A COMPARISON OF METHODS FOR CONSTRUCTING CONFIDENCE SETS OF PHYLOGENETIC TREES USING MAXIMUM LIKELIHOOD

by

Etai Markowski

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
March 2021

Table of Contents

List of Tables

List of Figures

0.1 Abstract

This thesis compares six phylogenetic tests, or equivalently, methods for con- structing confidence sets of phylogenetic trees : the Kishino-Hasegawa (KH) test statistic, Shimodaira–Hasegawa (SH), two versions of the Approximately Unbiased (AU) test, Chi-square and Bonferroni. The Bonferroni test is a new variation of the Chi-square test that corrects for selection bias. A variation of the AU test, AU Corrected, is considered that adjusts for difficulties arising when bootstrap support for trees is low. Confidence regions for each test are ex- amined using simulations from six and eight-taxon trees. We consider differing internal edge-lengths and challenging inference scenarios where some internal edge-lengths are equal to 0. In the second part of this thesis we apply the same tests to multiple real-world data sets, some with larger numbers of taxa, and make references to the observations and trends obtained in the previous part.

Chapter 1

Introduction

This thesis considers evolutionary inferences drawn from DNA or amino acid sequence alignments. A sequence alignment arranges sequences for a number of species into sites (Wu 2010). Although alignments are subject to error, the intent is that the i^{th} corresponds to a 'homologous position' across all species and across ancestral species. That is, the i^{th} site in all sequences in all living and ancestral species evolved from a common ancestral site in the earliest common ancestral species in the tree by vertical descent. The sequences are represented as rows in a matrix (see example Table 1.1) in such way to match as many nucleotides along columns.

1.1 Conventional Phylogenetic Models

For nucleotide data, each site can be one of A, C, T, G, the four nucleotide bases of DNA strands. Although there are possible errors in alignment, conventional evolutionary models usually ignore those errors and treat sites as corresponding to the same physical position. Thus a change of one of the nucleotides in a site is the consequence of evolutionary substitutions along the tree, and that is what we are trying to exploit in developing models of evolution.

For example, consider a subset of the α-globin plus β-globin gene sequences from Yang (1993). Table 1.1 shows sites 7-10 of the full 570 site alignment.

Table 1.1: DNA sequence example

mamal	...	site 7	site 8	site 9	site 10	...
Human	...	C	C	G	C	...
Rabbit	...	C	C	-	C	...
Rat	...	G	C	G	A	...

Note that to denote a **gap** a '-' is used. The reason for a gap is either a **deletion** or

an **insertion**. For instance, at site 9, the rabbit lineage may have lost a nucleotide at that position, making their gene shorter. Alternatively, the ancestor of Human, Rat may have gained a nucleotide.

Denoting x as a *'site pattern'*, in Table 1.1, the 7th column represents site pattern $x_7 = \text{CCG}$. Meaning, C was observed for Human and Rabbit, and G for Rat.

Conventional phylogenetic models make three main assumptions:

1. Evolutionary events at different sites occur independently of each other in an independent and identically distributed way.
2. Given the ancestral character state for an edge at a site, evolution along that edge occurs independently of all else.
3. Evolution along an edge occurs according to a continuous-time Markov chain.

Because of the first assumption it suffices to consider evolution separately for each site.

Considering *site 7* in the table above, we assume that this corresponds to the same physical location in the genes of the observed taxa and their ancestors. One possible explanation for the data is given in Figure 1.1.

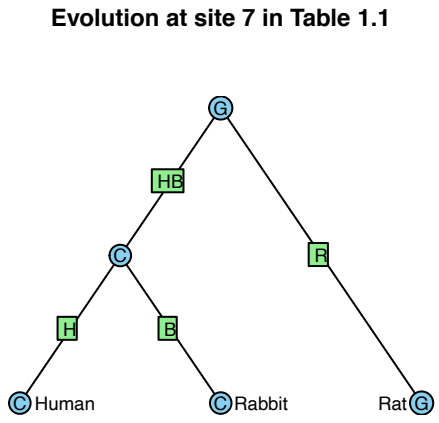

Figure 1.1: Site 7

In the illustration above a substitution occurred in the lineage leading to Human (H), Rabbit (B).

The reasons for differing nucleotides arising for different species is that mutations can arise and get fixed in the populations giving rise to these species. These would be considered substitutions and it is this substitution process that a Markov chain models.

Two other types of commonly occurring evolutionary events that affect sequences are insertions and deletions. Consider the Human and Rabbit sequences in Table 1.1, for illustration. The '-' at site 9 is a gap and arose because of either an insertion or a deletion. For instance sites 7-10 might have been `CCGC` in the ancestor HB of Human and Rabbit and then, in the lineage leading to Rabbit, the G was lost (a deletion event). Alternatively, sites 7-10 might have been `CCC` in HB and then a `G` was gained in the Human lineage, leading to `CCGC` (an insertion event). It is impossible to know which occurred, so the event is sometimes referred to as an indel event. Although considerable effort has been devoted to modeling such events (eg. Miklós et al. 2009) the models often involve approximations and are only computationally feasible for short sequence alignments and a small number of sequences. Consequently, in most analyses indels, or gaps, are treated as missing data.

The continuous-time Markov chain substitution model along edges is usually assumed to be a stationary, time-reversible Markov chain. The Markov chain is characterized by its rate matrix Q. Given that the ancestral nucleotide is i, the probability, $P_{ij}(t)$, that at time t it is j gives the ijth entry of the substitution matrix, $P(t)$. The substitution matrix is related to the rate matrix through the matrix exponentiation $P(t) = \exp[Qt]$ (Sheldon 1996).

