213. Hypothesis tests for comparing trees

Corresponding Author

Mark T. Holder 1200 Sunnyside Ave. Ecology and Evolutionary Biology, University of Kansas, 66045

E-mail: mtholder@ku.edu Telephone: 785-864-5789

Co-Author

Emily Jane McTavish 1200 Sunnyside Ave. Ecology and Evolutionary Biology, University of Kansas, 66045

E-mail: ejm@ku.edu Telephone: 785-864-5789

Abstract

Researchers often want to conduct statistical tests of competing hypotheses about genealogical relationships. Alternative tree topologies form a set of hypotheses that partially overlap with each other in a complex manner. This leads to difficult forms of the statistical problem of multiple tests. Furthermore, choosing an appropriate null distribution in tree testing is often difficult. Many methods for conducting statistical tests on topological hypotheses have been proposed. These approaches subtly differ in the questions they address and how they account for the manner in which the candidate set of alternative trees was generated. This entry reviews the current statistical approaches to comparing topological hypotheses, and discusses their appropriateness in a range of scenarios.

Keywords

Bootstrap
Likelihood ratio
Maximum Likelihood
Null distribution
Phylogenetics
Phylogeny
Selection bias
Species relationships
Statistical testing
Topology testing

Glossary

Likelihood Ratio test statistic: Twice the difference in log-likelihoods between two hypotheses. Note that the ratio of likelihoods of different hypotheses becomes a difference when log-transformed.
Kishino-Hasegawa test (KH test): A test of two competing topological hypotheses which is appropriate if both hypotheses are determined a priori
Shimodaira-Hasegawa test (SH test): An extension of the KH test which can be used to test one specific tree against another tree that was found by searching for the best tree among a large set of candidate trees.
Bootstrap: A resampling procedure used to approximate the variance of distribution. Many pseudoreplicate data sets are created by resampling the original data. An estimate is calculated for each pseudoreplicate. The standard error of the estimate on the actual data can be approximated by examining the variability across the collection of estimates from the pseudoreplicate data.
Bootstrap Proportion (BP): In the phylogenetic bootstrap, the BP is the proportion of bootstrap replicates that include a given clade. It is a measure of how well supported the clade is.
Resampling of estimated log-likelihoods (RELL): A fast approximation of bootstrapping to assess variability in likelihood ratio test statistics. Instead of resampling characters and conducting a full phylogenetic analysis, in RELL we simply resample the site-likelihoods from the original.
Maximum likelihood: See section on ML
Adjusted Bootstrap Proportion (aBP): A corrected bootstrap proportion of phylogenetic hypotheses which produces more accurate approximations of P-values when using the $P \approx 1 - BP$. The aBP accounts for the complex boundaries between phylogenetic hypotheses.

Body

Estimating the genealogical relationships between species is a computationally and statistically challenging endeavor. Whenever the tree that we estimate differs from the tree that we were expecting, a natural question is: "Can we explain the discrepancy between our estimate and a previous hypothesis solely on the basis of sampling error?" If we have a small amount of character data or our data set has a large amount of internal conflict, we should be wary of our estimate. In such cases we may want to publish our best estimate, but not reject previous hypotheses about the phylogenetic relationships. We can treat our previous expectations for the topology as an null hypothesis for statistical testing.

The standard recipe for approaching this type of question in the frequentist approach to statistics is to:

1. choose a test statistic that describes how the data deviates from expectations under the null hypothesis;
2. describe the distribution of values that we would expect for this statistic if the null hypothesis were true – this is the null distribution of the test statistic;
3. compare the value of the test statistic to the null distribution to find out the probability of seeing a result that is at least this surprising if the null were true – this probability is the P-value; and
4. reject the null hypothesis if the P-value is small.

0.1 Likelihood ratio is (almost always) the best test statistic

The choice of which test statistic to use in this form of testing is very open-ended. It may be tempting to use a test statistic that measures how different our estimated tree is from the trees that are a part of the null hypothesis. However, using only a tree-to-tree distance is suboptimal because it does not capture the strength of support for different groupings. Fortunately, the "law of likelihood" assures us that the ratio of the likelihoods for competing hypotheses captures all of the evidence in the data for one hypothesis over another (see "Maximum likelihood"). So, we can focus our testing on a procedure that uses the popular "delta" statistic:

$$\delta(T_1, T_0 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_0 \mid X)\right]$$

where T_1 and T_0 are the two trees that we are comparing, X represents the data, and and L represents the likelihood of a tree for a data set. Note that δ is twice the difference in log-likelihoods for our two hypotheses. It is a measure of how much more support one hypothesis has than the other. We often call this the likelihood-ratio test statistic. Note that the statistic is calculated on the log-scale. So, the logarithm of the ratio of two likelihoods becomes a difference between the log-likelihoods.

So what is the likelihood of a tree? The likelihood, $L(T_1 \mid X)$, is simply the probability that we would generate a data set identical to our data (X) if tree 1 is an accurate depiction of the phylogenetic relationships. You can think of the likelihood as a measurement of how well our model (the tree) matches our data. Trees that would cause us to expect the patterns of data that we actually observe have higher likelihoods.

Because the likelihood is the probability of the entire data set, it uses all of the information in the data. Test statistics that are not functions of the likelihood are less powerful because they do not capture all of this information. While DNA sequence data is commonly used to reconstruct phylogenies, these approaches can be applied to any data for which the likelihood of a tree can be calculated (see "Maximum likelihood" This is intended to be a cross-ref to another entry).

