[Leonard, 2010]

Misleading Evolutionary Signals: Recombination

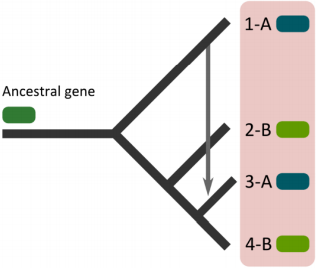


Misleading Evolutionary Signals: Hidden Paralogy/Incomplete Sampling



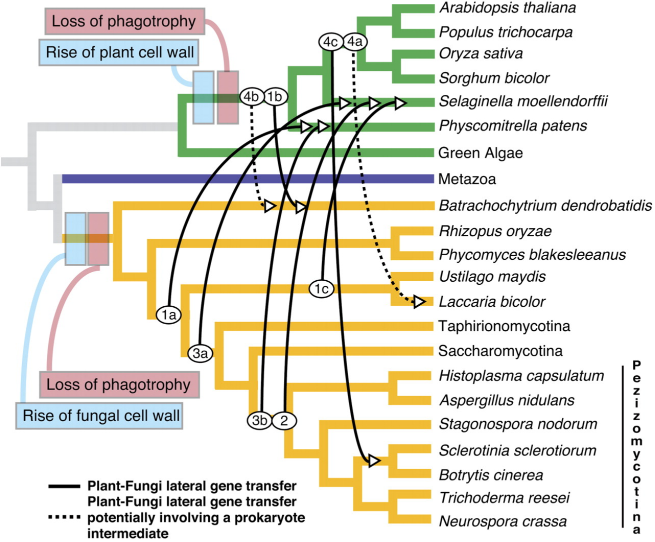
[Leonard, 2010]

Misleading Evolutionary Signals: Horizontal Gene Transfer



[Leonard, 2010]

Misleading Evolutionary Signals: Horizontal Gene Transfer



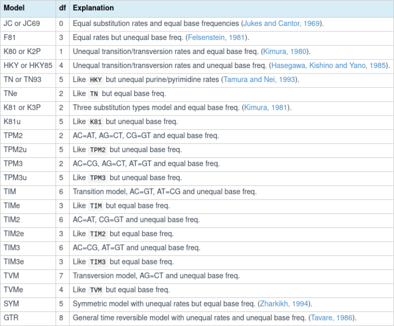
[Richards et al., 2009]

Phylogenetic Statistics: Overview

1. Sequence Evolution Model Selection
2. Testing Branch Support
3. Comparing Trees

Sequence Evolution Model Selection

Lots of Nucleotide Evolution Models

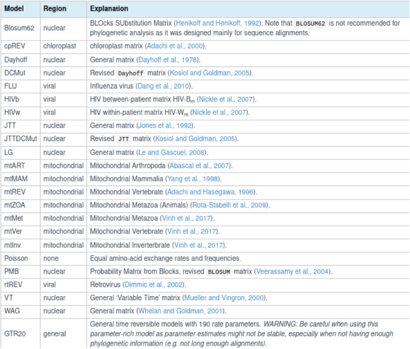


Model	df	Explanation
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).
TNe	2	Like `TN` but equal base freq.
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).
K81u	5	Like `K81` but unequal base freq.
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.
TPM2u	5	Like `TPM2` but unequal base freq.
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.
TPM3u	5	Like `TPM3` but unequal base freq.
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.
TIMe	3	Like `TIM` but equal base freq.
TIM2	6	AC=AT, CG=GT and unequal base freq.
TIM2e	3	Like `TIM2` but equal base freq.
TIM3	6	AC=CG, AT=GT and unequal base freq.
TIM3e	3	Like `TIM3` but equal base freq.
TVM	7	Transversion model, AG=CT and unequal base freq.
TVMe	4	Like `TVM` but equal base freq.
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).

`iqtree.org`

- +I, +G, +I+G, +R, +I, +Asc, +F, +FQ, +FO = 286 models
- 195 codon models
- Not even including partitioning or mixtures

Even More Protein Evolution Models



Model	Region	Explanation
Blosum62	nuclear	BLOcks SUbstitution Matrix (Henikoff and Henikoff, 1992). Note that `BLOSUM62` is not recommended for phylogenetic analysis as it was designed mainly for sequence alignments.
cpREV	chloroplast	chloroplast matrix (Adachi et al., 2000).
Dayhoff	nuclear	General matrix (Dayhoff et al., 1978).
DCMut	nuclear	Revised `Dayhoff` matrix (Kosiol and Goldman, 2005).
FLU	viral	Influenza virus (Dang et al., 2010).
HIVb	viral	HIV between-patient matrix HIV-B_m (Nickle et al., 2007).
HIVw	viral	HIV within-patient matrix HIV-W_m (Nickle et al., 2007).
JTT	nuclear	General matrix (Jones et al., 1992).
JTTDCMut	nuclear	Revised `JTT` matrix (Kosiol and Goldman, 2005).
LG	nuclear	General matrix (Le and Gascuel, 2008).
mtART	mitochondrial	Mitochondrial Arthropoda (Abascal et al., 2007).
mtMAM	mitochondrial	Mitochondrial Mammalia (Yang et al., 1998).
mtREV	mitochondrial	Mitochondrial Vertebrate (Adachi and Hasegawa, 1996).
mtZOA	mitochondrial	Mitochondrial Metazoa (Animals) (Rota-Stabelli et al., 2009).
mtMet	mitochondrial	Mitochondrial Metazoa (Vinh et al., 2017).
mtVer	mitochondrial	Mitochondrial Vertebrate (Vinh et al., 2017).
mtInv	mitochondrial	Mitochondrial Inverterbrate (Vinh et al., 2017).
Poisson	none	Equal amino-acid exchange rates and frequencies.
PMB	nuclear	Probability Matrix from Blocks, revised `BLOSUM` matrix (Veerassamy et al., 2004).
rtREV	viral	Retrovirus (Dimmic et al., 2002).
VT	nuclear	General 'Variable Time' matrix (Mueller and Vingron, 2000).
WAG	nuclear	General matrix (Whelan and Goldman, 2001).
GTR20	general	General time reversible models with 190 rate parameters. *WARNING: Be careful when using this parameter-rich model as parameter estimates might not be stable, especially when not having enough phylogenetic information (e.g. not long enough alignments).*

`iqtree.org`

- +I, +G, +I+G, +R, +I, +Asc, +F, +FQ, +FO = 539 models
- Also not including even more mixture models

What happens theoretically if the wrong model is specified?