The rate matrix (Wu 2010) given in Eq. (1.1) is the most general time-reversible nucleotide rate matrix, and is referred as the general time-reversible model (GTR). The general time-reversible (GTR) model of DNA substitution is the most general neutral, independent, finite-sites, time-reversible model possible. It was first described in a general form by Tavaré (1986).

$$Q = \{q_{ij}\} = \rho * \begin{bmatrix} \cdot & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \cdot & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \cdot & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \cdot \end{bmatrix} \tag{1.1}$$

This substitution model was used for the HIV real data analysis considered in this thesis.

If $a = c = d = f = 1$, $b = e = \kappa$, the GTR model reduces to the HKY85 model, which was proposed by Hasegawa, Kishino and Yano (1985). It has the following rate matrix:

$$Q = \{q_{ij}\} = \rho * \begin{bmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{bmatrix} \tag{1.2}$$

where κ is the transition/transversion ratio (ratio of rate of transition and transversion per site) parameter. This substitution model was used in simulations.

Although most of our analyses are of nucleotide data, we consider amino acid data sets in some of our real data examples. Amino acid alignments are used for protein-coding genes and effectively recode each triple of nucleotides into its corresponding amino acid. Site 2 in an amino acid alignment thus corresponds to sites 4-6 in the corresponding nucleotide alignment. Although some information is lost in converting nucleotide alignments to amino acid alignments, amino acid data tends to be less affected by model misspecification when dealing with longer evolutionary time scales. In any case, the model assumptions 1-3 are also used for amino acid models but the state space is now 20-dimensional because there are 20 amino acids. Because of the larger dimension of the state space, rather than estimating rate matrix parameters, empirical exchangeabilities derived from large data bases are used. The rate matrix entry Q_{ij} is related to the fixed empirical exchangeability S_{ij} via the equation $Q_{ij} = S_{ij}\pi_j$ where the stationary frequencies are usually estimated by the frequency with which they occurred in the alignment. The empirical exchangeabilities used in this thesis are the JTT rate matrix (Jones, Taylor, and Thornton 1992), the LG matrix (Le and Gascuel

2008) and the mtREV matrix (Adachi and Hasegawa 1996). The usual nomenclature for describing such models, which we use in the real data sections, is for instance, JTT+F or mtREV+F if the stationary frequencies are derived from the data and JTT or mtREV if the stationary frequencies come from the database.

To allow evolutionary processes to vary over sites, conventional models use mixtures. One such adjustment that we consider throughout the thesis is the gamma model of Yang (1994) that allows rates of evolution to vary across sites according to a discretized gamma distribution. In our examples, the model has $G = 4$ rates $r_1(\alpha), \ldots, r_4(\alpha)$ each of which arise with equal probability, $1/G$. The rate $r_j(\alpha)$ is the conditional mean of a $\Gamma(\alpha, 1/\alpha)$ random variable T, conditional upon T being between the $(j-1)/G$th and jth quantile of the $\Gamma(\alpha, 1/\alpha)$ distribution. Under this model, rates for sites are iid from the discretized gamma distribution. The substitution matrix for a site that has drawn rate $r_j(\alpha)$, is then $P(r_j(\alpha)t)$ instead of $P(t)$. The nomenclature adjustment for a gamma rates across sites model is to add a +G. So, for instance, JTT+F+G denotes a model where the JTT exchangeability is used, frequencies are estimated from the data set and a gamma rates-across-sites process is fit.

A final extension of base models considered in one of the real data examples, is to allow stationary frequencies to vary over sites via a mixture model. The model assumes that one frequency vector, π_c, in a fixed finite set of frequency vectors. Here the π_c were estimated as frequently occurring at sites in a large database (Le and Gascuel 2008). The weights or probabilities of the π_c occurring at sites are estimated from the data at hand. The nomenclature here is to add a C60 if, for instance, there were 60 frequency classes. For instance, LG+C60+F+G.

Throughout this thesis we denote a tree as τ, edge-lengths as t and other parameters as θ. The data for a site i, for instance `CCA` at site 10 in Table 1.1, is usually denoted x_i.

1.2 Maximum Likelihood Estimation

Using substitution probabilities for the Markov chain at a site (the change of a C $\rightarrow$ G for example), the Maximum Likelihood (ML) method infers the tree for which

the data was most probable. A tree that requires many substitutions to explain the observed tip data will usually be assigned a lower probability.

Evolution is assumed conditionally independent along lineages and usually assumed according to Markov Chain. This leads to a probability for the site pattern in Figure 1.1 as

$$\pi_G P_{GC}(t_{HB}) P_{CC}(t_H) P_{CC}(t_B) P_{GG}(t_R) \tag{1.3}$$

where π_G is the probability of having G as the root of the tree.

We don't actually know that the ancestral states are G & C, but we can calculate the product above for any choice x_R, x_{HB}:

$$J(x_R, x_{HB}) = \pi_G P_{x_R, x_{HB}}(t_{HB}) P_{x_{HB}C}(t_H) P_{x_{HB}C}(t_B) P_{x_R G}(t_R) \tag{1.4}$$

Since the only thing we observe is the tip data $x_7 = \text{CCG}$, the probability of the observed data is:

$$p(x_7; \epsilon, t, \theta) = \textstyle\sum_{x_R \in \{A,C,T,G\}} \sum_{x_{HB} \in \{A,C,T,G\}} J(x_R, x_{HB}) \tag{1.5}$$

Here we explicitly indicate dependence on the tree τ, edge-lengths t, and other parameters involved in the substitution process, θ. Assuming independent evolution across sites, the likelihood can be calculated after repeating the process above for all sites, as

$$L(\tau, t, \theta) = \prod_{k=1}^{n} p(x_n; \tau, t, \theta) \tag{1.6}$$

Equation (1.5) can get quite a bit more complicated with more taxa; A sum has to be included for each internal node. With m taxa there are $m-2$ internal nodes, so $4^{(m-2)}$ terms need to be summed over. This becomes prohibitive with many taxa. A pruning algorithm (Felsenstein, 1981) is available to reduce the number of terms summed over to something that is linear in the number of species, m.