0.1.1 Null hypotheses

Depending on the question at hand our null hypothesis may be one of several types. The null may be that multiple topologies are equally good explanations, that one topology corresponds to the true set of relationships, or that one specific group is not a part of the true tree. Different tests described below are formulated with different null hypotheses. In addition, our hypotheses generally are statements about 'topologies' - the relationships among taxa. However, we calculate likelihoods on trees - topologies with branch lengths optimized according to the model of evolution to maximize the likelihood of the data. Thus, the testing procedures must include some method of handling the estimation of branch lengths, despite the fact that the hypotheses being tested rarely refer to branch lengths explicitly.

0.1.2 Calculating the δ statistic in phylogenetics

To conduct a test, we also need an alternative tree to contrast with the null hypothesis. As we will see later, we run into problems if the alternative tree that we want to test is not known when we start the study. This often occurs if we initiate a study to test a hypothesis. We may suspect that a particular grouping (the null) is not true,

but we do not have a specific alternative. An appealing approach is to estimate the best tree for our data, and then perform a test to see if we can reject the null. If we use the data to select the alternative tree to test, then we must alter the threshold that we use for deciding when we reject the null hypothesis (see below).

However, if we start our study interested in comparing two specific trees, then calculating δ is easy: we simply choose the most appropriate models of character evolution for our data and then calculate the log-likelihood of each tree. Calculating δ is easy in this case, but to conduct a test we also need to know the null distribution of the δ test statistic.

0.2 The null distribution of the LR is not trivial to obtain

0.2.1 The null hypothesis that expected likelihood is the same for multiple trees

To think about the null distribution, we have to think clearly about what hypothesis we are testing. Consider the null hypothesis that trees with or without a particular group are expected to be equally good explanations of the data. This may seem like an odd null hypothesis – after all when we are about to perform a test we do not actually expect all of the trees to be tied.

Imagine, for example, hypotheses for the relationships between humans, chimps, and gorillas, with orangutans as an outgroup. (Figure 1). If our null was that in the true tree humans are the sister group to gorillas (Figure 1A), the most challenging form of this null would occur if the lineage that was the last common ancestor of only humans and gorillas persisted for an infinitesimal amount of time before speciating (Figure 1D). By "most challenging" null, we mean the null hypothesis that will be hardest to reject. In that case, where the true tree has a zero-length internal branch, we would expect no character changes in the data set to support the grouping of humans and gorillas. Thus, we should expect each of alternative trees (e.g. trees with humans sister to the chimps [Figure 1B], or trees with chimps sister to gorillas [Figure 1C]) to explain the data just as well as the true topology (Figure 1A). Frequentist hypothesis testing tries to give the null hypothesis the benefit of the doubt at every step of hypothesis testing by focusing on these most challenging cases. Because of this focus on the most challenging case, the testing procedure will let us make very general statements such as "either the null hypothesis is false or some rare form of sampling error occurred; we would only expect this extreme effect of sampling error in $P \times 100\%$ of tests..."

0.2.2 Using per-character δ's to generate a null distribution

Using the difference in log-likelihoods between the alternative and null hypothesis is appealing in terms of statistical power and making the best use of our data. Unfortunately, we do not have any clear, simple theoretical description of the null distribution of this statistic. In other words, if the null were true, we do not know how large the δ statistic might get solely as the result of sampling errors. Having a null distribution is crucial because it lets us convert an observed value of the test statistic to a P-value.

With no theoretical descriptions of the null distribution of the δ statistic, we seem to be incapable of conducting a hypothesis test. Fortunately, we can look at the phylogenetic signal within each character in our data set. This allows us to get a sense of whether the difference in log-likelihoods is coming from a few or many characters and whether or not there is much internal conflict. In almost all models used to describe character evolution in

phylogenetics, different characters are assumed to evolve independently of one another. Thus the likelihood is calculated by taking a product over characters. On the log-scale the value of the δ statistic is simply the sum of the per-site versions of the δ statistic. So, looking at the distribution of the per-site differences in log-likelihoods gives us a rich view of the per-datum quantities that make up our test statistic.

Imagine that we have 50 characters in our data set and a total log-likelihood difference of 5.0 between our null hypothesis and an alternative tree. Is this degree of support statistically significant? The answer actually depends on the variance in the support from character to character.

Our null hypothesis states that trees with or without the clade of interest (e.g. Humans and Gorillas) are expected to explain the data equally well. If our tree inference method is not biased, we could imagine that every informative character has a 50% chance of favoring the null tree, and a 50% chance of supporting the alternative tree (as we will see below, this assumption is too simplistic because it ignores the number of trees that correspond to each hypothesis, but this simple case is a good starting point).

Consider the scenarios shown in Figure 2. In case *a*, all 50 characters prefer the alternative tree by 0.1 log-likelihood units. In this case, the phylogenetic signal in the data is very repeatable and internally consistent. If the magnitude of character preference were only 0.1 and a tree with the clade was equally likely explain the data as a tree without the clade, what is the probability that we would see an overall difference in log-likelihoods of 5.0 from a data set of 50 characters? To get a log-likelihood difference of 5, each of the 50 randomly selected sites would have to favor the same tree. The probability is tiny – it is the equivalent of flipping a fair coin 50 times and always seeing the same face of the coin (the probability is 0.5^{49} if you conduct a typical, two-sided hypothesis test). Seeing a total log-likelihood difference of 5.0 out of fifty characters each of which has a preference of around 0.1, is an extremely unlikely event if the null hypothesis is true. So we have reason to reject the null.

