Lecture 15 – Hypothesis Testing

The development of explicitly statistical phylogenetic hypothesis testing has to rank up there with the development of PCR as causes of the renaissance that systematics has experienced over the last 25 years but the idea of testing hypotheses with phylogenies is an old one. Any evolutionary hypothesis that makes topological predictions can be the focus of phylogenetic hypothesis testing. These have traditionally been frequentist tests.

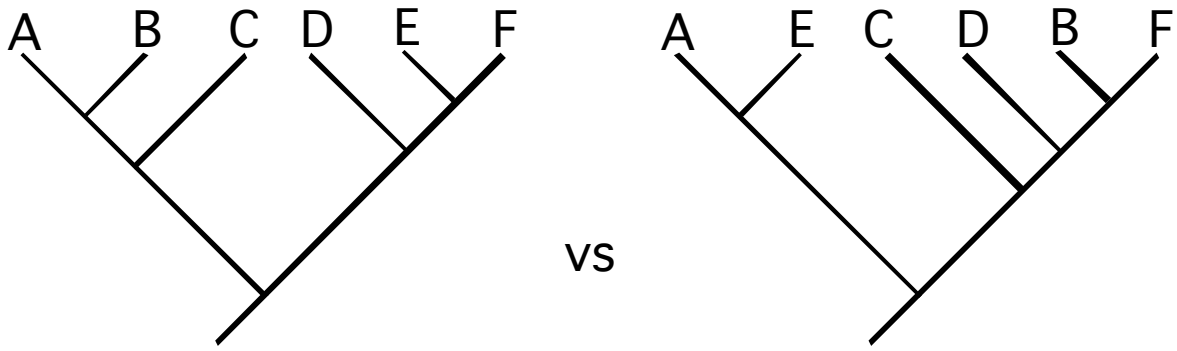
II. Tests of Topology – Probably the most common type of test involves testing topologies.

We may have a case where we're testing a single *a priori* hypothesis or we may have a pair of competing hypotheses. This latter case was dealt with first historically, so we'll do the same.

A. Tests of competing *a priori* hypotheses.

The early methods were based on comparisons of two trees that have been determined prior to collecting data.

These take the approach of asking "is the support for Tree A significantly different than the support for Tree B?"



Again, to be used correctly, these trees should be fully defined *a priori* based on some independent the data.

The **null hypothesis is that there is no difference in support for one tree over the other**.

So, it's simple enough to calculate the **test statistic**:

Under Parsimony (or minimum evolution), it's:

$$\delta = \textbf{Length}_A - \textbf{Length}_B$$

Under Likelihood, it's:

$$\delta = \ln L_A - \ln L_B$$

This is simply the difference in optimality score of the trees being compared. This is the difference in the sum across characters, which is the same as the sum of the pairwise differences.

There is a long history of studies on how to assess the magnitude of the test statistic. Rather than go through the entire literature, I'll point you to Goldman et al. (2000, *Syst. Biol.*, 49:652), who provide the best overview that I've seen of (most of) these older methods.

We'll focus, therefore, on common currently-used methods, starting with the test of Kishino and Hasegawa (i.e., the KH test; Kishino & Hasegawa 1989. J. Mol. Evol., 29:170), with a null distribution assumed to follow a normal distribution or with it estimated using the RELL bootstrap.

The **parametric version** of these tests works as follows:

The null expectation is that the difference in optimality score for the two trees is zero.

To generate a null distribution, early methods looked at the difference in optimality between the two trees on a **site-by-site basis.** Thus, these are called **paired-sites method.**

Character	Length		Difference
	Tree A	Tree B	
4	2	1	1
5	2	3	-1
6	1	2	-1
10	4	2	2
Total	9	8	1

Under the null hypothesis, one might expect the single-likelihood differences to be **normally distributed** across characters. That is, for any character, the expected difference between trees is 0. Shimodaira (2002) provides a justification for this based on *infinite data*.

Here's an example comparing two trees.

```
Tree              1           2
-ln L    2558.60898  2529.81009    [So the test statistic is 28.8 lnL units]

Kishino-Hasegawa test:
  KH test using normal approximation, two-tailed test

                                        KH-test
    Tree      -lnL          Diff -lnL     P
-----------------------------------------------
       1    2558.60898      28.79889   0.000*
       2    2529.81009        (best)
  * P < 0.05
```

Here, Tree 2 is better supported than Tree 1.