1.3 Phylogenetic ML Tests

1.3.1 The Two-Tree Problem

Consider a two-tree comparison $H_o : \tau = \tau_o$ vs $H_A : \tau = \tau_1$, for two fixed trees a one-sided test is considered testing whether there is a significant evidence against tree τ_0 and in favour of tree τ_1. KH and Chi-square (discussed later below) are two tests that consider only two trees at a time, which can create a selection bias, as we explain below.

1.3.2 Confidence Set of Trees

A confidence set of trees C is a random set satisfying that $P(\tau_0 \in C) \approx 1 - \alpha$, where τ_o is the true tree.

A confidence set is said to be conservative if $P(\tau_o \in C) > 1 - \alpha$. The advantage of such a set is that it is likely to contain the true tree. The disadvantage is that it often contains too many trees.

There is a duality or 1-1 correspondence between tests and confidence sets. If $p(\tau_o)$ is a p-value for a test of the null hypothesis $H_o : \tau = \tau_o$ and can be applied for any choice of τ_o, then a $(1 - \alpha) \times 100\%$ confidence set, C, can be constructed as the set of τ_o with $p(\tau_o) \geq \alpha$.

We consider a number of settings where there are multiple true trees due to some subset of internal edge-lengths being set to 0. As an extreme example, if all of the internal edge-lengths are set to 0 in the true tree, then every tree topology is true. For multiple true trees, coverage is the average probability, over all true trees, that the true tree is contained in the confidence set. It can be approximated by what we refer to as average observed coverage over simulations, which is the average, over true trees, of the observed coverage: the proportion of simulated data sets for which the true tree was in the confidence set.

1.3.3 Selection Bias

One possible way of using a two-tree test to construct a confidence set is to apply the approach above with $p(\tau_o)$ for the test of $H_o : \tau = \tau_o$ vs $H_A : \tau = \hat{\tau}$, where $\hat{\tau}$ is an estimated tree (usually the ML tree). The problem with this approach is that, for two tree tests, τ_1 is supposed to be fixed in advance. In this approach, $\hat{\tau}$ is *selected* based on the data. When $\hat{\tau}$ is selected in this way it becomes too easy to reject H_o, resulting in $P(\tau_o \in C) \ll 1 - \alpha$. Tests that will be described to test the two-tree problem are the KH test and Chi-square test detailed in the tests section below.

1.3.4 Bootstrap Support

The bootstrap was originally developed by Efron (1982). Its use in phylogenetics was initiated by Felsenstein (1985). With bootstrap support we resample sites, observational units, at random with replacement. The bootstrap principle used in obtaining approximations is that parameter estimates from the original data should be treated as if they were true and estimates from bootstrap samples as estimated quantities. For instance, a standard use of bootstrapping is to approximate the distribution of $\hat{\theta} - \theta$ by the empirical distribution of $\hat{\theta}^* - \hat{\theta}$ over repeated bootstrap samples, where $\hat{\theta}^*$ is a bootstrap estimated quantity. Bootstrap support for a tree is defined as the proportion of times that the tree was estimated over repeated bootstrap sample data sets. Large bootstrap support means that there is little sampling error in the tree estimate.. Bootstrap support is a non-standard use of bootstrap principles.

Because estimation and even likelihood calculation is computationally intensive in phylogenetics, an approximation to full bootstrapping was developed referred to as Resampling Estimated Log-Likelihoods, or **RELL** (Kishino, Miyata, and Hasegawa 1990). Rather than re-estimating parameters for each bootstrap, the approach uses the estimated parameters from the original data, thus effectively resampling the maximized site likelihoods for the original data.

To illustrate, we contrast what would be treated as the maximized log likelihoods for a bootstrapped data set using the RELL approach or the standard bootstrap:

$x_1, \ldots, x_n$ original data

$x_1^*, \ldots, x_n^*$ after resampling with replacement (bootstrap data)

$l(x_1^*, \tau, t, \theta), \ldots, l(x_n^*, \tau, t, \theta)$ log-likelihood for θ contributions from first to nth observation.

The ML estimates, $(\hat{\theta}, \hat{t})$, and ML bootstrap estimates, $(\hat{\theta}^*, \hat{t}^*)$, are:

$$(\hat{\theta}^*, \hat{t}^*) = \arg\max_{\theta,t} \sum l(x_i^*, \tau, t, \theta)$$

$$(\hat{\theta}, \hat{t}) = \arg\max_{\theta,t} \sum l(x_i, \tau, t, \theta)$$

For the standard bootstrap, the maximized log likelihood for τ is:

$$\max l = \sum l(x_i^*, \tau, \hat{t}^*, \hat{\theta}^*)$$

whereas for RELL bootstrap it is:

$$l = \sum l(x_i^*, \tau, \hat{t}, \hat{\theta})$$

Because RELL does not require re-estimating parameters for each bootstrapped data set it is much faster and computationally cheaper than the standard bootstrap.

1.4 Tests

In the subsections that follow we describe the existing tests that will be considered in this thesis. Two other likelihood-based tests are available. The SOWH test is a parametric bootstrap test that was described in detail in Goldman, Anderson, and Rodrigo (2000) and is named after its authors (Swofford et al. 2004). The Single Distribution Nonparametric Bootstrap (SDNB) test is another bootstrap-based approach defined in Shi et al. (2005). Owing to the computational expense of bootstrapping in phylogenetics, these tests are not as frequently reported as some of the tests below. Moreover, the computational expense of these tests makes it difficult to include them in the simulations reported here, particularly in 8-taxon tree simulations which included a large number of trees.

