Testing Hypotheses: Topology Comparisons

Last updated 04/09/12
Measures of Support (continued)

- Non-parametric bootstrap
- Bremer support
- Bayesian posterior probabilities
- Likelihood ratio tests (LRT)
  - Cannot be used because hypotheses (i.e., trees) are not nested
- Approximate LRT (implemented in PhyML)
- Topology comparison tests
  - Paired-sites tests:
    - Parsimony
    - Likelihood
  - Parametric bootstrap (SOWH test)
Measures of Support (continued)

• Non-parametric bootstrap
• Bremer support
• Bayesian posterior probabilities
• Likelihood ratio tests
  – Cannot be used because hypotheses (i.e., trees) are not nested
• Topology comparison tests
  – Parametric bootstrap (SOWH test)
  – Paired-sites tests
Paired-sites tests

• The basic question:
  – Is tree A significantly better than tree B or are the differences (in parsimony or likelihood scores) within the expectations of random error?

1. Compare two trees for either parsimony or likelihood scores
   • Number of steps or likelihood score at each site

2. Decide whether the difference is significant by comparing to an assumed distribution (binomial or normal), or creating a null distribution using bootstrap sampling techniques
Paired-sites tests: parsimony

• Developed by Alan Templeton (originally for restriction site data; 1983)
• Winning sites test (Prager and Wilson 1988) PAUP*
  – a simpler version of Templeton
  – for each site, score which tree is better so that each site is assigned either + or - (or 0 if they are equivalent). Use a binomial distribution to test whether the fraction of + versus - is significantly different from 0.5
• Wilcoxon signed-ranks test (Templeton 1983) PAUP*
  – Replace absolute values of differences by their ranks, then re-applies signs
• Kishino-Hasegawa test (1989) PAUP*
  – Assume all sites are i.i.d
  – Test statistic $T = \Sigma T_{(i)}$
    • Where $T_{(i)}$ is the difference in the minimum number of substitutions on the two trees at the $i$th informative site.
    • Expectation for $T$ (under the null hypothesis that the two trees are not significantly different) is zero. The sample variance can be obtained with an equation and tested with a $t$-test with $n-1$ degrees of freedom ($n$ = number of informative sites)
  – If no a priori reason to suspect one tree better than other (two-tailed)
  – If a priori reason (such as one tree being the most parsimonious), test is not valid
Paired-sites tests: **parsimony**

Example

- One dataset
- Two trees:
  - Tree 1: 1153 steps (Better)
  - Tree 2: 1279 steps (Worse)
- Question: Are the two trees significantly different?
Winning sites: PAUP* output

Comparison of tree 1 (best) to tree 2:

<table>
<thead>
<tr>
<th>Character</th>
<th>Tree 1</th>
<th>Tree 2</th>
<th>Difference</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>888</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>890</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
</tbody>
</table>

Totals

Winning-sites (sign) test:

- N = 166
- Number of positive differences = 144
- Number of negative differences = 22
- Test statistic = 0.867470
- P < 0.0001* (normal approximation with z = 9.391421)
Wilcoxon signed-ranks (Templeton) test: PAUP* output

Comparison of tree 1 (best) to tree 2:

<table>
<thead>
<tr>
<th>Character</th>
<th>Tree 1</th>
<th>Tree 2</th>
<th>Difference</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>888</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>890</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td></td>
<td>-81.5</td>
</tr>
</tbody>
</table>

Totals 126 12068 -1793

Templeton (Wilcoxon signed-ranks) test:

N = 162
Test statistic (T) = 1467
P < 0.0001* (normal approximation with z = -9.899495)
Kishino-Hasegawa test: PAUP* output

<table>
<thead>
<tr>
<th>Tree</th>
<th>Length</th>
<th>Length diff</th>
<th>s.d.(diff)</th>
<th>t</th>
<th>P*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1153</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1279</td>
<td>126</td>
<td>12.02005</td>
<td>10.4825</td>
<td>&lt;0.0001*</td>
</tr>
</tbody>
</table>

* Probability of getting a more extreme T-value under the null hypothesis of no difference between the two trees (two-tailed test). Asterisked values in table (if any) indicate significant difference at P < 0.05.