- Increased Inaccuracy (wrong tree more often)

What happens theoretically if the wrong model is specified?

- Increased Inaccuracy (wrong tree more often)
- Inconsistency (adding more data converges to wrong tree)

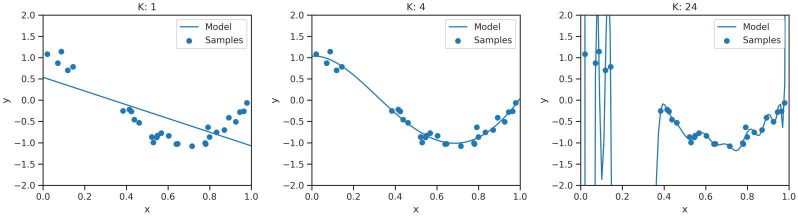
What happens theoretically if the wrong model is specified?

- Increased Inaccuracy (wrong tree more often)
- Inconsistency (adding more data converges to wrong tree)
- Wrong branch lengths (important for certain analyses)

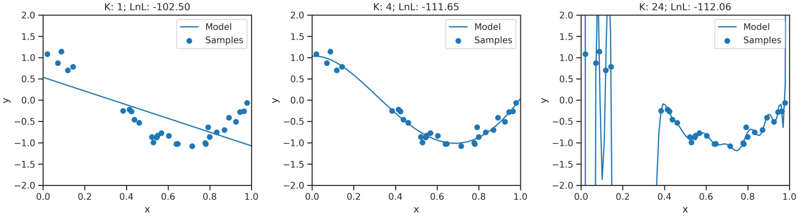
What happens theoretically if the wrong model is specified?

- Increased Inaccuracy (wrong tree more often)
- Inconsistency (adding more data converges to wrong tree)
- Wrong branch lengths (important for certain analyses)
- Wrong tree support values

How do we select a model?

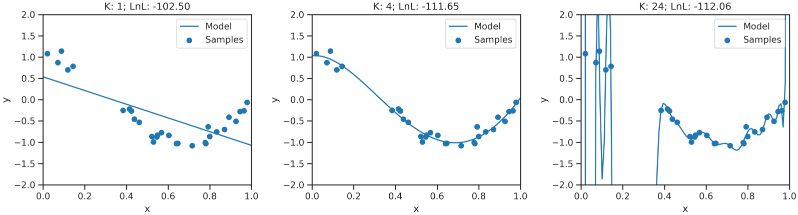


Model Likelihood



- $\mathcal{L}(\theta|X) = p(X|\theta)$

Model Likelihood



- $\mathcal{L}(\theta|X) = p(X|\theta)$
- $\ell = ln(\mathcal{L})$

Information Criterion

- Model Maximum Likelihood ($\hat{\ell}$) + Model Complexity (K) Penalty Term

Information Criterion

- Model Maximum Likelihood ($\hat{\ell}$) + Model Complexity (K) Penalty Term
- Akaike Information Criterion (AIC)

Information Criterion

- Model Maximum Likelihood ($\hat{\ell}$) + Model Complexity (K) Penalty Term
- Akaike Information Criterion (**AIC**)
- **AIC** = $-2\hat{\ell} + 2K$

Information Criterion

- Model Maximum Likelihood ($\hat{\ell}$) + Model Complexity (K) Penalty Term
- Akaike Information Criterion (**AIC**)
- **AIC** = $-2\hat{\ell} + 2K$
- Applying to phylogenies: τ and θ are the parameters we care about

Information Criterion

- Model Maximum Likelihood ($\hat{\ell}$) + Model Complexity (K) Penalty Term
- Akaike Information Criterion (**AIC**)
- **AIC** = $-2\hat{\ell} + 2K$
- Applying to phylogenies: τ and θ are the parameters we care about
- Infer quick phylogeny on your alignment: $\ell(\tau, \theta | X) = p(X | \tau, \theta)$

Other criteria

- **AIC**: minimise $-2\hat{\ell} + 2K$

Other criteria

- **AIC**: minimise $-2\hat{\ell} + 2K$
- **AICc**: AIC + $\frac{2K(K+1)}{n-K-1}$

Other criteria

- **AIC**: minimise $-2\hat{\ell} + 2K$
- **AICc**: AIC + $\frac{2K(K+1)}{n-K-1}$
- **BIC**: minimise $-2\hat{\ell} + K \ln n$

Other criteria

- **AIC**: minimise $-2\hat{\ell} + 2K$
- **AICc**: AIC + $\frac{2K(K+1)}{n-K-1}$
- **BIC**: minimise $-2\hat{\ell} + K \ln n$
- **DT**: BIC + relative branch-length error

Other criteria

- **AIC**: minimise $-2\hat{\ell} + 2K$
- **AICc**: AIC + $\frac{2K(K+1)}{n-K-1}$
- **BIC**: minimise $-2\hat{\ell} + K \ln n$
- **DT**: BIC + relative branch-length error
- **BF**: ratio of marginal likelihoods

Limitations

- What if everything fits poorly?

Limitations

- What if everything fits poorly?
- Relative goodness of fit NOT absolute

Limitations

- What if everything fits poorly?
- Relative goodness of fit NOT absolute
- Parametric Bootstrapping/Posterior Predictive Simulation

Limitations

- What if everything fits poorly?
- Relative goodness of fit NOT absolute
- Parametric Bootstrapping/Posterior Predictive Simulation
- If the model is reasonable then data simulated under should resemble the empirical data

Does it REALLY matter?

nature
communications

Article | Open Access | Published: 25 February 2019

Model selection may not be a mandatory step for phylogeny reconstruction

Shiran Abadi, Dana Azouri, Tal Pupko & Itay Mayrose

Nature Communications **10**, Article number: 934 (2019) | Cite this article

[Abadi et al., 2019]

Does it REALLY matter?