1.4.1 KH

The KH test (Kishino and Hasegawa 1989) is a two-tree test that is based on a reformulation of the hypotheses of interest in terms of the differences between site log likelihoods for the two trees.

$$d_i = l(x_i; \tau_1, \hat{t}, \hat{\theta}) - l(x_i; \tau_2, \hat{t}, \hat{\theta}) \tag{1.7}$$

Here $\hat{\theta}$ represents any parameters other than edge-lengths that had to be estimated (such as substitution rate etc.), and $l(x_i; \tau_j, \hat{t}, \hat{\theta})$ denotes the log probability of the site pattern x_i for the i^{th} site in the sequence alignment.

If the two trees explain the data equally well, then on average, the site log likelihoods for one should not tend to be larger than another. On the other hand, if Tree 1 is correct and Tree 2 is not, then we expect that the observed data will tend to be more likely under Tree 1 than Tree 2. This leads to the hypotheses

$$H_0 : E[d_i] = 0 \qquad \text{against} \qquad H_A : E[d_i] > 0 \tag{1.8}$$

Treating the d_i as independent and identically distributed, by the Central Limit Theorem, for n large, the approximate distribution of $\bar{d}$ is $\bar{d} \sim N(0, s_d^2/n)$ where s_d^2 denotes the sample variance of the d_i. The KH test computes the p-value

$$p = P(Z > n\bar{d}) \tag{1.9}$$

to check whether the KH test statistic is larger than expected from a $N(0, ns_d^2)$ distribution.

In reality the d_i are not iid, so the z-test is not strictly justified (Susko 2014), and indeed the test turns out to be highly conservative, as we will illustrate.

A later variation to the test replaced $N(0, s_d^2/n)$ with a bootstrap distribution (Kishino and Hasegawa 1989). However, because this test tends to give very similar p-values to the original KH test, this thesis will focus on the original KH test.

1.4.2 SH

The Shimodaira–Hasegawa test, SH (Shimodaira and Hasegawa 1999) was created as a modification to the KH test. The KH test was originally designed to compare a pair of trees, but ended up being used to make inferences about multiple null hypothesis trees. Moreover, in the absence of a clear alternative hypothesis tree, the ML tree is used as the alternative tree in the log-likelihood ratio. As a tree that is not fixed in advance but rather selected based on the data, such a choice can induce a selection bias. The SH modifies KH to correct for this selection bias.

Let $L_1, \ldots, L_M$ denote the maximized log likelihoods for a fixed set of M candidate trees; in most of our examples, these would be all trees. Let $\hat{\tau}$ denote the maximum-likelihood topology. We have to consider the effect of selection of $\hat{\tau}$ to derive the distribution of $L_{\hat{\tau}} - L_\tau$.

A method to test multiple-comparison of trees to statistical model selection was introduced by Shimodaira (1993, 1998) using the following steps:

1. Calculate the test statistics

$$T_\tau = \max\{L_1 - L_\tau, \ldots, L_M - L_\tau\} = L_{\hat{\tau}} - L_\tau \qquad \text{for} \quad \tau = 1, \ldots, M \quad (1.10)$$

2. Create N bootstrap replicates of $(L_1, \ldots, L_M)$ for an $M \times N$ matrix for bootstrapping, and use the RELL method described in Section 1.3.4.

3. *Centering* - in order to overcome the issue where bootstrapping is effectively like simulating from whatever the true process was, which might correspond to the null or might correspond to the alternative hypothesis, centering the data forces it to correspond to the null hypothesis. SH corrects by centering the log likelihoods for each tree so that the mean difference in log likelihoods for any pair of trees is 0 under the null hypothesis. This is done by subtracting from each element in a row the average of the row to create $\widetilde{R}_{\tau\times i}$, a replicate of L_τ generated under the least favorable configuration, LFC (which is the star tree, explained more later).

$$\widetilde{R}_{\tau\times i} = \widetilde{L}_{\tau\times i} - \frac{1}{N}\sum_{j=1}^{N} \widetilde{L}_{\tau\times j} \quad (1.11)$$

Here i indexes the bootstrap replicate.

4. Now for each column in $\widetilde{R}_{\tau i}$ calculate a replicate of the vector T_τ.

$$\widetilde{S}_{\tau i} = \max\{\widetilde{R}_{1i} - \widetilde{R}_{\tau i}, \ldots, \widetilde{R}_{Mi} - \widetilde{R}_{\tau i}\} \tag{1.12}$$

5. Calculating p-value: For each row $\tau = 1, \ldots, M$ in $\widetilde{S}_{\tau i}$ count the number of entries that exceeded T_τ:

$$P_\tau = \frac{\text{number of } \{\widetilde{S}_{\tau i} > T_\tau\}}{N} \tag{1.13}$$

Note that for the case that there are two true trees in a set, we should expect to see the same results for KH vs SH, however, due to Monte Carlo we see some small variations.

1.4.3 AU

The Approximately Unbiased test (AU) for regions with general smooth boundaries Shimodaira (2002) was developed based on the theory of Efron, Halloran, and Holmes (1996), Efron and Tibshirani (1998).

The test was based on the results about bootstrap support for regions in normal mean problems. In this setting the data are $Y \sim N(\mu, \Sigma)$. The alternative hypothesis of interest is that $\mu \in R$ for some region R. The bootstrap support for a region is defined as the proportion of bootstrap samples for which the bootstrap mean is in the region. Efron and Tibshirani (1998) showed that bootstrap support for regions with smooth boundaries are approximately the same as 1 minus the p-value for a test of the null hypothesis that the true mean is not in the region R. The AU test was developed as a higher-order correction to this p-value. Strictly speaking, the motivation below applies only to this normal means setting. The hope was that the results would extrapolate to bootstrap support in phylogenetics. However, it was later shown that bootstrap support does not give an approximate p-value in phylogenetic settings (Susko 2009). Thus the properties of the AU test are not well understood in this setting. Nevertheless, the AU test is frequently used in practice. Thus we include it among our comparator methods.