Remember: this test is only valid if there is no a priori reason to believe that one tree is better than the other.
Paired-sites tests: likelihood

- Winning sites test (Prager and Wilson 1988)
  - Similar to the parsimony approach, but using likelihood scores for each site
- $z$ test (and the practically identical $t$ test) PAUP* (KH test with normal approximation)
  - If the likelihood difference between the two trees is >1.96 times its standard error, the two trees are significantly different
- Wilcoxon signed-ranks test (Templeton 1983)
- Kishino-Hasegawa (Kishino & Hasegawa 1989)
  - Normal approximation (see $z$ test above)
  - Using bootstrap resampling to create a null distribution
    - RELL approximation PAUP* (KH test RELL)
    - Full optimization PAUP* (KH test Full)
- Shimodaira-Hasegawa PAUP* (RELL & Full optimization)
- Approximately unbiased AU (Shimodaira 2002) CONSEL
  - Similar to SH but uses multiscale bootstrap to correct for the bias caused by standard bootstrap resampling
Example

Is the ML tree (right) significantly better than the accepted phylogeny (left)?

\[ \ln L_1 = -5073.75 \]

\[ \ln L_{ML} = -5069.85 \]

See Goldman et al. 2000
Example: KH bootstrap test

Observed difference in lnL = 3.9 (P = 0.384)

See Goldman et al. 2000
Example: frequency distribution of log-likelihood differences

KH bootstrap histogram
P = 0.384

KH Normal approximation
P = 0.384
(var = no. of sites x variance of site lnLs)
RELL approximation vs. Full optimization

- In all paired sites tests, the sum of the likelihoods for each site is compared among the different trees

- **Full optimization** (option in PAUP*):
  - For each bootstrap replicate, branch lengths are optimized for each tree to be compared

- **RELL (Resampling Estimated Log Likelihoods) approximation** (option in PAUP*):
  - For each bootstrap replicate, assume the same branch lengths obtained from the original data
  - Keeps track of likelihoods at each individual sites and then adds them up according to the sites that were resampled (much faster)
The Kishino-Hasegawa (KH) test is only valid if the two compared trees are specified before-hand.

However, the common practice is to estimate the ML tree from dataset and test alternative trees (from other authors/datasets) against the ML tree.

Also, comparing more than two trees leads to a multiple comparison problem that cannot be solved by Bonferroni corrections.

Use of KH test in these cases will lead to false rejections of the non-ML trees (selection bias) pointed out by Shimodaira & Hasegawa (1999) and Goldman et al. (2000).

Selection bias leads to overconfidence in the wrong trees.

The Shimodaira-Hasegawa (SH) corrects for this selection bias and accounts for the multiple comparisons issue, but is very conservative.

Test constructed under the least favorable scenario:

- Null: that all hypotheses are equivalent
- For example: if we have 10 reasonable trees to evaluate, but we add another 90 implausible trees to the comparison, they make the test more conservative
- Thus, too many trees will dilute the power
Shimodaira-Hasegawa (SH) test (PAUP* and CONSEL)

- The conservative behavior is partially alleviated with the Weighted Shimodaria Hasegawa (WSH) CONSEL
  - The test statistic is standardized
- The null distribution against which the test statistic is compared is obtained by bootstrap re-sampling via Full optimization or RELL approximation
- However, bootstrap is biased (Zharkikh & Li 1992; Li & Zharkikh 1994; Hillis & Bull 1993; Felsenstein & Kishino 1993; Efron, Halloran & Holmes 1996; Newton 1996)
- Shimodaira (2002) introduced a modification of the bootstrap to reduce the bias when conducting the SH test
- Known as the Approximately Unbiased (AU) test (CONSEL)
- Uses multi-scale bootstrap to correct p-value estimates
  - A series of bootstraps of different sizes (still uses RELL approximation)
  - Fits curves through resulting p-values to obtain a correction formula
**KH and SH tests (RELL): PAUP* output**

<table>
<thead>
<tr>
<th>Tree</th>
<th>-ln L</th>
<th>Diff -ln L</th>
<th>KH-test P</th>
<th>SH-test P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5988.05924</td>
<td>(best)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6256.51841</td>
<td>268.45917</td>
<td>0.000*</td>
<td>0.000*</td>
</tr>
<tr>
<td>3</td>
<td>6263.25718</td>
<td>275.19794</td>
<td>0.000*</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

* P < 0.05

Time used to compute likelihoods = 0.22 sec

Kishino-Hasegawa and Shimodaira-Hasegawa tests:

KH test using **RELL bootstrap**, two-tailed test
SH test using **RELL bootstrap** (one-tailed test)
Number of bootstrap replicates = 1000
KH and SH tests (Full): PAUP* output

<table>
<thead>
<tr>
<th>Tree</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ln L</td>
<td>5988.05924</td>
<td>6256.51841</td>
<td>6263.25718</td>
</tr>
</tbody>
</table>

Time used to compute likelihoods = 0.22 sec

Kishino-Hasegawa and Shimodaira-Hasegawa tests:
  KH test using bootstrap with full optimization, two-tailed test
  SH test using bootstrap with full optimization (one-tailed test)

Number of bootstrap replicates = 1000

<table>
<thead>
<tr>
<th>Tree</th>
<th>-ln L</th>
<th>Diff -ln L</th>
<th>KH-test</th>
<th>SH-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5988.05924</td>
<td>(best)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6256.51841</td>
<td>268.45917</td>
<td>0.000*</td>
<td>0.000*</td>
</tr>
<tr>
<td>3</td>
<td>6263.25718</td>
<td>275.19794</td>
<td>0.000*</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