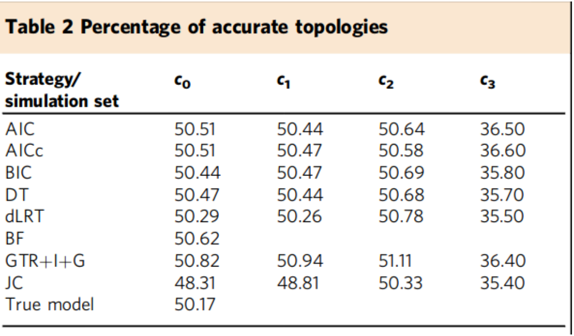


Table 2 Percentage of accurate topologies

Strategy/ simulation set	c_0	c_1	c_2	c_3
AIC	50.51	50.44	50.64	36.50
AICc	50.51	50.47	50.58	36.60
BIC	50.44	50.47	50.69	35.80
DT	50.47	50.44	50.68	35.70
dLRT	50.29	50.26	50.78	35.50
BF	50.62			
GTR+I+G	50.82	50.94	51.11	36.40
JC	48.31	48.81	50.33	35.40
True model	50.17			

[Abadi et al., 2019]

Does it REALLY matter?

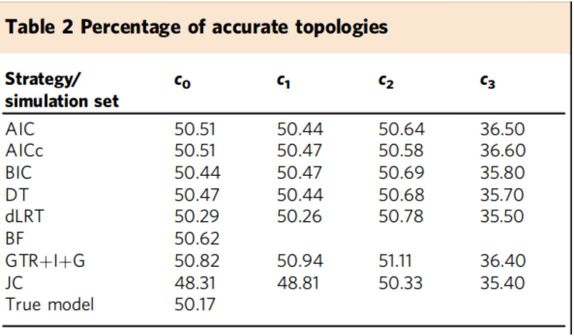


Table 2 Percentage of accurate topologies

Strategy/ simulation set	c_0	c_1	c_2	c_3
AIC	50.51	50.44	50.64	36.50
AICc	50.51	50.47	50.58	36.60
BIC	50.44	50.47	50.69	35.80
DT	50.47	50.44	50.68	35.70
dLRT	50.29	50.26	50.78	35.50
BF	50.62			
GTR+I+G	50.82	50.94	51.11	36.40
JC	48.31	48.81	50.33	35.40
True model	50.17			

[Abadi et al., 2019]

- Criteria are inconsistent (BIC/AIC disagree in 62% of cases)

Does it REALLY matter?

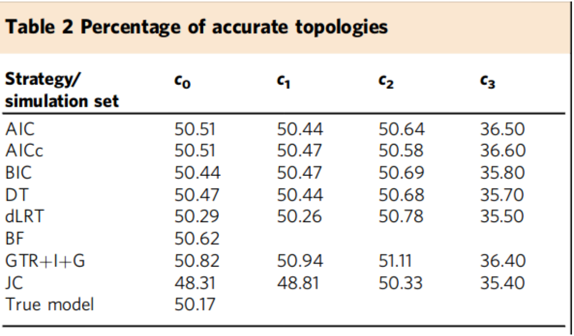


Table 2 Percentage of accurate topologies

Strategy/ simulation set	c_0	c_1	c_2	c_3
AIC	50.51	50.44	50.64	36.50
AICc	50.51	50.47	50.58	36.60
BIC	50.44	50.47	50.69	35.80
DT	50.47	50.44	50.68	35.70
dLRT	50.29	50.26	50.78	35.50
BF	50.62			
GTR+I+G	50.82	50.94	51.11	36.40
JC	48.31	48.81	50.33	35.40
True model	50.17			

[Abadi et al., 2019]

- Criteria are inconsistent (BIC/AIC disagree in 62% of cases)
- Different models change the distance matrix trivially.

Does it REALLY matter?

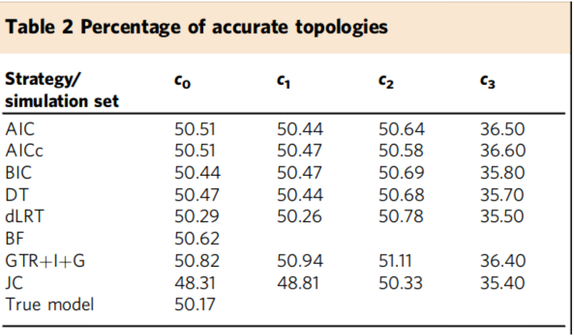


Table 2 Percentage of accurate topologies

Strategy/ simulation set	c_0	c_1	c_2	c_3
AIC	50.51	50.44	50.64	36.50
AICc	50.51	50.47	50.58	36.60
BIC	50.44	50.47	50.69	35.80
DT	50.47	50.44	50.68	35.70
dLRT	50.29	50.26	50.78	35.50
BF	50.62			
GTR+I+G	50.82	50.94	51.11	36.40
JC	48.31	48.81	50.33	35.40
True model	50.17			

[Abadi et al., 2019]

- Criteria are inconsistent (BIC/AIC disagree in 62% of cases)
- Different models change the distance matrix trivially.
- ALL models lead to generally similar topologies.

But!

- Likely still matters in complex cases

But!

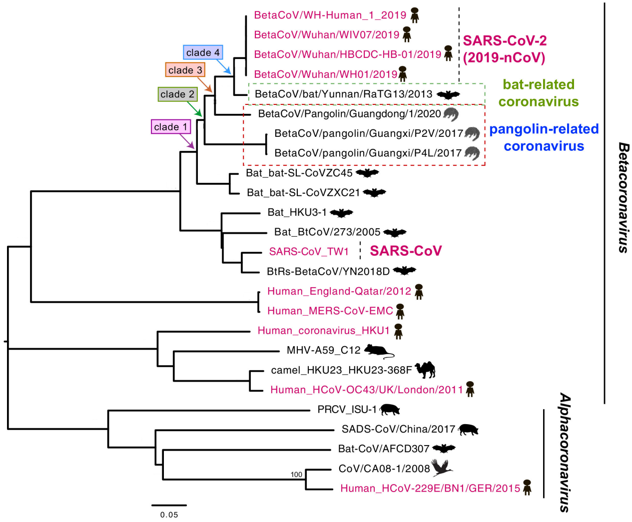
- Likely still matters in complex cases
- Definitely matter a lot for inferences on tree e.g. branch lengths

But!