AU's methodology is to compare bootstrap probability (BP) of regions:

1. **Bootstrapping** is repeated with varying sample sizes, some of which differ from that of the original data.

2. For a given region of interest and each sample size, count the number of bootstrap samples for which the bootstrapped mean is in the region to obtain **BP values** for different sequence lengths.

3. The AU test calculates the approximately unbiased p-value from the **change in the BP values** along the changing sample sizes, based on the motivation below.

It follows from the Corollary of Efron (1985) that the $p-$ value, 1-AU, where AU is defined as:

$$AU = 1 - \Phi(d - c) \tag{1.14}$$

is an $O(n^{-3/2})$ approximation to a valid p-value for the alternative hypothesis that the true mean is in the region of interest. Here d is the *signed Euclidean distance* between the sample mean for the data and the projection of that sample mean onto the boundary of the region. It is the implicit test statistic for the procedure. The quantity c is related to the curvature at the boundary, and $\Phi(\cdot)$ denotes standard normal cumulative distribution function. Although d and c can, in principle be calculated, the AU procedure instead utilizes their relationship to BP to approximate them indirectly. Shimodaira uses an argument from Efron and Tibshirani (1998) that

$$BP \approx 1 - \Phi(d + c) \tag{1.15}$$

Shimodaira (2002) notes that if BP is calculated for a fraction r of the original sample size, then the resulting BP, BP_r should satisfy

$$BP_r \approx 1 - \Phi(d\sqrt{r} + c/\sqrt{r}) \tag{1.16}$$

Since BP_r and BP can be calculated, this gives two equations in two unknowns which can be solved to approximate d and c which can then be substituted into (1.14). The actual procedure used in Shimodaira is a little more complicated and uses BP_r for multiple choices of r to approximate d and c. 1-AU from (1.14) approximates an exact p-value up to $O(n^{-3/2})$ which is considered third-order accurate; the usual error in approximation, for instance, from (1.15) is $O(n^{-1/2})$. Often what is reported is AU, which is referred to as the support or confidence level for the region, not to be confused with the level, α, of a $(1-\alpha)\times 100\%$ confidence set.

Susko (2009) argues that BP is not first order correct when used in a phylogenetic setting and that, consequently, the AU test might not have even approximately correct coverage. Thus the properties of the AU test are not well understood in this setting. Nevertheless, the AU test is frequently used in practice. Its actual coverage properties are an issue that will be discussed in the next chapter.

1.4.4 Chi-square

The chi-square test was developed in R. Ota et al. (2000) for the case where two trees differ by a single edge. That is, when a single edge-length in Tree 1 is set to 0, it gives a special case of Tree 0.. A conservative extension that applies more generally was developed in Susko (2014).
Consider the topology τ_0, compared with the alternative τ_1, which is the tree that is believed to be the true tree:

$$H_0 : \tau = \tau_0 \qquad\qquad H_A : \tau = \tau_1 \tag{1.17}$$

Let the log likelihood for the best topology τ_1 be l_{τ_1} and l_{τ_0} for τ_0. Then the log ratio statistic is $\quad lrs = 2\{l_{\tau_1} - l_{\tau_0}\}$. The distribution of lrs, under the null hypothesis, is treated as being chi-square distributed with *df* degrees of freedom in Susko (2014), where the *df* is the number of branches that had to be adjusted to 0 in order for the two topologies to be equal.

In Figure 1.2, topology τ_0 on the left and topology τ_1 on the right differ in that the placements of taxa 2 and 4 have been reversed. Consequently, the edges ab in red

correspond to two different splits of the taxa into two groups.

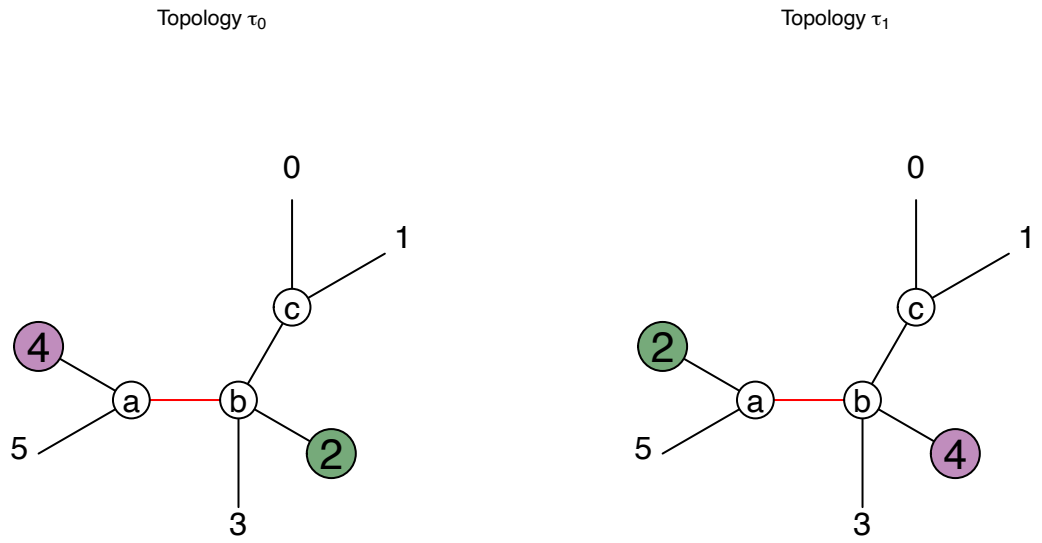

Figure 1.2: Equivalent Topologies

Defining topology τ_0' as the topology τ_0 with the only change being internal edge length ab set to 0, gives the tree in Figure 1.3 where the upper branch with tips [1, 0] are separated from the rest of the tips [2, 3, 4, 5]. The same is true for τ_1; topology τ_1' is identical to τ_0'.