Figures 2*b* and 2*c* depict two other possible outcomes of looking at the per-character log-likelihood differences. Note that the alternative tree is favored by 5.0 log-likelihood units in each of the figures 2*a*, *b*, and *c*. In *b*, 49 of the characters have no preference between the trees, and one character prefers the alternative tree by 5.0 log-likelihood units. In other words all of our signal is coming from one of our 50 data points. Unsurprisingly, this should not lead to a significant test result. If our character data is sampled from a universe with these sort of properties (probability of 0.02 of getting an informative site, and informative sites are strongly informative), then we would expect to see quite a bit of data set-to-data set variability in which tree is preferred. In fact there is a very strong chance (≈ 0.63) that one of the trees will be preferred by at least 5.0 log-likelihood units even under the null hypothesis. Hence we do not have enough support in the real data set to make us reject the null hypothesis.

In Figure 2*c*, we see a case of strong internal conflict in the data. One character has no preference between the hypotheses; 25 characters prefer the alternative tree; and 24 characters prefer the null tree. Among the 49 characters which have a preference for one tree over another, the magnitude of the preference is strong – centered around 5.0 log-likelihood units. If this distribution described the signal in our data, then we should not be surprised at all to see a preference of about 5 log-likelihood units. Due to sampling error, we would almost always see this much of a difference.

The interactive web page http://phylo.bio.ku.edu/mephytis/lrt-null-nonparametric.html allows you to construct the null distribution for these three scenarios of variation in data signal.

0.2.3 KH test

The previous section tried to provide some intuition for why the distribution of differences in log-likelihoods per site has a strong effect on whether or not we consider a result such as $\delta = 10$ to be enough evidence for us to reject the null hypothesis (recall that the δ is twice the difference in log-likelihoods; so in the previous example a difference in log likelihoods of 5 implies $\delta = 10$). To derive the null distribution requires more care than our intuitive exercise. Early efforts (Cavender, 1978; Templeton, 1983) had considered calculating a P-value from the total difference in scores using data that had integer scores only (or which scoring systems in which a character could just be defined as supporting or conflicting with a tree). Kishino and Hasegawa (1989) tackled this problem in a way that fully used the magnitude of the difference in log-likelihoods.

Kishino and Hasegawa (1989) reasoned that, because the difference in log-likelihoods between trees is simply the sum of many per-site differences, the statistic should (according to the Central Limit Theorem) follow a normal distribution if the data set is large. They used the empirically estimated per site log-likelihoods to estimate a variance of the normal distribution. They note that the sampled distribution of differences in log-likelihoods is not identical to the null distribution. However, the sampled distribution should have a similar variance to the null distribution. Under the null hypothesis, both trees should explain the data equally well. So, the the null distribution should be centered at 0. Combining the mean of 0 with the sampled variance provides a normal distribution which approximates the null distribution.

Kishino and Hasegawa (1989) also noted that we can avoid the explicit assumption of normality by simply re-sampling the per-site differences in log-likelihoods many times to generate a distribution of δ statistics. This distribution will not be centered around 0, because whatever tree is favored by our real data will tend to be favored in the resampling procedure. We can correct for this bias very simply – by subtracting from each sampled δ value the mean over all of the simulation replicates. This enforces the null hypothesis expectation that the two trees would explain the data equally well. This procedure of resampling the per site likelihoods is referred to as the "resampling of estimated log-likelihoods" (RELL) bootstrap, and testing the significance of the difference between trees in this way is the "KH test", after Kishino and Hasegawa (1989). Using the "subtract mean" option for centering in the `http://phylo.bio.ku.edu/mephytis/lrt-null-nonparametric.html` demonstration mimics this RELL approach to performing the KH test.

Susko (2014) questioned the assumption that the difference in log-likelihoods should follow a normal distribution. He has found the KH test using the normal as the null distribution and the RELL approximation to be too conservative. He introduced two new approaches for generating the null distribution for the KH-test: a procedure that uses simulation from normal distributions to approximate the shape of the log-likelihood surface on phylogenetic trees, and a procedure using a null distribution created by mixing χ^2 (chi-squared) distributions. The mixture of χ^2 distributions is suitable for testing two trees that are very topologically similar to each other; Susko also introduced some conservative approximations that can be used in more general cases.

0.2.4 Selection bias problems

The largest impediment to the widespread use of the KH test is that it assumes that both the null hypothesis tree and the alternative are known a priori. In phylogenetics it is much more common for the null hypothesis to correspond to a set of trees and for the alternative hypothesis to be implicitly defined as "all trees that do not fit the null." This introduces a multiple testing problem: if you select the ML tree to use as your alternative,

then you are implicitly wading through a huge number of possible alternative trees and selecting the one that is most likely to lead to rejection of the null. Selecting the most promising alternative tree makes sense in terms of constructing the most powerful test, but this selection of an alternative has a large effect on the null distribution of the test statistic. Failing to account for the multiple testing aspect of phylogenetic topology testing can lead to gross exaggerations of the strength of support against the null – this problem can be referred to as selection bias.

