Hypothesis testing and phylogenetics

Woods Hole Workshop on Molecular Evolution, 2017

Mark T. Holder
University of Kansas

Thanks to Paul Lewis, Joe Felsenstein, and Peter Beerli for slides.
Motivation

Pipeline of:

Data collection $\rightarrow$ Alignment $\rightarrow$ Model selection $\rightarrow$ Tree estimation
gives us a point estimate of the best tree.

This lecture/lab covers:

1. How confident should we be that this tree is correct?
2. Can we be confident in some aspects of the tree (e.g. some clades)?
3. If our question of interest is about evolution of characters, can we answer that question despite our uncertainty about the tree?
Conclusions 1 - confidence on trees

1. Non-parametric bootstrapping: useful for assessing sampling error, but a little hard to interpret precisely.
   - Susko’s aBP gives $1 - aBP \approx P$-value for the hypothesis that a recovered branch is not present in the true tree.
2. “How should we assign a $P$-value to tree hypothesis?” is surprisingly complicated.
   - Kishino-Hasegawa (KH-Test) if testing 2 (a priori) trees.
   - Shimodaira’s approximately unbiased (AU-Test) for sets of trees.
   - Parametric bootstrapping (can simulate under complex models)
Conclusions 2 - confidence about evo. hypotheses

If $H_0$ is about the evolution of a trait:

1. $P$-value must consider uncertainty of the tree:
   - can be large $P$ over confidence set of trees.
   - Bayesian methods (covered tomorrow) enable prior predictive or posterior predictive $P$-values.
Conclusions 3 - simulate your own null distributions

(the focus of the lab)

1. In phylogenetics we often have to simulate data to approximate $P$-values
2. Designing the simulations requires care to make a convincing argument.
Reasons phylogenetic inference might be wrong

1. **Systematic error** – Our inference method might not be sophisticated enough
2. **Random error** – We might not have enough data – we are misled by sampling error.
The bootstrap

(unknown) true value of $\theta$

(unknown) true distribution

Estimate of $\theta$

Empirical distribution of sample
The bootstrap

(unknown) true value of \( \theta \)

(unknown) true distribution

\[
\text{(unknown) true distribution}
\]

\[
\text{empirical distribution of sample}
\]

\[
\text{Distribution of estimates of parameters}
\]

\[
\text{Bootstrap replicates}
\]

\[
\text{estimate of } \theta
\]

Slide from Joe Felsenstein
The bootstrap for phylogenies

Original Data

Bootstrap sample #1

Bootstrap sample #2

Sample same number of sites, with replacement (and so on)

Slide from Joe Felsenstein
The majority-rule consensus tree

Trees:

How many times each partition of species is found:

<table>
<thead>
<tr>
<th>Partition</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>BCDF</td>
</tr>
<tr>
<td>ACE</td>
<td>BDF</td>
</tr>
<tr>
<td>ACEF</td>
<td>BD</td>
</tr>
<tr>
<td>AC</td>
<td>BDEF</td>
</tr>
<tr>
<td>AEF</td>
<td>BCD</td>
</tr>
<tr>
<td>ADEF</td>
<td>BC</td>
</tr>
<tr>
<td>ABCE</td>
<td>DF</td>
</tr>
</tbody>
</table>

Slide from Joe Felsenstein
From Hasegawa's analysis of 232 sites D-loop
http://phylo.bio.ku.edu/mephytis/boot-sample.html
http://phylo.bio.ku.edu/mephytis/parsimony.html
http://phylo.bio.ku.edu/mephytis/bootstrap.html
Bootstrapping for branch support

- Typically a few hundred bootstrap, pseudoreplicate datasets are produced.

- Less thorough searching is faster, but will usually artificially lower bootstrap proportions (BP). However, Anisimova et al. (2011) report that RAxML’s rapid bootstrap algorithm may inflate BP.

- “Rogue” taxa can lower support for many splits – you do not have to use the majority-rule consensus tree to summarize bootstrap confidence statements; See also (Lemoine et al., 2017)
Bootstrap proportions have been characterized as providing:

• a measure of repeatability,
• an estimate of the probability that the tree is correct (and bootstrapping has been criticized as being too conservative in this context),
• the P-value for a tree or clade
Frequentist hypothesis testing: coin flipping example

\( N = 100 \) and \( h = 60 \)

Can we reject the fair coin hypothesis? \( H_0 : \Pr(\text{heads}) = 0.5 \)

The “recipe” is:

1. Formulate null (\( H_0 \)) and alternative (\( H_A \)) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic: \( f_H = \) fraction of heads in sample.
   \( f_H = 0.6 \)
4. Characterize the null distribution of the test statistic
5. Calculate the \( P \)-value: The probability of a test statistic value more extreme than \( f_H \) arising even if \( H_0 \) is true.
6. Reject \( H_0 \) if \( P \)-value is \( \leq \) your Type I error rate.
Null distribution
$P$-value $\approx 0.058$
Making similar plots for tree inference is hard.