In the resulting tree we see that all tips [2, 3, 4, 5] are now branched from the same node a without distinction between τ_0' and τ_1'.

A tree with internal edge-lengths set to 0, call it T_0, that makes the topologies equivalent is on the boundary between the parameter space for topology τ_0 and the parameter space for topology τ_1. Because T_0 is on the boundary of the parameter space it is the appropriate tree for computing p-values. In a similar sense, when we test $H_0 : \mu \leq 0 \quad$ against $\quad H_A : \mu > 0$, we calculate p-values using the value $\mu = 0$ on the boundary between the null $(-\infty, 0]$ and alternative $(0, \infty)$ parameter spaces. As discussed in Susko (2014), the p-value, p, can in principal be calculated using the distribution of lrs under the null hypothesis. That distribution is a weighted mixture of chi-squares, $\sum_{j=0}^{df} w_j P[\chi_j^2 \leq x]$, where df is the number of edge-lengths set to 0 in T_0. The weights, w_j, which are positive and sum to 1, are not easily calculated. But the resulting p-value, p, satisfies that

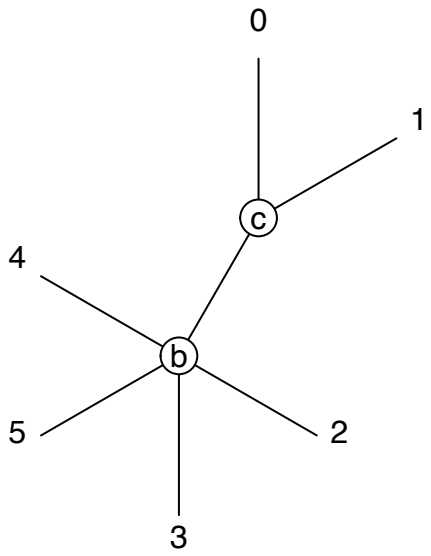

Figure 1.3: Setting Internal Edge To 0

$$p = \sum_{j=0}^{df} w_j P[\chi_j^2 > lrs] \leq \sum_{j=0}^{df} w_j P[\chi_{df}^2 > lrs] = P(\chi_{df}^2 > lrs) = p_{\chi^2} \tag{1.18}$$

for large sequence. The quantity p_{χ^2} can be calculated. Since $p \leq p_{\chi^2}$, then the probability

$$P(p_{\chi^2} < \alpha) \leq P(p < \alpha) \tag{1.19}$$

Since approximate p-values satisfy that $P(p < \alpha) \approx \alpha$, the probability of false rejection satisfies that

$$P(p_{\chi^2} < \alpha) \leq P(p < \alpha) \approx \alpha \tag{1.20}$$

Thus as a two-tree test, the chi-square test is conservative because p_χ^2 always larger than p. We illustrate the calculation leading to this conclusion in Figure 1.4.

The chi-square confidence sets are constructed using the chi-square test for two trees, taking $\tau_1 = \hat{\tau}$ the ML topology. For any fixed topology τ_0 being considered for inclusion in the confidence set,