The original versions of the KH-test made this assumption of a normal distribution, but we can use the test without making that assumption by using a bootstrap resampling method called RELL. This simply generates bootstrap replicates by resampling the single-site likelihoods (Resampling Estimated Log Likelihoods). This method is really fast, because nothing is recalculated for the bootstrap replicates. The SSLs *for each tree* are stored and we resample them (more on this in a minute).

```
Tree              1           2
-ln L   2558.60898  2529.81009

Time used to compute likelihoods = 0.12 sec

Kishino-Hasegawa test:
  KH test using RELL bootstrap, two-tailed test
  Number of bootstrap replicates = 1000
                                          KH-test
    Tree    -ln L          Diff -ln L      P
-----------------------------------------------
       1    2558.60898      28.79889   0.396
       2    2529.81009        (best)
```

Interestingly, **the choice of assuming a normal distribution versus using RELL bootstrap alters the results a lot.** This is likely because of two things; the data are finite and the distribution is non-normal, with a higher variance that is captured by the bootstrap resampling process.

The KH-test is really only appropriate if the trees being compared are **chosen *a priori***.

However, **most applications** of these tests have involved **comparing a suboptimal tree to the best tree.**

This use violates all the null distributions we've talked about. It's still done, but several papers have pointed out the dangers of doing this.

Remember that the null hypothesis in the earlier tests is that there is no difference between the two trees.

However, if we are comparing a suboptimal tree to the best tree, the null distributions will be shifted.

Shimodaira and Hasegawa (1999; Mol. Biol. Evol., 16:1114) developed a test (**the SH Test**) that corrects for multiple comparisons and corrects for the *a priori* requirement.

The SH test (Shimodaira and Hasegawa. 1999. Mol. Biol. Evol., 16:1114) uses bootstrap resampling (typically RELL) to generate an average ln*L* for the **collection** of trees (n > 2) being considered.

It uses this average to center the null distribution and correct for the bias of finding the ML tree and comparing suboptimal trees to it.

Since this method relies on a collection of trees, the null hypothesis is a little different than above. Here, the null hypothesis is that all the trees in the set are equally good explanations of the data.

This leads the SH test to be very conservative.

Shimodaira (2002. Syst. Biol. 51:492) has applied multiscale bootstrapping to correct for this bias in producing the AU (almost unbiased) test and this test sees lots of use.

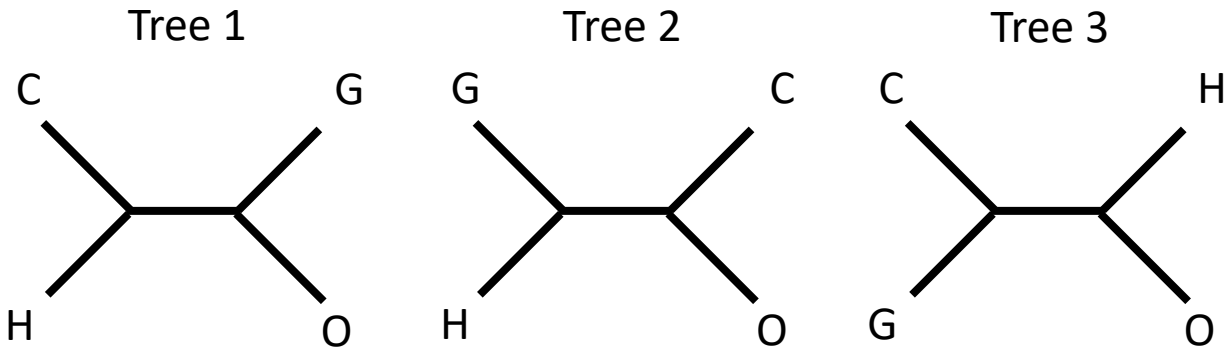
B. More on Resampling Estimated Log-Likelihoods

Hasegawa & Kishino (1989, 1994) developed an approximation to the traditional bootstrap that has seen lots of use. The idea is to save SSLs on each tree one is interested in. They used the tree possible resolutions of the human, chimp, gorilla tree (with orangutan as OG).

Site	Tree 1	Tree 2	Tree 3
1	1.35225388	1.35753553	1.3555375
2	1.35225388	1.35753553	1.3555375
3	2.62926496	2.6423911	2.6375218
.	.	.	.
.	.	.	.
898	1.78856476	1.80062479	1.79603731
-lnL	**2192.99696**	2201.66391	2200.37977

These SSLs, rather than the columns in the alignment, are then resampled with replacement, and the ML tree for each replicate is identified. This is fast, because the SSLs for each tree are already calculated and don't need to be re-estimated.