These considerations are even stronger in the case of another common pattern in phylogenetic data analysis: a researcher estimates a tree, and then would like to highlight the groupings that are significantly supported. In this case neither the null nor the alternative is specified a priori.

The Shimodaira-Hasegawa test (SH test) was the first phylogenetic testing procedure to deal with the problems related to selection bias.

0.2.5 SH Test

Like the KH test, the SH test uses the likelihood ratio test statistic, and it uses a resampling procedure to generate the null distribution. In order to account for selection bias, the researcher must inform the testing machinery of the candidate set of trees that were considered plausible before the data set was examined. The SH test will test each of these trees in a way that acknowledges the fact that the researcher has searched through this set of trees to find the one with the highest score.

Recall that when performing the KH test we used the null hypothesized expectation of 0 for the difference in log-likelihoods. Because we are using the best of many trees as the alternative tree in the SH test, we can no longer assume that the δ statistic will be centered around 0. Clearly, the maximum likelihood tree will have a δ statistic that favors it when we compare it to any other tree - otherwise it would not be the maximum likelihood estimate of the tree. In the SH test we allow each of the candidate trees to influence our null distribution for the delta statistic. Then we calculate a P-value for each tree individually to ascertain if its difference from the maximum likelihood tree's score can be explained under the null hypothesis.

We can enforce the null hypothesis that a tree is no better than other members of the candidate set by:

1. resampling the sites to generate a estimate of the sampling distributions of log-likelihoods for each tree in the candidate set.
2. calculating the mean log-likelihood for each tree across all resampled replicates, and
3. subtracting that mean log-likelihood for a tree from each of its resampled log-likelihood values.

This centers the distribution of log-likelihoods for each tree across resampling replicates around 0.

To mimic the selection bias, for each replicate we find the highest centered log-likelihood over all of the trees. We use that value to stand in for the maximized likelihood. For each bootstrap replicate, each candidate tree's centered likelihood is compared to the highest centered log-likelihood for that replicate. This generates a sample of the null δ statistic for each tree.

For each topology in each bootstrap replicate we have generated a δ value under the expectation that on average that tree is no better or worse than the ML tree. For any candidate topology, we can then compare the observed

δ to this distribution. The P-value for each tree is the fraction of the RELL bootstrap in which the resampled δ statistic for that tree is more extreme than that tree's δ statistic calculated on the real data set. In this way we are able to capture not only the effects of the number of candidate trees we are considering on the expectation of δ, but also the variance structure of the log-likelihoods for individual trees.

The SH test makes the pessimistic assumption that we should treat every member of the candidate set as equally likely a priori and that each contributes in an independent manner to the multiple testing problem. Because a few sites may strongly disfavor a large number of candidate trees, centering each tree's set of scores around 0 is probably too generous an assumption with respect to the worst trees in the candidate set. The result of making this cautious assumption is that the test tends to be too conservative.

A more pressing empirical problem with the test is that the researcher has to specify the candidate set of trees honestly. Omitting a tree from the candidate set will tend to exaggerate the significance of the trees tested. For even a moderate number of tips in a tree, there are a very large number of possible, and a priori plausible, trees. Thus, the test can be infeasible to apply because it would requre storing and resampling the likelihood for too many trees.

0.2.6 Parametric bootstrapping

The methods described above leverage the variation in the likelihoods across sites in observed data to generate distributions of expected δ values. This gives us the ability to estimate P-values. Alternatively, we can use our understanding of evolutionary process to generate expected distributions of data sets under a topology of interest. First, a substitution model is selected for the data. When we estimate a phylogeny using likelihood or Bayesian methods, we also estimate a model of evolution under which our sequences have evolved. For example, if we estimate a phylogeny using a general time reversible model for DNA sequences (see "Evolutionary models and model selection" cross ref intended - should be checked), we estimate the rate of change between each class of nucleotides. The choice of substitution model is often determined by a likelihood ratio tests or an information criterion. The parameter values for the chosen model, θ, are estimated for the null hypothesis topology.

Next a series of data sets are simulated on that topology under the parameters (θ) of that inferred model of evolution. A search for the maximum likelihood (ML) tree is performed on each of these simulated data sets. In addition, the the likelihood of the optimized null topology is calculated for each simulated data set. The difference in likelihood score, δ, between the ML topology and the topology generating the data, is calculated for each simulated data set. This procedure produces a distribution of δ values that would be expected if the null tree were the correct topology. By comparing the observed δ to the distribution of simulated δ you can calculate the probability that the observed difference in likelihood between the ML tree and the null would have been observed if the null were the true tree. Testing phylogenies in this manner is often called the SOWH test, after Swofford, Olson, Waddell, and Hillis who first described using parametric bootstrapping to test topologies (Swofford et al., 1996; Goldman et al., 2000).

There are several advantages to using parametric bootstrapping for topology testing. Because these tests use the inferred model of evolution the SOWH test has much higher power than the non-parametric approaches described above (Goldman et al., 2000; Buckley, 2002). Also, conveniently, the SOWH test doesn't rely on topologies having been determined a priori. As the data sets are generated with respect to a specific topology, in comparison to the observed ML tree, these tests may be performed on the ML tree without the necessity of attempting to encompass all potentially plausibly trees, as is appropriate for the SH test. There may be millions of plausible trees, so this

is a valuable advantage.