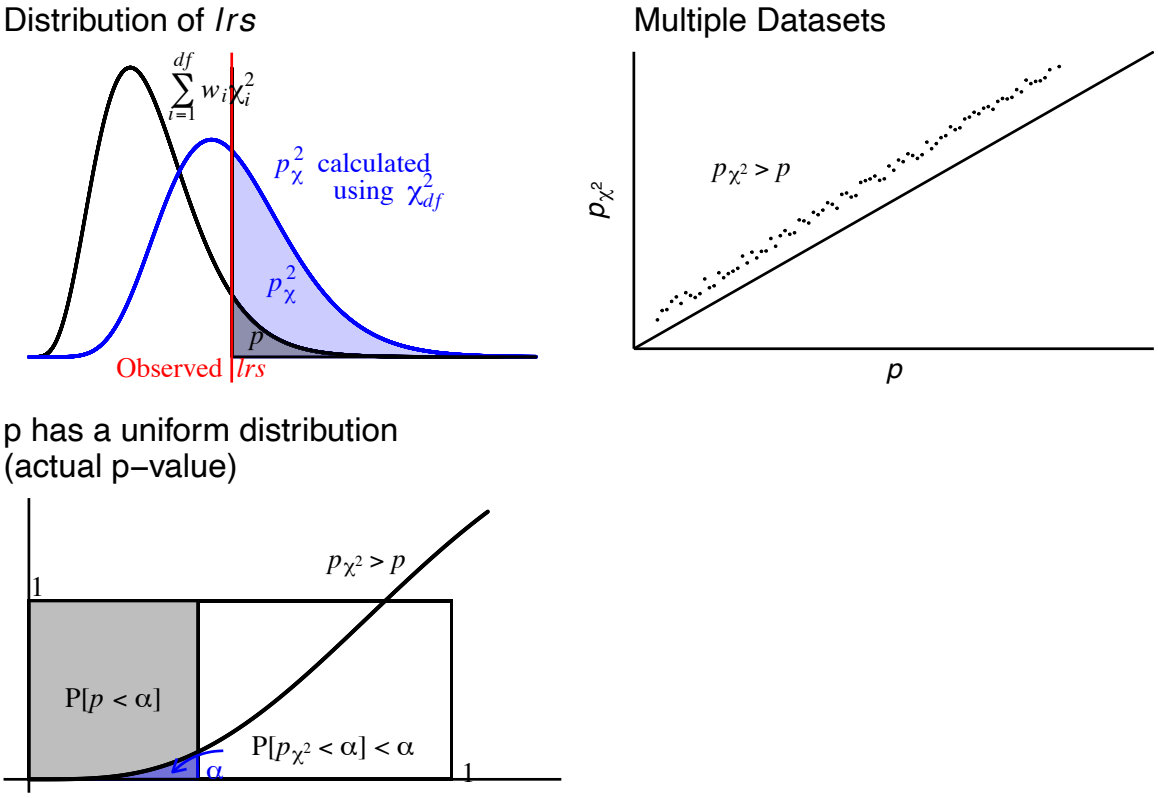

Figure 1.4: Illustrating why the chi-square test is conservative

1. Obtain the consensus tree, T_0, with just enough edge-lengths set to 0 to make τ_0 and τ_1 equivalent. The only information required is setting df = the number of edge-lengths set to 0
2. Calculate p_{χ^2} with this degrees of freedom df and topology $\tau_1 = \hat{\tau}$
3. Include τ_0 in the confidence set if $p_{\chi^2} \geq \alpha$

Because the ML topology is used as the alternative tree, a selection bias is expected and as a confidence set procedure, the chi-square test is no longer guaranteed to be conservative. In fact, in our simulations it showed to be rather aggressive in comparison to the other tests.

1.5 Thesis Structure

In this thesis we compare four existing models, KH, SH, AU and Chi-square introduced in Chapter 1 as well as two new methods: AU-corrected and a Bonferroni correction to Chi-square, derived in Chapter 2, over multiple metrics to assess favourability. We define coverage and mean set size of a resulting test, and discuss the complexity of calculating a confidence interval for coverage due to correlation of trees appearing in a resulting set. We slice and dice the data under different circumstances of tree type for both 6-taxa and 8-taxa, for each type setting a different number of internal edge lengths to zero, as well as looking at the difference between two adjacent zero-length internal edges versus non-adjacent. We also compare varying internal edge lengths (the non-zero ones of course) to observe the effect of the length on the tests' abilities to identify trees, And lastly we compare confidence levels of the more common 0.05 level as well as 0.1 and 0.01, with the expectation that the tests may behave more conservatively when setting a higher threshold of confidence level. We contrast high coverage performance of tests with their ability to keep the resulting set size of trees closer to the expectation of the tree type, and also discuss the significance of having a coverage higher than the confidence level set. Chapter 3 shows the detailed results of all of our findings, showing side-by-side output for the tests. In Chapter 4 we apply the same tests to four real-world datasets. Chapter 5 concludes the thesis by providing a summary and discussion of the findings as we outline the common behaviour observed for each of the tests.

Chapter 2

Methodology

2.1 Bonferroni Correction

Bonferroni here refers to a new approach of calculating confidence sets that utilizes the Chi-square test of two trees and adjusts for selection bias. We use the ML tree to determine a relevant set of comparisons. It is related to but not quite the same as conventional Bonferroni corrections for multiple tests. Let T_0 be the tree with edge-lengths set to 0 to make $\hat{\tau}$ equivalent to τ_0. Let $A(T_0)$ be the set topologies that are consistent with T_0 .

In the following example we look at comparing topology τ_0 with the ML topology $\hat{\tau}$:

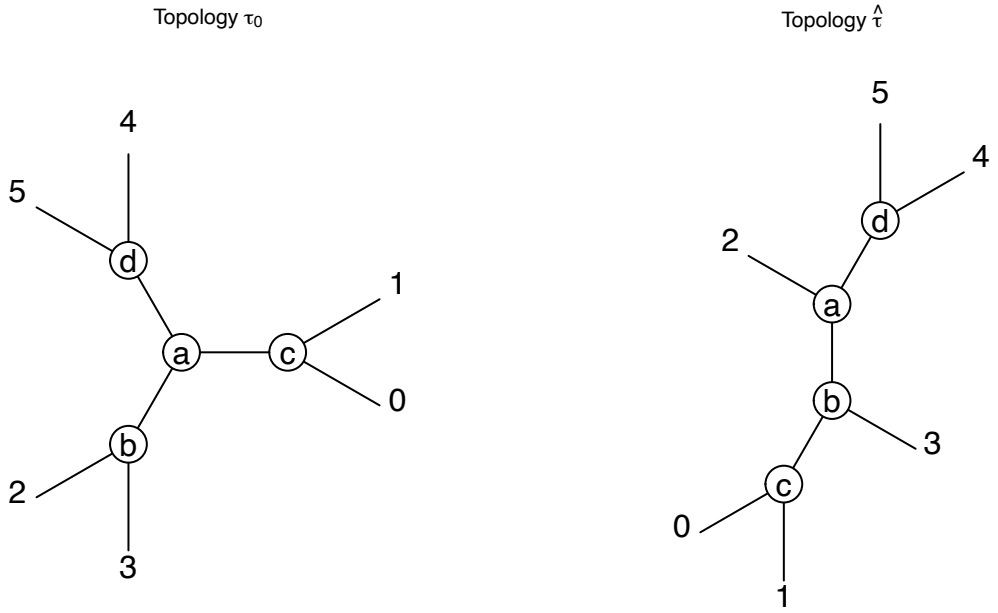

Figure 2.1: Bonferroni - Example Topology

In order to make the two equivalent we need to shorten the internal edge length ab to 0. That results in Figure 2.2 consensus tree (topology T_0):