- Likely still matters in complex cases
- Definitely matter a lot for inferences on tree e.g. branch lengths
- Pretty quick and easy to do these days, so why not do it?

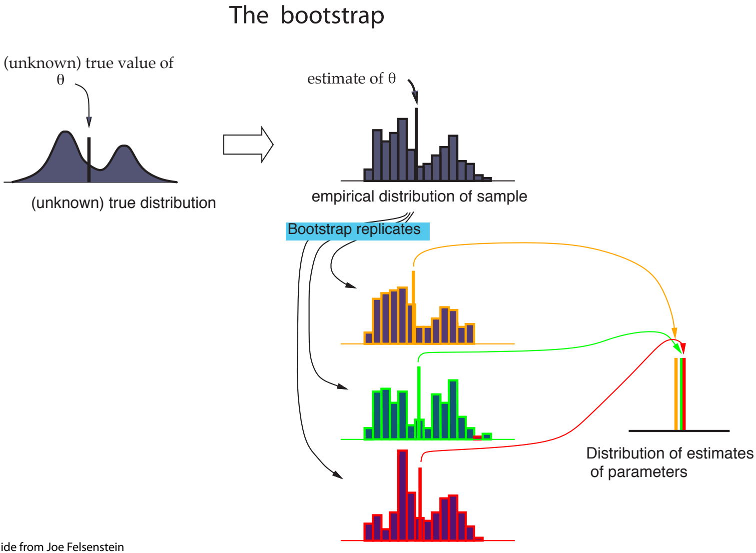
Testing Branch Support

Pangolin or Bat?



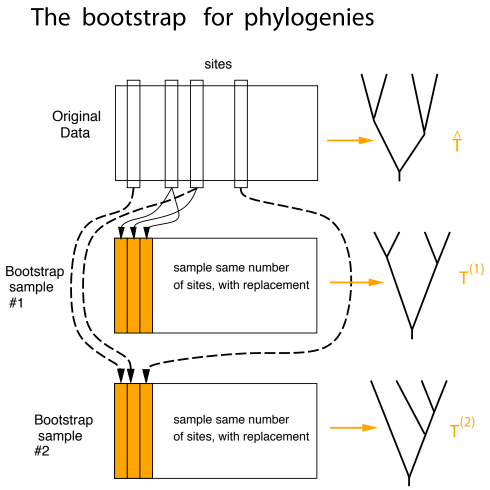
[Zhang et al., 2020] (disclaimer: better analyses exist but this was simplest)

Non-Parametric Bootstrapping in General



Slide from Joe Felsenstein

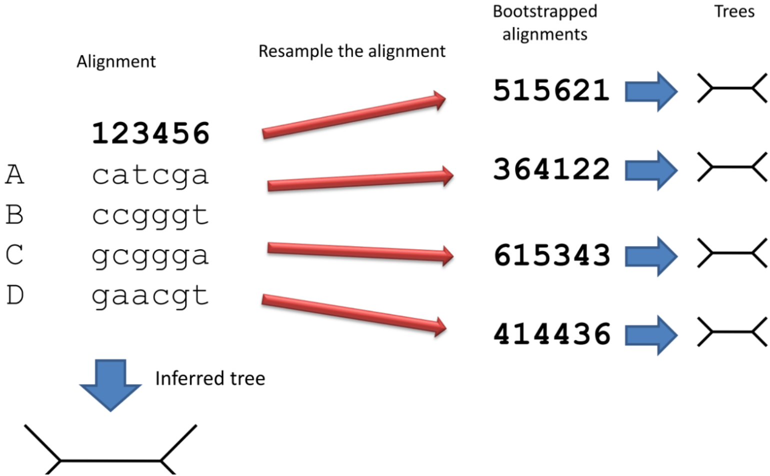
Bootstrapping Phylogenies



Slide from Joe Felsenstein

(and so on)

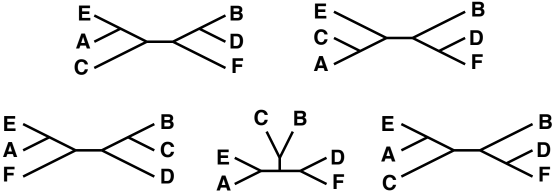
Bootstrapping Phylogenies



Bootstrapping Phylogenies

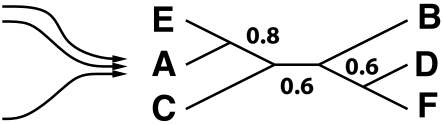
The majority-rule consensus tree

Trees:



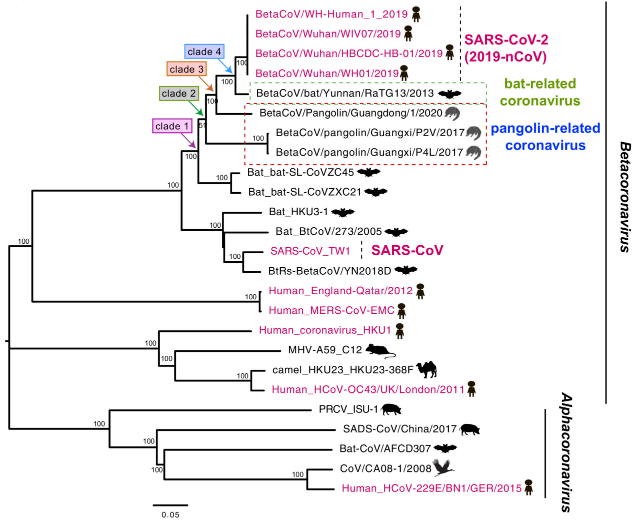
How many times each partition of species is found:

AE I BCDF	4
ACE I BDF	3
ACEF I BD	1
AC I BDEF	1
AEF I BCD	1
ADEF I BC	2
ABCE I DF	3



Slide from Joe Felsenstein

Combining the Results



[Zhang et al., 2020]

What are Non-Parametric Bootstraps Doing?