- Our parameter space is trees and branch lengths.
- Our data is a matrix of characters.
- It is hard to put these objects in the same space. You can do this “pattern frequency space”.

Some cartoons of projections of this space are posted at:
coin flipping

\[ N = 100 \text{ and } H = 60 \]

Can we reject the hypothesis of a fair coin?

We can use simulation to generate the null distribution (we could actually use the binomial distribution to analytically solve this one)...
A simulation of the null distribution of the # heads

P-value $\approx 0.029$
We discussed how bootstrapping gives us a sense of the variability of our estimate.

It can also give a tail probability for $\Pr(f^{(\text{boot})}_H \leq 0.5)$.

Amazingly (for many applications):

$$\Pr(\hat{f}_H \geq 0.6 \mid \text{null is true}) \approx \Pr(f^{(\text{boot})}_H \leq 0.5)$$

In other words, the $P$-value is approximate by the fraction of bootstrap replicates consistent with the null.
Distribution of the # heads in bootstrap resampled datasets

\[ \Pr(p^{(boot)} \leq 0.5) \approx 0.027 \]

\[ BP = \text{frac. of } p^{(boot)} \geq 0.5 \]

\[ BP \approx 0.973 \]

\[ P\text{-value} \approx 1 - BP \]
A simulation of the null distribution of the # heads

Distribution of the # heads in bootstrap resampled datasets
• When you decide between trees, the boundaries between tree hypotheses can be curved

• When the boundary of the hypothesis space is curved, $1 - BP$ can be a poor approximation of the $P$-value. – Efron et al. (1996)
In the straight border case, symmetry implies that:

The actual $P$-value (blue region) $\approx 1 - BP$

$(1 - BP$ is the blue below)
In the curved border case, the symmetry breaks down:

The actual $P$-value (blue region) $\neq 1 - BP$

$1 - BP$ is the blue below
• Efron et al. (1996) proposed a computationally expensive multi-level bootstrap (which has not been widely used).

• Shimodaira (2002) used the same theoretical framework to devise a (more feasible) Approximately Unbiased (AU) test of topologies.
  – Multiple scales of bootstrap resampling (80% of characters, 90%, 100%, 110%...) are used to detect and correct for curvature of the boundary.
  – Implemented in the new versions of PAUP*
Susko (2010) adjusted BP – aBP

- Susko agrees with curvature arguments of Efron et al. (1996) and Shimodaira (2002), but points out that they ignore the sharp point in parameter space around the polytomy.
- He correct bootstrap proportions: $1 - aBP$ accurately estimates the $P$-value.
- The method uses the multivariate normal distributions based on calculations about the curvature of the likelihood surface.
- You need to perform a different correction when you know the candidate tree a priori versus when you are putting BP on the ML tree.
- BP may not be conservative when you correct for selection bias.
aBP for each BP (5 model conditions)

aBP with selection bias correction for each BP (5 model conditions)
Conclusions – bootstrapping

1. Non-parametric bootstrapping proportions help us see which branches have so little support that they could be plausibly explained by sampling error.
2. BPs are a little hard to interpret precisely.
3. Susko has and adjustment ("aBP") so that $1 - aBP \approx P$-value for the hypothesis that a recovered branch is not present in the true tree.
Can we test trees using the LRT?

1. Should we calculate the LRT as:
   \[
   \delta_i = 2 \left[ \ln L(t = \hat{t}, T_i \mid X) - \ln L(t = 0, T_i \mid X) \right]
   \]

2. And can we use the $\chi_1^2$ distribution to get the critical value for $\delta$?
1. Should we calculate the LRT as:
\[ \delta_i = 2 \left[ \ln L(t = \hat{t}, T_i | X) - \ln L(t = 0, T_i | X) \right] \]

No. \( t = 0 \) might not yield the best alternative \( \ln L \)

2. And can we use the \( \chi^2 \) distribution to get the critical value for \( \delta \)?

No. Constraining parameters at boundaries leads to a mixture such as: \[ \frac{1}{2} \chi^2_0 + \frac{1}{2} \chi^2_1 \]

See Ota et al. (2000).
Can we test trees using the LRT?