However, there are several important disadvantages to a parametric bootstrapping approach as well. Firstly, there is a large computational burden of estimating the maximum likelihood phylogeny for hundreds of simulated data sets. If the individual ML searches do not find the best topologies, δ values in the null distribution will be underestimated; this can lead the test to reject too frequently. Although there are some shortcuts which can decrease the computational time, these may affect rejection rates in undetermined ways (Goldman et al., 2000). Nonetheless, with increases in computational power, and the application of large computing clusters the technical issue of searching for ML trees can be overcome. More troubling is the issue that available models of sequence evolution under which we are able to simulate, rarely capture the complexity of observed empirical data (Buckley, 2002). This oversimplification of simulated data likely results in phylogenetic inference on these data sets having an easier time inferring the correct tree. This will result in smaller values of δ statistics in the null distribution, and make it easer to reject the alternative topology as the generating tree. Researchers should be aware that if they reject a null hypothesis based on parametric bootstrapping, the appropriate conclusion is that the degree of support observed is too large to be easily explained by chance assuming that the simulated model of sequence evolution adequately mimics the level of conflicting signal in the true generating process. Indeed, in many comparisons of SH and KH tests to SOWH tests, parametric bootstrapping strongly rejects the null in cases where those non-parametric tests are unable to do so (Goldman et al., 2000). This may be due to some differences in how the null and alternative hypotheses are generated under these different tests (rejecting one alternative tree vs rejecting all other trees) and the general result that parametric tests have more statistical power. The more worrisome possibility is that SOWH may have excessive type I error when the model of evolution used is too simplistic (Buckley, 2002).

Susko (2014) has made some suggestions to improve the SOWH test and decrease the prevalence of type I error. If instead of using a fully resolved topology for the null distribution, a constrained topology in which branches differentiating the two hypotheses are collapsed (e.g. Figure 1D), the test would better capture the borderline between support for one topology or the other. By comparing this borderline topology, this modification gives a fairer chance to the null hypothesis and lowers the excessive rejection rates of the SOWH test (Susko, 2014).

0.2.7 Quick and dirty methods

Finding the likelihood ratio between the maximum likelihood tree and the best tree that lacks a particular grouping can be computationally expensive. Often a nearest neighbor of the ML tree which has all of the groups with the exception of the clade of interest will be the "next best" tree. A few testing methods have exploited this to produce quick statements of support for a branch in the ML tree by only considering the neighboring trees that lack that branch. This makes the δ test statistic quick to calculate. If a mixture of χ^2 distributions is used as a null distribution, this approach is referred to as an "approximate Likelihood Ratio Test" (aLRT) (Anisimova and Gascuel, 2006). Anisimova et al. (2011) found that the aLRT rejects the null too frequently, and they recommend another fast approximate statement of support, the aBayes statistic. To calculate the aBayes score for a branch in the ML tree, one divides the likelihood of the tree containing the branch by the sum of the likelihood of that tree and the two trees that are nearest neighbors of the tree but lack the branch in question (see Figure 3). Simulations by Anisimova et al. (2011) indicate that using a 0.95 threshold for aBayes does not reject the null too often, but there are no theoretical results to indicate that this conclusion is general.

0.3 Pure resampling approaches

The methods discussed above all use the likelihood ratio test statistic - they only differ in how they generate the null distribution for that test statistic. It is also possible to take a fully resampling based approach in many cases. Felsenstein (1985) introduced the general statistical technique of bootstrapping to the field of phylogenetics as a method for determining confidence for different groupings in the tree. In the bootstrapping procedure, you create an artificial data set the same size as your data by resampling the characters (randomly and with replacement) from your original data set. You can estimate a tree on this bootstrap-pseudoreplicate data set (see "Tree Support Measures"). This process can be repeated many times to produces a collection of bootstrap trees. Clades that show up in all or almost all of the bootstrap trees must have substantial support. Even in the face of the sampling error of the bootstrapping procedure, these clades are almost always recovered. Thus, the amount of support for these clades in your data must be larger than the amount of noise that is typically produced by sampling error. If a clade appears across many different replicate subsamples of your data, you can have some confidence that the clade is not just an artifact of sampling error.

For some statistical problems, one can even approximate a P-value using 1.0 minus the bootstrap proportion (BP). So, if a result was found in 96% of the bootstrap replicates, its P-value would be approximately 0.04. Unfortunately, this simple approach does not apply to phylogenetic problems. The connection between the P-value and $1 - BP$ is complicated (see Alfaro et al., 2003; Newton, 1996, and references therein). Correction for the $1 - BP$ approach have been developed based on theories about the geometry of tree space. This body of theory underlies the Approximately Unbiased test for phylogenies (Shimodaira, 2002) and the multli-level bootstrap method of Efron et al. (1996).

0.3.1 aBP

Susko (2010) and (2014) developed improved phylogeny testing procedures that pay particular attention to the boundaries between hypotheses where the branches that differ between the trees all have branch lengths of 0. At that point, the different topologies make exactly the same prediction of what data we would expect to see – so it is impossible to distinguish data arising from one topology from data that arose from other trees. If we are treating one of the trees as the null, this point should be the hardest point of the null hypothesis to reject. Not only are the boundaries between hypotheses curved, but more than two trees' boundaries come together at this point of zero branch lengths. Susko's approaches evaluate the curvature of the likelihood function for different trees at this point. After fitting normal distributions to mimic the shape of the likelihood surface, his "adjusted bootstrap proportion" (aBP) can correct for the complexities of the hypothesis testing problem. Thus $1 - aBP$ provides a better estimate of the P-value than the simple $1 - BP$ approach.