They provided bootstrap values for full trees:



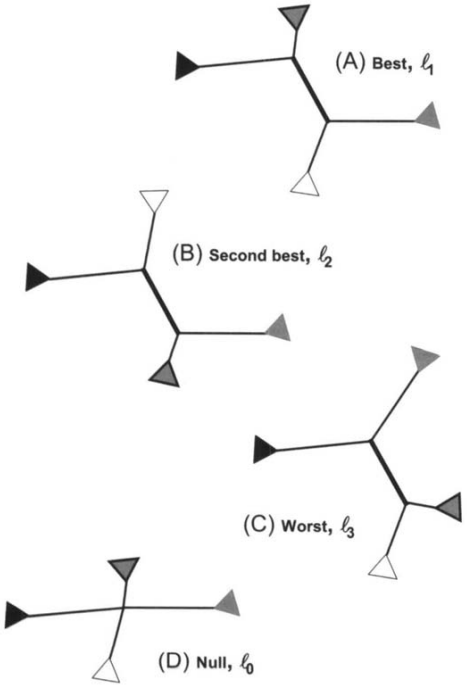
RELL Boot = 0.644 RELL Boot = 0.333 RELL Boot = 0.003

IQ-TREE uses RELL to generate nodal support values by saving SSLs on all trees it examines during a tree search (UFBoot); no searches are done on bootstrap replicates.

C. An Approximate LRT

Anisimova and Gascuel (2006. Syst. Biol., 55:539) have produced and developed a fast and approximate approach that is seeing increased usage in phylogenomics and has been built into PhyML.

The idea is that most phylogenetic hypotheses rely the presence/ absence of a particular internal branch, and we can test the significance of that branch.



The older idea is that we can compare the likelihood of the ML tree (Tree A, left) where a branch occurs with the ML score with that branch length collapsed to zero (Tree D).

This represents a nested pair of hypotheses, with the special case being a boundary value, so we could use the half χ^2 distribution to assess significance.

Thus, the test statistic would be $2(\ln l_1 - \ln l_0)$.

However, there's bias because we use the data to rank the trees.

A&G circumvent that bias by using a more conservative test statistic: $2(\ln l_1 - \ln l_2)$.

This is a really fast test because they only re-optimize the 4 adjacent branches.

In general, I don't like the idea of introducing an opposing bias to correct for an original bias, and I was the AE on the original paper. However, the authors did extensive power and accuracy analyses, and even an examination robustness to model violations. The method provides reasonably good approximations to results that would be achieved from more computationally intensive methods.

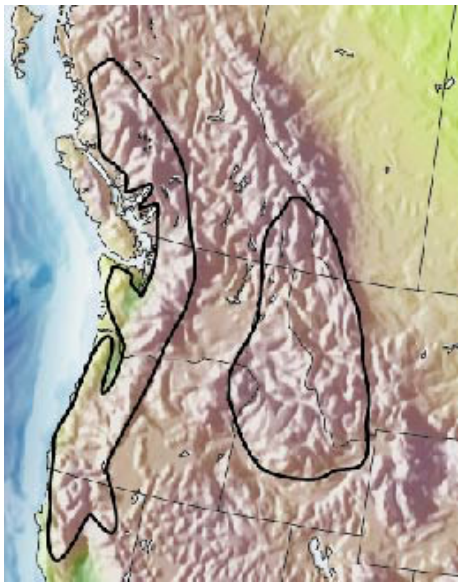
The approach has been built into PHYML and IQ-TREE, commonly used packages for phylogeny estimation from very large data sets. It does this fast test for all internal branches in the final tree to provide nodal support values.

C. Bayesian Hypothesis Testing

We can contrast these frequentist approaches with a Bayesian test of explicit hypotheses.

Remember, the probability that we're calculating in Bayesian statistics is the probability that some hypothesis examined is correct (conditional on the data, model of sequence evolution and assuming convergence of the MCMC).

I'll go through an example from a study of phylogeography of temperate rainforest ecosystems in the PNW.



These forests occur in two distinct areas, and many species have populations in both areas.

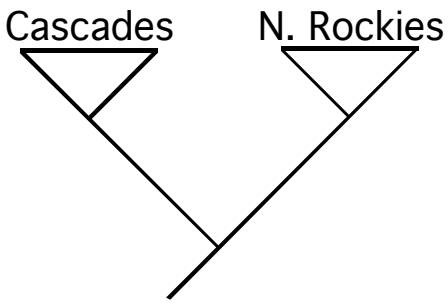
There have been several hypotheses proposed to explain how the distribution has arisen.