No, tree hypotheses are not nested!
aLRT of Anisimova and Gascuel (2006)

- For a **branch** $j$, calculate $\delta_j^{\dagger}$ as twice the difference in $\ln L$ between the optimal tree (which has the branch) and the best NNI neighbor.
- This is very fast.
- They argue that the null distribution for each LRT around the polytomy follows a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution.
- They introduce Bonferroni-correction appropriate for correcting for the selection of the best of the three resolutions.
- They find aLRT to be accurate and powerful in simulations, but Anisimova et al. (2011) report that it rejects too often and is sensitive to model violation.
\[ aLRT = 2 \left[ \ln \ell_1 - \ln L(T_2 \mid X) \right] \]
\[ \ell_1 = L(T_1 \mid X) \]

Image from Anisimova and Gascuel (2006)
Simulation studies of Anisimova et al. (2011) show it to have the best power of the methods that do not have inflated probability of falsely rejecting the null.

It is sensitive to model violation.

This is similar to “likelihood-mapping” of Strimmer and von Haeseler (1997)
coin flipping example (again, for inspiration)

\[ N = 100 \text{ and } H = 60 \]

Can we reject the hypothesis of a fair coin?

We can use simulation to generate the null distribution (we could actually use the binomial distribution to analytically solve this one)...
A simulation of the null distribution of the # heads

P-value $\approx 0.029$
The simplest phylogenetic test would compare two trees

Null: If we had no sampling error (infinite data) $T_1$ and $T_2$ would explain the data equally well.

Test Statistic:

$$\delta(T_1, T_2 \mid X) = 2 \left[ \ln L(T_1 \mid X) - \ln L(T_2 \mid X) \right]$$

Expectation under null:

$$\mathbb{E}_{H_0} \left[ \delta(T_1, T_2 \mid X) \right] = 0$$
Using 3000 sites of mtDNA sequence for 5 primates

\[ T_1 \text{ is } ((\text{chimp, gorilla}), \text{human}) \]
Total \( \ln L(T_1) - \ln L(T_2) = -1.58134 \)
Scatterplot of site lnL on 2 trees.

Total lnL(T1) – lnL(T2) = -1.58134
Using 3000 sites of mtDNA sequence for 5 primates

$T_1$ is ((chimp, gorilla), human) \quad \ln L(T_1 | X) = -7363.296

$T_2$ is ((chimp, human), gorilla) \quad \ln L(T_2 | X) = -7361.707

$\delta(T_1, T_2 | X) = -3.18$
To get the $P$-value, we need to know the probability:

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

$$\delta(T_1, T_2 \mid X) = -3.18 \quad \mathbb{E}(\delta) \quad -\delta(T_1, T_2 \mid X) = 3.18$$
1. Examine the difference in $\ln L$ for each site: $\delta(T_1, T_2 \mid X_i)$ for site $i$.

2. Note that the total difference is simply a sum:

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

3. 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.
$\delta(T_1, T_2 \mid X_i)$ for each site, $i$. 
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:

1. use assumptions of Normality (by appealing to the Central Limit Theorem\(^1\)). Or
2. use bootstrapping to generate a cloud of pseudo-replicate $\delta(T_1, T_2 \mid X^*)$ values, and look at their variance.

\(^1\)Susko (2014) recently showed that this is flawed and too conservative.
\( \delta \) for many (RELL) bootstrapped replicates of the data
The (RELL) bootstrapped sample of statistics. Is this the null distribution for our $\delta$ test statistic?

$\delta(T_1, T_2 \mid X^*)$
Bootstrapping gives us a reasonable guess of the variance under $H_0$

Subtracting the mean of the bootstrapped $\delta(T_1, T_2 \mid X^*)$ values gives the 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.
\[ \delta(T_1, T_2 \mid X^{(j)}) - \bar{\delta}(T_1, T_2 \mid X^*) \]

for many (RELL) bootstrapped replicates of the data
Approximate null distribution with tails (absolute value $\geq 3.18$) shown

$$\delta(T_1, T_2 \mid X^*) - \bar{\delta}(T_1, T_2 \mid X^*)$$
Other ways to assess the null distribution of the LR test statistic

- Bootstrapping then centering LR, and
- Using normality assumptions.

are both clever and cute solutions.

They are too conservative (Susko, 2014) - more complicated calculations from the Normal [KHns] or mixtures of $\chi^2$ distributions [chi-bar].