The demonstration at `http://phylo.bio.ku.edu/mephytis/bootstrap.html` allows you to visualize bootstrapping in a parameter space that describes the frequency of different types of data patterns. This is a projection of the true (high dimensional) space down into a coordinate system that show the relative frequency of the data patterns that support each of three possible trees when you analyze data under the parsimony criterion. The boundary point in the middle of the triangle corresponds to the trees when their internal branch lengths are 0. This is the null hypothesis point for tree testing which is the focus of aBP correction of Susko (2014).

1 Further considerations

This article has outlined methods for considering sampling error when phylogenetic hypotheses are tested. There are many other sources of error that a practitioner must be aware of that are completely ignored by the tests presented here. For instance:

- If the method of tree inference is not sophisticated enough to interpret the data, then the tree inference may be subject to systematic bias. For example, for some sets of branch lengths parsimony is guaranteed to estimate the wrong tree if given enough data. For you to have confidence in your estimated tree, you must be confident that it is not the result of either sampling error or systematic error.
- If you have data from one or a few loci, you may have confidence in your estimates of the gene trees. However, you should be aware that the species tree may differ from the gene tree due to hybridization, lateral gene transfer, problems of paralogy, or simply the failure of polymorphisms to fix during the duration of an ancestral species lineage. The topology tests described above do not accommodate these effects.
- Errors in the construction of the data set are ignored by the methods above. For example, the true alignment is not known and phylogenetic biases of alignment algorithms can affect the strength of support. See Karin et al. (2014) for recommendations about incorporating alignment uncertainty in parametric bootstrapping.

References

Alfaro, M. E., Zoller, S., and Lutzoni, F. (2003). Bayes or bootstrap? a simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular Biology and Evolution, 20(2):255–266.

Anisimova, M. and Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Systematic Biology, 55(4):539–552.

Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C., and Gascuel, O. (2011). Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Systematic Biology.

Buckley, T. (2002). Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Systematic Biology, 51(3):509–523.

Cavender, J. A. (1978). Taxonomy with confidence. Mathematical Biosciences, 40(3-4):271–280.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences of the United States of America, 93:13429–13434.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution, 39(4):783–791.

Goldman, N., Anderson, J., and Rodrigo, A. (2000). Likelihood-Based Tests of Topologies in Phylogenetics. Systematic Biology, 49(4):652–670.

Karin, E. L., Susko, E., and Pupko, T. (2014). Alignment errors strongly impact likelihood-based tests for comparing topologies. Molecular Biology and Evolution, page msu231.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution, 29:170–179.

Newton, M. A. (1996). Bootstrapping Phylogenies: Large Deviations and Dispersion Effects. Biometrika, 83(1):315–328.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Systematic Biology, 51(3):492–508.

Susko, E. (2010). First-Order Correct Bootstrap Support Adjustments for Splits that Allow Hypothesis Testing When Using Maximum Likelihood Estimation. Molecular Biology and Evolution, 27(7):1621–1629.

Susko, E. (2014). Tests for two trees using likelihood methods. Molecular Biology and Evolution, page msu039.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference.

Templeton, A. R. (1983). Phylogenetic Inference From Restriction Endonuclease Cleavage Site Maps with Particular Reference to the Evolution of Humans and the Apes. Evolution, 37(2):221–244.

List of Relevant Websites

Any websites relevant to this article should be cited here, not within the reference section.
http://phylo.bio.ku.edu/mephytis/lrt-null-nonparametric.html

Cross References

Please provide suggestions for cross-references to other articles within the Work. Full Table of Contents will be available on http://editorial.elsevier.com/ once we are fully commissioned.
"Maximum Likelihood"
"Evolutionary models and model selection"
"Tree Support Measures"

Biography and photo

A biography and photo of each author, to be included with the article when published online.

Mark T. Holder is an associate professor in the Department of Ecology and Evolutionary Biology at the University of Kansas.

Emily Jane McTavish is a postdoc in the Holder lab at the University of Kansas and a Humboldt Research Fellow at the Heidelberg Institute for Theoretical Studies.

Figures

Figure 1.