- Randomly reweighing the sites in an alignments

What are Non-Parametric Bootstraps Doing?

- Randomly reweighing the sites in an alignments
- Probability of a site being excluded $1 - \frac{1}{n}n$

What are Non-Parametric Bootstraps Doing?

- Randomly reweighing the sites in an alignments
- Probability of a site being excluded $1 - \frac{1}{n}n$
- Goal to simulate an infinite population (number of alignment columns)

Limitations

- Typically underestimates the true probabilities

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding (redoing full inference for each replicate)

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding (redoing full inference for each replicate)
- Assumes independence of sites

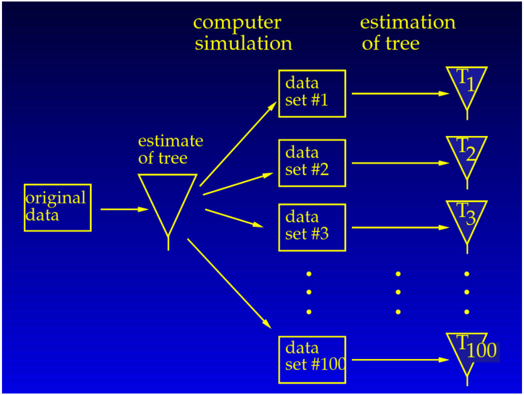
Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding (redoing full inference for each replicate)
- Assumes independence of sites
- Reliant on good input data

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding (redoing full inference for each replicate)
- Assumes independence of sites
- Reliant on good input data
- Only answers to what extent does input data support a given part of the tree

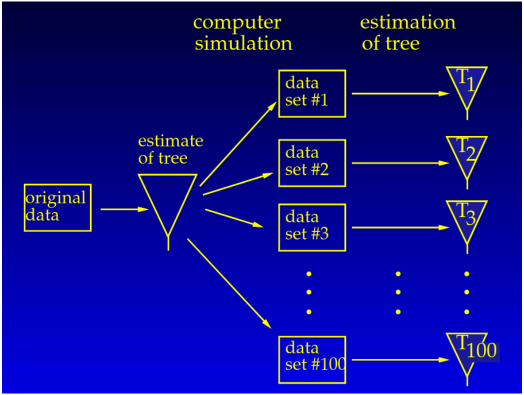
Parametric Bootstraps



Joe Felsenstein

- Key for many more sophisticated tests.

Parametric Bootstraps



Joe Felsenstein

- Key for many more sophisticated tests.
- Can be used to generate *p*-values, but non-trivial

Speeding-up Bootstraps

- Rapid Bootstraps (RBS): optimise branch length/model once and focus only on topology

Speeding-up Bootstraps

- Rapid Bootstraps (**RBS**): optimise branch length/model once and focus only on topology
- Re-sampling estimated log-likelihoods (**RELL**): keep site-wise ℓ from tree-search

Speeding-up Bootstraps

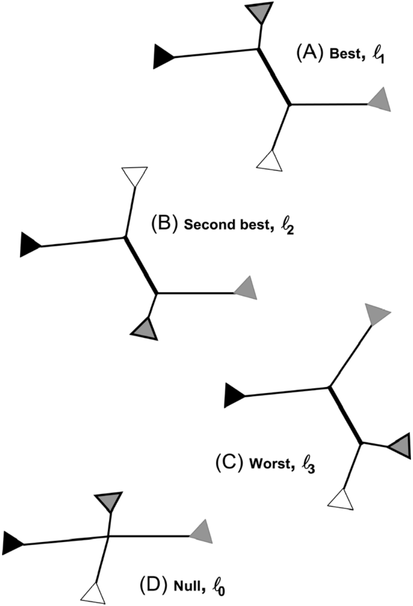
- Rapid Bootstraps (RBS): optimise branch length/model once and focus only on topology
- Re-sampling estimated log-likelihoods (RELL): keep site-wise ℓ from tree-search
- Ultrafast Bootstraps (UFBoot): Use trees from tree-search + RELL

Speeding-up Bootstraps

- Rapid Bootstraps (**RBS**): optimise branch length/model once and focus only on topology
- Re-sampling estimated log-likelihoods (**RELL**): keep site-wise ℓ from tree-search
- Ultrafast Bootstraps (**UFBoot**): Use trees from tree-search + RELL
- Always remember to recycle your computational results!

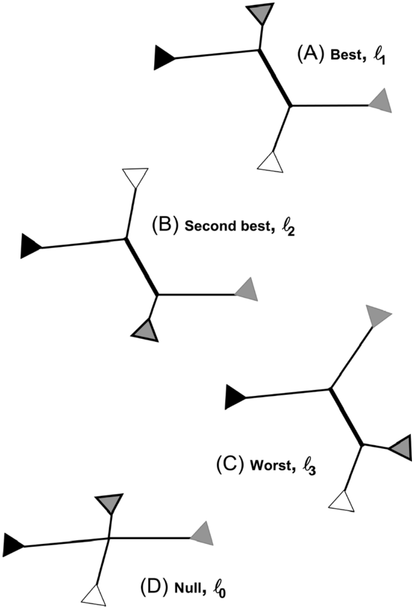
Is there a more efficient way?

Likelihood Tests



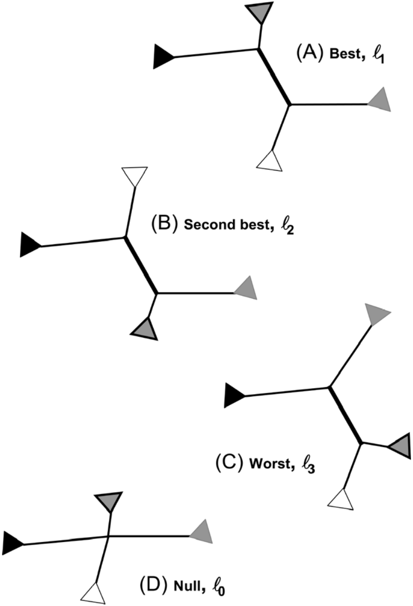
- Comparing the 3 nearest NNIs to a given branch:

Likelihood Tests



- Comparing the 3 nearest NNIs to a given branch:
- Parametric **aLRT**: χ^2 of δ for branch vs. closest NNIs

Likelihood Tests



- Comparing the 3 nearest NNIs to a given branch:
- Parametric **aLRT**: χ^2 of δ for branch vs. closest NNIs
- Non-parametric **SH-aLRT** based on RELL

Bayesian inference

- $p(\theta, \tau | X) = \frac{p(X|\theta,\tau)p(\theta)}{p(X)}$

Bayesian inference

- $p(\theta, \tau | X) = \frac{p(X|\theta,\tau)p(\theta)}{p(X)}$
- Use MCMC to deal with intractable $p(X)$

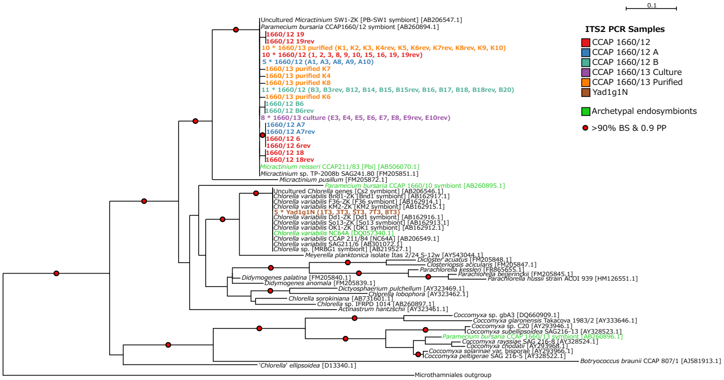
Bayesian inference

- $p(\theta, \tau | X) = \frac{p(X|\theta,\tau)p(\theta)}{p(X)}$
- Use MCMC to deal with intractable $p(X)$
- Frequency of clade in samples at stationarity: **posterior probability** of clade

Bayesian inference

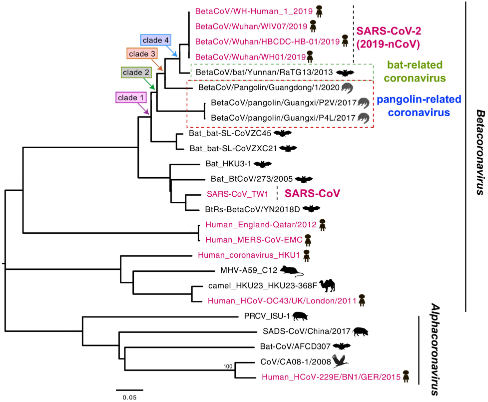
- $p(\theta, \tau|X) = \frac{p(X|\theta,\tau)p(\theta)}{p(X)}$
- Use MCMC to deal with intractable $p(X)$
- Frequency of clade in samples at stationarity: **posterior probability** of clade
- **aBayes**: estimate **PP** per branch from NNIs

Standard Approach: Use Multiple Methods



Comparing Trees

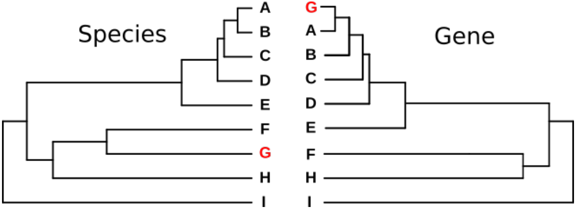
Competing Hypotheses



Rival Hypotheses

- ((Human, Bat), Pangolin)
- ((Human, Pangolin), Bat)

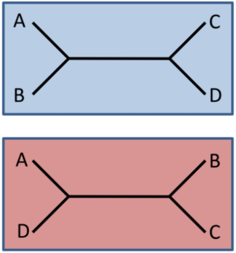
Rejecting Species Tree



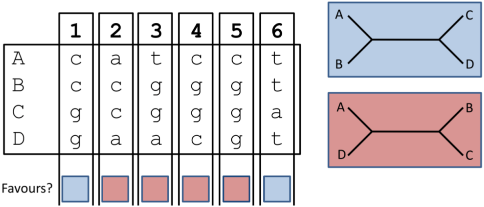
Lateral Gene Transfer

Simplistic Comparison

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?	🟦	🟥	🟥	🟥	🟥	🟦



Qualitative Comparison

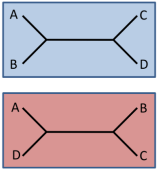


	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?	🟦	🟥	🟥	🟥	🟥	🟦

- 4 sites favour the red tree, 2 favour the blue

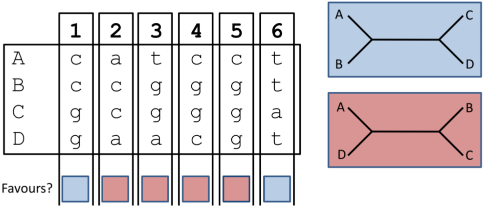
Qualitative Comparison

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?	🟦	🟥	🟥	🟥	🟥	🟦



- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k} p^k (1-p)^{n-k}$

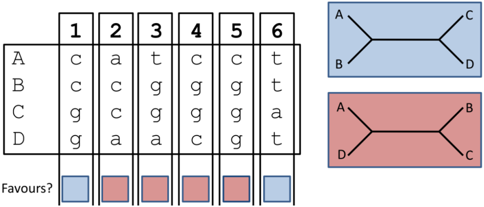
Qualitative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?	🟦	🟥	🟥	🟥	🟥	🟦

- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k} p^k (1-p)^{n-k}$
- 4 out of 6 $p = 0.6875$

Qualitative Comparison

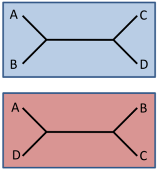


	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?	🟦	🟥	🟥	🟥	🟥	🟦

- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k}p^k(1-p)^{n-k}$
- 4 out of 6 $p = 0.6875$
- 40 out of 60 $p = 0.0124$

Qualitative Comparison

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?	🟦	🟥	🟥	🟥	🟥	🟦



- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k}p^k(1-p)^{n-k}$
- 4 out of 6 $p = 0.6875$
- 40 out of 60 $p = 0.0124$
- 400 out of 600 $p = 2.3 * 10^{-16}$

More robust approach

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.