They do not match the null distribution under any model of sequence evolution.
Mini-summary

• $\delta(T_1, T_2 \mid X) = 2 [\ln L(T_1 \mid X) - \ln L(T_2 \mid X)]$ is a powerful statistic for discrimination between trees.

• We can assess confidence by considering the variance in signal between different characters.

• Bootstrapping helps us assess the variance in $\ln L$ that we would expect to result from sampling error.
**Scenario**

1. A (presumably evil) competing lab scoops you by publishing a tree, $T_1$, for your favorite group of organisms.

2. You have just collected a new dataset for the group, and your ML estimate of the best tree, $T_2$, differ’s from $T_1$.

3. A KH Test shows that your data **significantly** prefer $T_2$ over $T_1$.

4. You write a (presumably scathing) response article.

Should a *Systematic Biology* publish your response?
What if start out with only one hypothesized tree, and we want to compare it to the ML tree?

The KH Test is **NOT** appropriate in this context (see Goldman et al., 2000, for discussion of this point)

**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.
Using the ML tree in your test introduces selection bias

Even when the $H_0$ is true, we do not expect

$$2 \left[ \ln L(\hat{T}) - \ln L(T_1) \right] = 0$$

Imagine a competition in which a large number of equally skilled people compete, and you compare the score of one competitor against the highest scorer.
Experiment: 70 people each flip a fair coin 100 times and count # heads.

\[ h_1 - h_2 \]
Experiment: 70 people each flip a fair coin 100 times and count # heads.

\[ h_1 - h_2 \]

Null dist.: difference in # heads any two competitors

\[ \max(h) - h_1 \]

Null dist.: difference highest – random competitor

<table>
<thead>
<tr>
<th>Diff # heads</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-30</td>
<td>0</td>
</tr>
<tr>
<td>-20</td>
<td>200</td>
</tr>
<tr>
<td>-10</td>
<td>400</td>
</tr>
<tr>
<td>0</td>
<td>600</td>
</tr>
<tr>
<td>10</td>
<td>800</td>
</tr>
<tr>
<td>20</td>
<td>1000</td>
</tr>
<tr>
<td>30</td>
<td>1200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diff # heads</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-30</td>
<td>500</td>
</tr>
<tr>
<td>-20</td>
<td>1000</td>
</tr>
<tr>
<td>-10</td>
<td>1500</td>
</tr>
<tr>
<td>0</td>
<td>2000</td>
</tr>
<tr>
<td>10</td>
<td>2500</td>
</tr>
<tr>
<td>20</td>
<td>3000</td>
</tr>
<tr>
<td>30</td>
<td>3500</td>
</tr>
</tbody>
</table>
Shimodaira and Hasegawa proposed the SH test which deals the “selection bias” introduced by using the ML tree in your test.

Requires set of candidate trees - these must not depend on the dataset to be analyzed.

$H_0$: each tree in the candidate set is as good as the other trees.

The test makes worst-case assumptions - it is conservative.

AU test is less conservative (still needs a candidate set)
real data → test stat commands → real test stat

null fitting commands

fit null

simulated data
real data → test stat commands → real test stat

null fitting commands

fit null

simulated data → null distrib.

null distrib.
real data \rightarrow \text{test stat commands} \rightarrow \text{real test stat}

\text{null fitting commands}

\text{simulated data} \rightarrow \text{null distrib.} \rightarrow P\text{-value}

\text{fit null}
Parametric bootstrapping to generate the null distribution for the LR statistic

1. find the best tree and model pair that are consistent with the null,

2. Simulate many datasets under the parameters of that model,

3. Calculate \( \delta^{(j)} = 2 \left[ \ln L(\hat{T}^{(j)} | X^{(j)}) - \ln L(\hat{T}_0^{(j)} | X^{(j)}) \right] \)
   for each simulated dataset.

   • the \((j)\) is just an index for the simulated dataset,
   • \(\hat{T}_0^{(j)}\) is the tree under the null hypothesis for simulation replicate \(j\)
Parametric bootstrapping

This procedure is often referred to as SOWH test (in that form, the null tree is specified *a priori*).

Huelsenbeck et al. (1996) describes how to use the approach as a test for monophyly.

Intuitive and powerful, but not robust to model violation (Buckley, 2002).

Can be done manually\textsuperscript{2} or via SOWHAT by Church et al. (2015). Optional demo [here](https://molevol.mbl.edu/index.php/ParametricBootstrappingLab).