* P < 0.05
### KH, SH & AU tests (RELL): CONSEL output

**P-values for several tests comparing 3 topologies**

<table>
<thead>
<tr>
<th>rank</th>
<th>item</th>
<th>obs</th>
<th>au</th>
<th>np</th>
<th>bp</th>
<th>pp</th>
<th>kh</th>
<th>sh</th>
<th>wkh</th>
<th>wsh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-9.7</td>
<td>0.967</td>
<td>0.931</td>
<td>0.933</td>
<td>1.000</td>
<td>0.929</td>
<td>0.981</td>
<td>0.929</td>
<td>0.995</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>9.7</td>
<td>0.056</td>
<td>0.063</td>
<td>0.061</td>
<td>6e-005</td>
<td>0.071</td>
<td>0.344</td>
<td>0.071</td>
<td>0.123</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>46.0</td>
<td>0.007</td>
<td>0.006</td>
<td>0.007</td>
<td>1e-020</td>
<td>0.008</td>
<td>0.013</td>
<td>0.008</td>
<td>0.010</td>
</tr>
</tbody>
</table>
References

• Chapter 21 “Paired Sites tests” and pp. 346–352 in Inferring Phylogenies textbook.
Non-parametric bootstrap (review)

(from Felsenstein, 2004)
Parametric Bootstrap

(from Felsenstein, 2004)
Parametric Bootstrap

• Simulate the data under the most likely tree and model of evolution
• Infer the ML tree for each simulated replicate
• Compute a majority-rule consensus tree with $P$ values for branches (same as with non-parametric bootstrap)
• Concerns:
  – Close reliance on the correctness of statistical model evolution
  – When model is correct, variation from parametric bootstrap will be very similar to that of non-parametric bootstrap
• Alternative use (and much more widely used) to compare alternative topologies to the most likely tree (SOWH test or Parametric Bootstrap Test of Monophyly)
Parametric Bootstrap of Monophyly

• Most useful:
  – In many cases, the overall tree structure, rather than a particular branch may suggest that a null hypothesis is incorrect
  – If no single branch is particularly well supported, but the cumulative effects of many branches contain enough phylogenetic signal to reject a particular null hypothesis
ML tree

A
  C
  B
  O

Morphology tree (Hypothesis A or null)

A
  B
  C
  O
simulate evolution on this tree to generate $N$ datasets of the same size and base composition as the original dataset
Parametric bootstrap test of monophyly (SOWH test)

- For each of the simulated dataset:

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Likelihood score</th>
<th>ML analyses</th>
<th>Likelihood score</th>
<th>ML analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>no constraints</td>
<td>Hyp A (or null Hypothesis)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$x_0$</td>
<td>$x_A$</td>
<td>$\delta_1$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$x_0$</td>
<td>$x_A$</td>
<td>$\delta_2$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$x_0$</td>
<td>$x_A$</td>
<td>$\delta_3$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>$x_0$</td>
<td>$x_A$</td>
<td>$\delta_{100}$</td>
<td></td>
</tr>
</tbody>
</table>
Parametric Bootstrap of Monophyly

- Original data
  - Estimate of tree under null (constrained)
    - Dataset #1
    - Dataset #2
    - Dataset #2
    - Dataset #2
    - Dataset #100
    - ML estimate of tree (unconstrained)
Parametric bootstrap test of monophyly (SOWH test)
Parametric bootstrap test of monophyly (SOWH test)

(a) Difference in score

Observed difference is not significantly different from the values in the distribution.
Can’t reject hypothesis A

(b) Difference in score

Observed difference is significantly different from the values in the distribution (p < 0.01).
Reject hypothesis A
If hypothesis A were true, we could often recover alternative hypotheses in an analysis.

If hypothesis A were true, we would be unlikely to recover alternative hypotheses in an analysis... Thus, Hyp A is unlikely; at least with our dataset (and all the assumptions we made in the process)
Parametric Bootstrap of Monophyly

• It is a way of testing whether we have enough power with the data at hand to reject a hypothesis

• There are many shortcut options for the SOWH test (Goldman et al. 2000 discusses implications of some of these):
  – use parsimony instead of likelihood for comparison of trees
  – use fixed topologies instead of constrained (to the null or Hyp A) and unconstrained searches
  – partial vs. full optimization of parameters and branchlengths
Additional References


Parametric Bootstrap test of Monophyly (SOWH test)

- https://molevol.mbl.edu/wiki/index.php/ParametricBootstrappingLab
- http://www.ebi.ac.uk/goldman/tests/SOWHInstr.html