More robust approach

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(T_1, T_2 \mid X) = 2\left[\ell(T_1 \mid X) - \ell(T_2 \mid X)\right]$

More robust approach

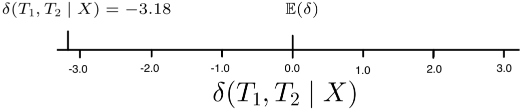
- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(T_1, T_2 \mid X) = 2\left[\ell(T_1 \mid X) - \ell(T_2 \mid X)\right]$
- Expectation under null: $\mathbb{E}\left[\delta(T_1, T_2 \mid X)\right] = 0$

More robust approach

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(T_1, T_2 \mid X) = 2\left[\ell(T_1 \mid X) - \ell(T_2 \mid X)\right]$
- Expectation under null: $\mathbb{E}\left[\delta(T_1, T_2 \mid X)\right] = 0$
- Slides modified from Mark Holder

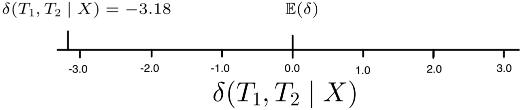
Using our SARS-CoV-2 example

- T_1 is ((Human, Bat), Pangolin) $\ell(T_1 \mid X) = -7363.296$



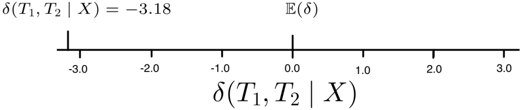
Using our SARS-CoV-2 example

- T_1 is ((Human, Bat), Pangolin) $\ell(T_1 \mid X) = -7363.296$
- T_2 is ((Human, Pangolin), Bat) $\ell(T_1 \mid X) = -7361.707$



Using our SARS-CoV-2 example

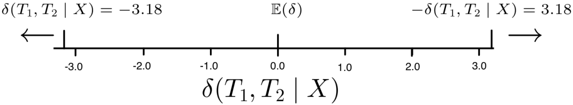
- T_1 is ((Human, Bat), Pangolin) $\ell(T_1 \mid X) = -7363.296$
- T_2 is ((Human, Pangolin), Bat) $\ell(T_1 \mid X) = -7361.707$
- $\delta(T_1, T_2|X) = -3.18$



Getting a p-value?

To get the *P*-value, we need to know the probability:

$$\Pr\left(\left|\delta(T_1, T_2 \mid X)\right| \geq 3.18 \middle| H_0 \text{ is true}\right)$$



Kishino-Hasegawa Test

- Examine the difference in ℓ for each site: $\delta(T_1, T_2 \mid X_i)$ for site i.

Kishino-Hasegawa Test

- Examine the difference in ℓ for each site: $\delta(T_1, T_2 \mid X_i)$ for site i.
- Note that the total difference is simply a sum:

$$\delta(T_1, T_2 \mid X) = \sum_{i=1}^{M} \delta(T_1, T_2 \mid X_i)$$

Kishino-Hasegawa Test

- Examine the difference in ℓ for each site: $\delta(T_1, T_2 \mid X_i)$ for site i.
- Note that the total difference is simply a sum:

$$\delta(T_1, T_2 \mid X) = \sum_{i=1}^{M} \delta(T_1, T_2 \mid X_i)$$

- The variance of $\delta(T_1, T_2 \mid X)$ will be a function of the variance in "site" $\delta(T_1, T_2 \mid X_i)$ values.

KH Test - the variance of $\delta(T_1, T_2 \mid X)$

To approximate variance of $\delta(T_1, T_2 \mid X)$ under the null, we could:

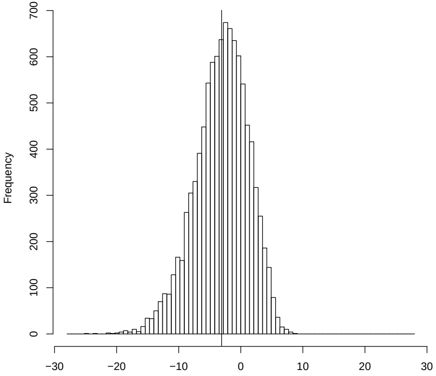
- Use assumptions of Normality (Central Limit Theorem)

KH Test - the variance of $\delta(T_1, T_2 \mid X)$

To approximate variance of $\delta(T_1, T_2 \mid X)$ under the null, we could:

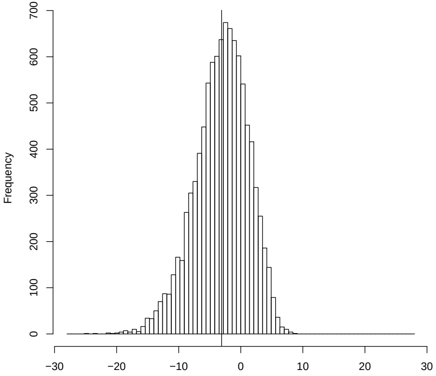
- Use assumptions of Normality (Central Limit Theorem)
- Use bootstrapping to generate a cloud of pseudo-replicate $\delta(T_1, T_2 \mid X^*)$ values, and look at their variance

Many RELL bootstraps later:



$$\delta(T_1, T_2 \mid X^*)$$

Many RELL bootstraps later:



$$\delta(T_1, T_2 \mid X^*)$$

- Is this our null?

KH Test - ‘centering’

- H_0 gives us the expected value:

$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

KH Test - ‘centering’

- H_0 gives us the expected value:

$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

- Bootstrapping gives us a reasonable guess of the variance under H_0

KH Test - ‘centering’

- H_0 gives us the expected value:

$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

- Bootstrapping gives us a reasonable guess of the variance under H_0
- By subtracting the mean of the bootstrapped $\delta(T_1, T_2 \mid X^*)$ values, we can create a null distribution.