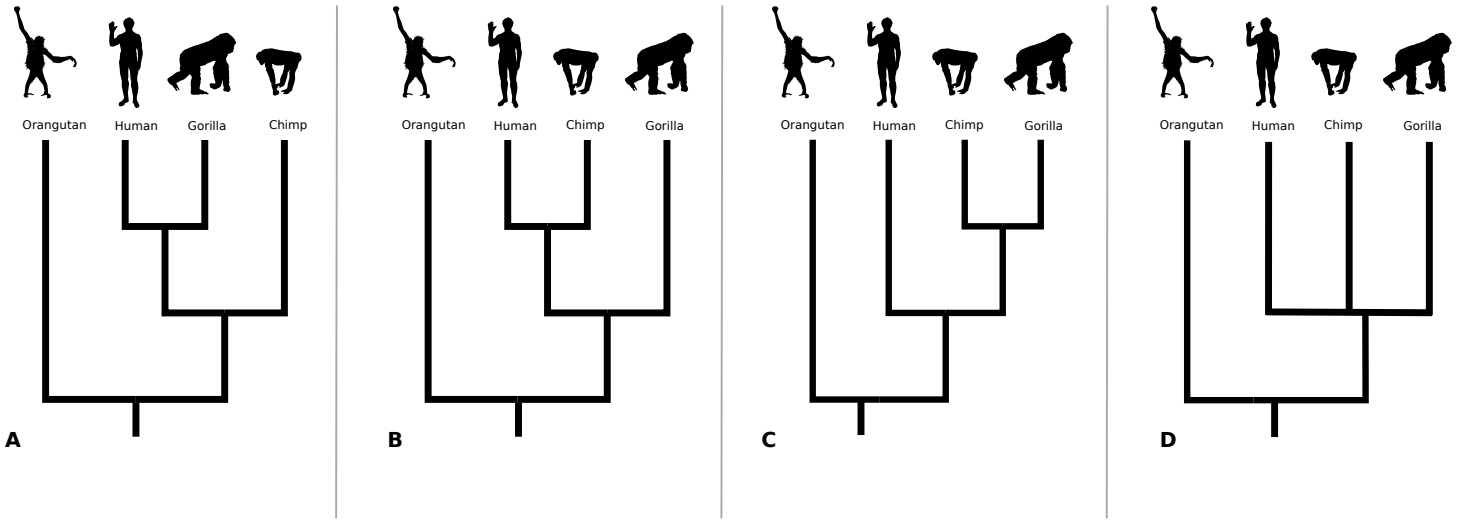

Figure 1: Three example trees demonstrating alternative potential topologies for the relationships between humans, chimps, and gorillas, with orangutans as an outgroup. (A), (B), and (C) are potential hypotheses, and (D) represents the border between these hypotheses. All branches which differ between (A), (B), and (C) are collapsed in (D). Silhouettes are from phylopic.org, credit Mike Keesey and Gareth Monger CC BY 3.0

Figure 2.