The most obvious hypothesis is that mesic forests were formerly continuous, and the rain shadow created by the Cascades caused the Columbia Basin to dry out.

This would have fragmented the formerly continuous ecosystem into the disconnected units we see today.

If this is the case, we expect the species that occur in the ecosystem to exhibit monophyly of each entity. Cascades populations should be monophyletic, as should populations from the inland rainforest ecosystems. Furthermore, that divergence is predicted to be pre-pleistocene.

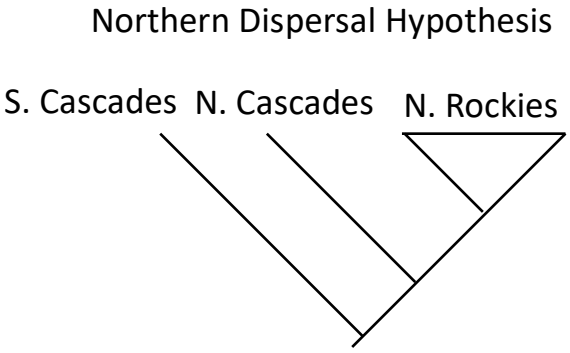
Here's the tree that is predicted by that hypothesis:

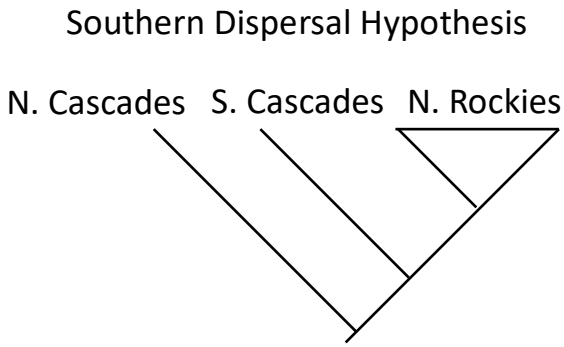


The populations from the Cascades should be monophyletic and the populations from the Northern Rockies should be monophyletic.

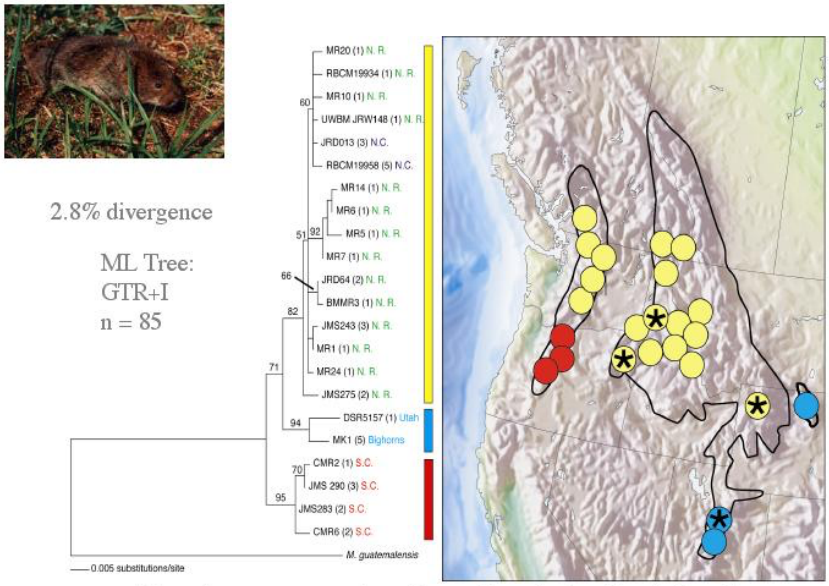
Alternatively, some have postulated that the rainforests that support this (and hundreds of other rainforest taxa) were eliminated from the Inland Northwest during Pleistocene glaciations and that the current disjucnt distribution has been achieved by post Pleistocene dispersal either from the northern or southern Cascades.

These hypotheses predict the following:





So, we go out, trap voles from across their range, collect DNA sequence data and do ML analyses, we get the following tree.



These data support a northern dispersal hypothesis, although from the Rockies to the Cascades.

If we run a typical MCMC and sample topologies to represent the posterior distribution of trees, we can use tree filters in PAUP* to assess the proportion of the trees in the distribution that are consistent with the topological predictions of each hypothesis.

If we do this for the water vole data, we derive the following results:

Bpp_{AV}	Bpp_{IDN}	Bpp_{IDS}
< 0.001	> 0.999	< 0.01

Remember these probabilities are different than conventional frequentist statistics. They represent conditional probabilities that each of the various hypotheses is correct, conditional on the data, the model and the priors (and convergence of the mcmc to the target distribution).