KH Test - ‘centering’

- H_0 gives us the expected value:

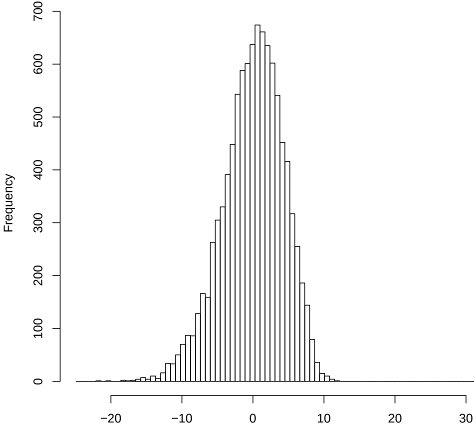
$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

- Bootstrapping gives us a reasonable guess of the variance under H_0
- By subtracting the mean of the bootstrapped $\delta(T_1, T_2 \mid X^*)$ values, we can create a null distribution.
- For each of the j bootstrap replicates, we treat:

$$\delta(T_1, T_2 \mid X^{*j}) - \bar{\delta}(T_1, T_2 \mid X^*)$$

as draws from the null distribution.

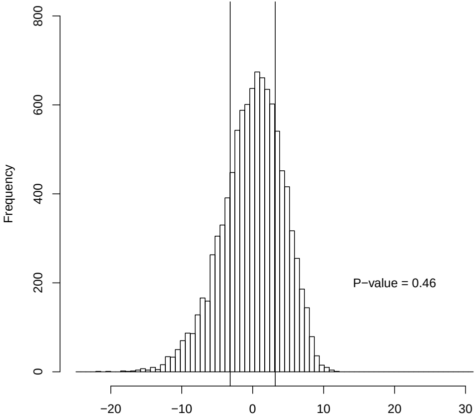
A null at last



$$\delta(T_1, T_2 \mid X^{(j)}) - \bar{\delta}(T_1, T_2 \mid X^*)$$

A null at last

Approximate null distribution with tails (absolute value ≥ 3.18) shown:



Downsides of the KH-test

- **Multiple Comparisons**: lots of trees increases the variance of $\delta(\hat{T}, T_1 \mid X)$

Downsides of the KH-test

- **Multiple Comparisons**: lots of trees increases the variance of $\delta(\hat{T}, T_1 \mid X)$
- **Selection bias**: Picking the ML tree to serve as one of the hypotheses invalidates the centering procedure of the KH test

Downsides of the KH-test

- **Multiple Comparisons**: lots of trees increases the variance of $\delta(\hat{T}, T_1 \mid X)$
- **Selection bias**: Picking the ML tree to serve as one of the hypotheses invalidates the centering procedure of the KH test
- $\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right]! = 0$ if choosing best vs rest

Alternative tests

- Shimodaira-Hasegawa Test:

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets

Alternative tests

- Shimodaira-Hasegawa Test:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**:

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**:
- Achieves weighting by varying bootstrap size for each tree.

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**:
- Achieves weighting by varying bootstrap size for each tree.
- Better for larger comparisons, but slow.

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**:
- Achieves weighting by varying bootstrap size for each tree.
- Better for larger comparisons, but slow.
- **Swofford–Olsen–Waddell–Hillis**:

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**:
- Achieves weighting by varying bootstrap size for each tree.
- Better for larger comparisons, but slow.
- **Swofford–Olsen–Waddell–Hillis**:
- Same idea but uses parametric bootstraps instead.

Alternative tests

- **Shimodaira-Hasegawa Test**:
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**:
- Achieves weighting by varying bootstrap size for each tree.
- Better for larger comparisons, but slow.
- **Swofford–Olsen–Waddell–Hillis**:
- Same idea but uses parametric bootstraps instead.
- Sensitive to model misspecification

Conclusion

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity
    - Matters little for some things and lots for others

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity
    - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity
    - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches
    - Likelihood tests offer a quicker alternative

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
  - Higehst likelihood model penalised by complexity
  - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches
  - Likelihood tests offer a quicker alternative
  - Branch support is built into Bayesian inferences

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
  - Higehst likelihood model penalised by complexity
  - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches
  - Likelihood tests offer a quicker alternative
  - Branch support is built into Bayesian inferences
  - Use multiple lines of evidence

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity
    - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches
    - Likelihood tests offer a quicker alternative
    - Branch support is built into Bayesian inferences
    - Use multiple lines of evidence
- You can compare trees statistically, but:

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity
    - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches
    - Likelihood tests offer a quicker alternative
    - Branch support is built into Bayesian inferences
    - Use multiple lines of evidence
- You can compare trees statistically, but:
    - Tree-space makes things non-trivial

Summary

- Phylogenies can be wrong for many reasons: bad data, sampling, pesky evolution
- Standard model selection principles apply to selection of sequence evolution models
    - Higehst likelihood model penalised by complexity
    - Matters little for some things and lots for others
- Bootstrapping in its various guises is a good way to estimate robustness of your branches
    - Likelihood tests offer a quicker alternative
    - Branch support is built into Bayesian inferences
    - Use multiple lines of evidence
- You can compare trees statistically, but:
    - Tree-space makes things non-trivial
    - Computationally demanding and easily biased

Questions?

References i

- Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. (2019).
  Model selection may not be a mandatory step for phylogeny reconstruction.
  Nature communications, 10(1):1–11.
- Barton, N. H. (2007).
  ***Evolution*.**
  Cold Spring Harbor Laboratory Press.
- Dimmic, M. W., Rest, J. S., Mindell, D. P., and Goldstein, R. A. (2002).
  rtrev: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny.
  Journal of molecular evolution, 55(1):65.

References ii

- Leonard, G. (2010).
  Development of fusion and duplication finder blast (fdfblast): a systematic tool to detect differentially distributed gene fusions and resolve trifurcations in the tree of life.
- Marwick, B. (2012).
  A cladistic evaluation of ancient thai bronze buddha images: six tests for a phylogenetic signal in the griswold collection.
  Connecting empires, pages 159–176.
- Richards, T. A., Soanes, D. M., Foster, P. G., Leonard, G., Thornton, C. R., and Talbot, N. J. (2009).
  Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi.
  The Plant Cell, 21(7):1897–1911.

References iii

Zhang, J., Jia, W., Zhu, J., Li, B., Xing, J., Liao, M., and Qi, W. (2020).
Insights into the cross-species evolution of 2019 novel coronavirus.
Journal of Infection.