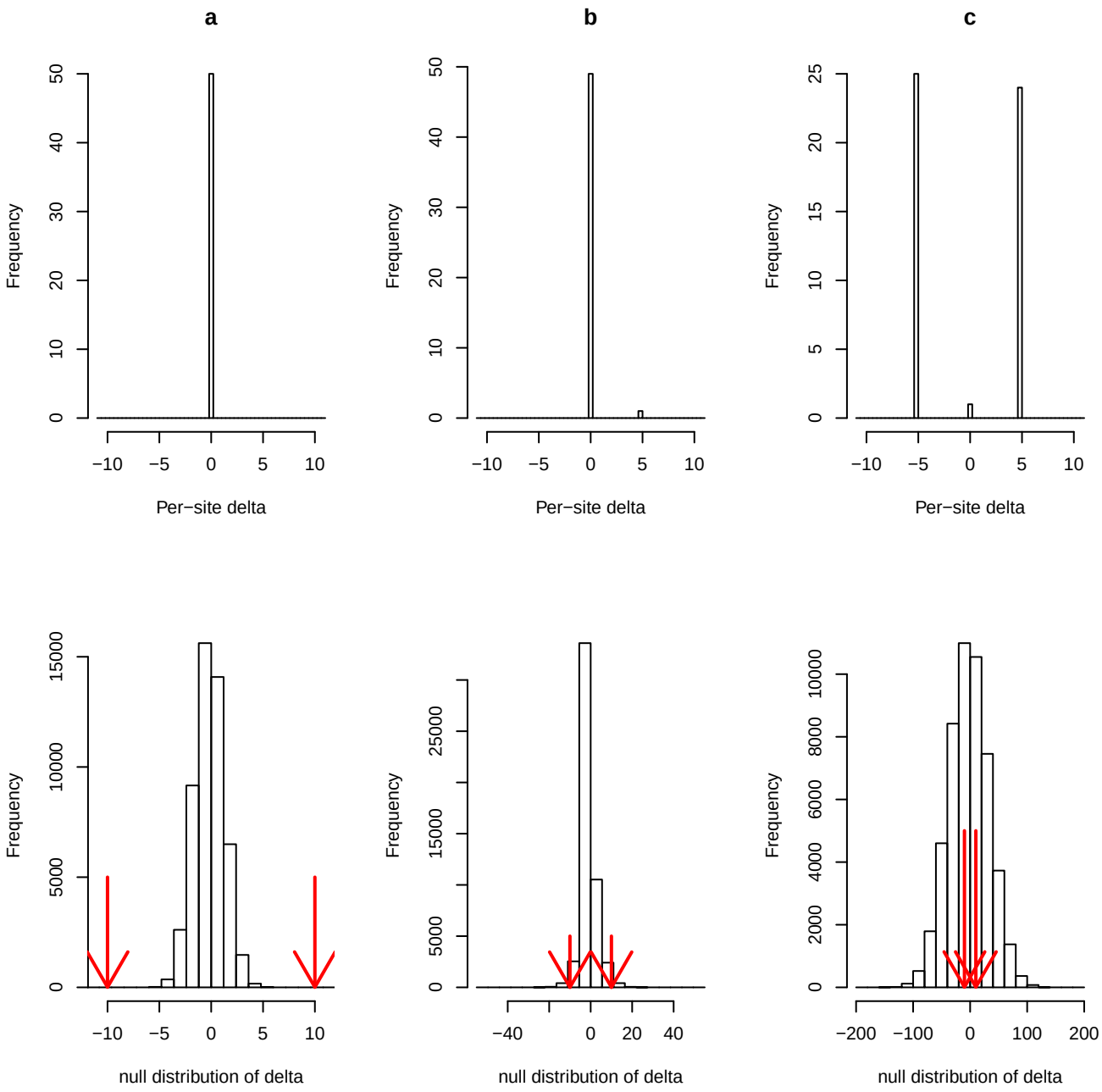

Figure 2: Null distributions under three different scenarios describing the distribution of per-site LR statistics. Each column shows the per site delta distribution (top) and an approximation of the null distribution (bottom). Note that the x-axis differs for each of the null distribution plots; red arrows on each plot indicate the test statistic value of 10.0 and -10.0 (the P-value is the probability of falling in the tails of the distribution more extreme than these arrows). In case a, all 50 characters prefer the tree by 0.1 log-likelihood units (per site $\delta = 0.2$). In b, 49 of the characters have no preference ($\delta_i = 0$) between the trees, and one character prefers the alternative tree by 5.0 log-likelihood units. In c, we see a case of strong internal conflict in the data. One character has no preference between the hypotheses; 25 characters prefer the alternative tree; and 24 characters prefer the null tree. Among the 49 characters which have a preference for one tree over another, the magnitude of the preference is strong – centered around 5.0 log-likelihood units

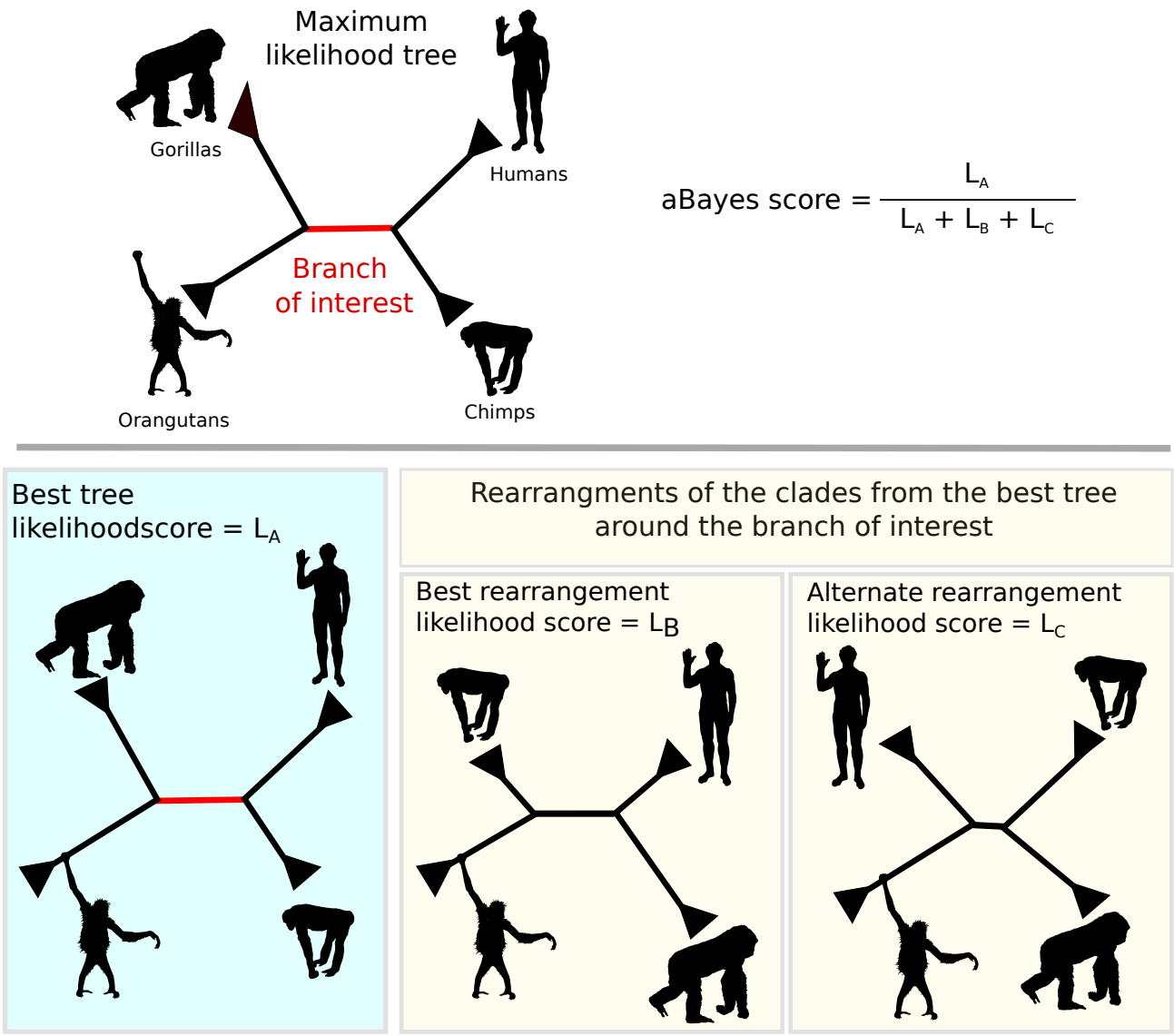

Figure 3: To calculate the aBayes score for a branch in the ML tree, one divides its likelihood by the sum of the likelihood of that tree and the two trees that are nearest neighbors of the tree but lack the branch in question. Silhouettes are from phylopic.org, credit Mike Keesey and Gareth Monger CC BY 3.0