Susko (2014): collapse optimize null tree with 0-length contraints for the branch in question (to avoid rejecting too often)

\textsuperscript{2}instructions in [https://molevol.mbl.edu/index.php/ParametricBootstrappingLab](https://molevol.mbl.edu/index.php/ParametricBootstrappingLab)
Null distribution of the difference in number of steps under GTR+I+G

<table>
<thead>
<tr>
<th>Difference in steps</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>−15</td>
<td></td>
</tr>
<tr>
<td>−10</td>
<td></td>
</tr>
<tr>
<td>−5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td></td>
</tr>
</tbody>
</table>
Null distribution of the difference in number of steps under JC

Difference in steps

Frequency

-15  -10  -50
0  200  400  600  800  1000
We often don’t want to test tree topologies

- If we are conducting a “comparative method” we have to consider phylogenetic history,

- ideally we would integrate out the uncertainty in the phylogeny

Tree is a “nuisance parameter”
Suppose we are using some positive test statistic, \( s \).

If we observe a value of \( s = X \) on our data:

Remember that \( P = \Pr(s \geq X \mid H_0) \), which is usually described as the tail probability.

\[
P = \int_X^{\infty} f(s \mid H_0) \, ds
\]

But what if the probability density of \( s \) depends on a nuisance parameter \( T \) and we don’t know the value of \( T \)?
P-values with nuisance parameters

\[ P = \int_{X}^{\infty} f(s \mid H_0) \, ds \]

We could:

• take the max \( P \) over all \( T \) – this is way too conservative

• make a confidence \((1 - \beta)\%\) set of \( T \) and take \( P \) to be \( \beta + \) the largest \( P \) in that set (Berger and Boos method)

• do a Bayesian analysis and get a posterior predictive \( P \) value:

\[ P = \int_{X}^{\infty} \left( \int f(s \mid T, H_0)p(T) \, dT \right) \, ds \]
rate CI → char. evo. rate → sim data → (usually bigger) \( P \)-value

tree topology

dge lengths
rate CI  \rightarrow \text{char. evo. rate} \rightarrow \text{sim data} \rightarrow \text{(usually biggest) } P\text{-value}

tree CI  \rightarrow \text{tree topology}

edge CI  \rightarrow \text{edge lengths}
Significantly different genealogy $\neq$ different phylogeny

- True “gene tree” can differ from true “species tree” for several biological reasons:
  - deep coalescence,
  - gene duplication/loss (you may be comparing paralogs),
  - lateral gene transfer.
There are lots of simulators out there

- Seq-gen, PAUP, Indelible ... - substitutions
- DAWG, Indelible ... - alignment
- ms and multi-species coalescent simulators for within population samples
- Ginkgo, DIM SUM... biogeography simulators,
- ...

Parametric bootstrapping is very flexible and lots of tools are available.
You have to think about what sources of error are most relevant for your data!
Errors modeling multiple hits
Assembly + asc. bias errors
Alignment errors
Paralogy errors
Errors from deep coalescence
No signal
Long ago
Time
Now
Errors from deep coalescence
Paralogy errors
Alignment errors
Assembly + asc. bias errors
Errors modeling multiple hits
Horizontal gene transfer
No signal
Now
Long ago
Time
Tree and parameter sensitivity interact

From ?

<table>
<thead>
<tr>
<th>Tree</th>
<th>Origins of viviparity</th>
<th>1X</th>
<th>2X</th>
<th>3X</th>
<th>4X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyron and Burbrink tree</td>
<td>106–129</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ExaML tree</td>
<td>103–126</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Branch-length improved tree</td>
<td>103–124</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Best tree</td>
<td>98–129</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Conclusions 1 - confidence on trees

1. Non-parametric bootstrapping: useful for assessing sampling error, but a little hard to interpret precisely.
   - Susko’s aBP gives $1 - aBP \approx P$-value for the hypothesis that a recovered branch is not present in the true tree.

2. “How should we assign a $P$-value to tree hypothesis?” is surprisingly complicated.
   - Kishino-Hasegawa (KH-Test) if testing 2 (a priori) trees.
   - Shimodaira’s approximately unbiased (AU-Test) for sets of trees.
   - Parametric bootstrapping (can simulate under complex models)
Conclusions 2 - confidence about evo. hypotheses

If $H_0$ is about the evolution of a trait:

1. $P$-value must consider uncertainty of the tree:
   - can be large $P$ over confidence set of trees.
   - Bayesian methods (covered tomorrow) enable prior predictive or posterior predictive $P$-values.
Conclusions 3 - simulate your own null distributions

(the focus of the lab)

1. In phylogenetics we often have to simulate data to approximate $P$-values
2. Designing the simulations requires care to make a convincing argument.