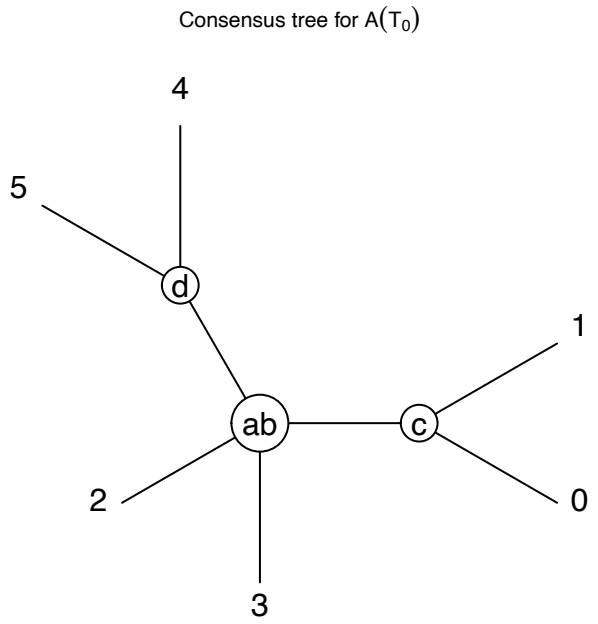

Figure 2.2: Bonferroni - Setting Internal Edge To 0

The next step is to list all topologies that would result in topology T_0 above if we shorten only a single internal edge length.

The resulting set $A(T_0)$ is given in Figure 2.3.

Let $n(A)$ be the number of trees in $A(T_0)$ and let df be the number of edges that were set to 0 in T_0. Then the Bonferroni confidence set includes τ_0 in the confidence set if

$$1 - (1 - P[\chi^2_{df} > 2(l_{\hat{\tau}} - l_{\tau_0})])^{n(A)} \geq \alpha \tag{2.1}$$

The algorithm to get the Bonferroni confidence set calculates the p-value for each choice of τ_0. Suppose that the current tree being considered is τ_0 in Figure 2.1. The procedure to get a p-value for τ_0 would be as follows:

1. Determine the ML tree (given in Figure 2.1 for the example)

    1.1. Determine the log-likelihood for τ_0 and $\hat{\tau}$ and use these to determine $lrs = 2\{l_{\tau_1} - l_{\tau_0}\}$. As a hypothetical example, we suppose this ended up being $\quad lrs = 2.3$

2. Determine the df as the number of edge-lengths required to make $\hat{\tau}$ and τ_0

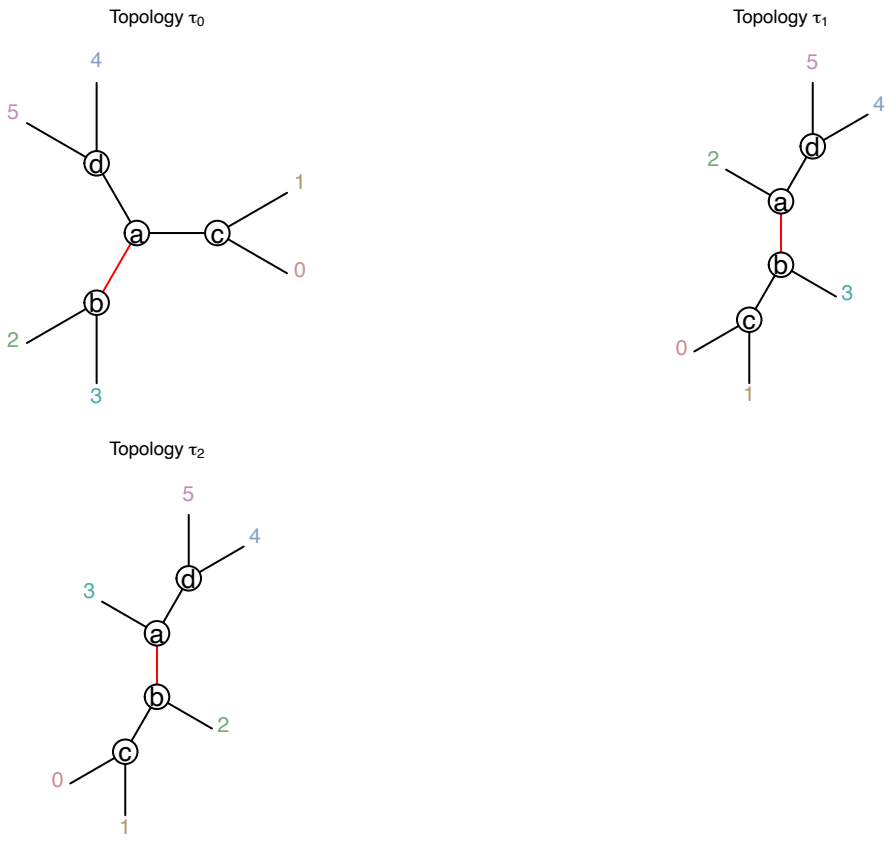

Figure 2.3: Bonferroni - The trees in $A(T_0)$

equivalent. In the example, the consensus tree making τ_0 and $\hat{\tau}$ equivalent is given in Figure 2.2 where we see that $df = 1$

3. Determine the set $A(T_0)$. In the example $A(T_0)$ set is given in Figure 2.2

4. Determine $n(A)$, the number of trees that are compatible with T_0. In the example $n(A) = 3$ as in Figure 2.3

5. Calculate the p-value using Equation (2.1). In the example, plugging in the values from previous steps we get a p-value of 0.9999673.

To motivate the Bonferroni confidence set we make two approximations:

C1. We treat $A(T_0)$ as the set of plausible trees that the ML estimate could have come from. The idea here is that if the null hypothesis is correct and T_0 is the correct tree, then in theory, by consistency of tree estimation, if sequence length is large enough, the ML tree will be in $A(T_0)$ most of the time.

C2. We set df to the number of 0 edge-lengths in T_0 for the test of

$$H_0 : \tau = \tau_0 \quad \text{vs} \quad H_A : \tau = \tau_i \quad \text{for some} \quad i \quad \text{in} A(T_0) \tag{2.2}$$

and use the chi-square test. Note that the usual chi-square test would set df to the number of 0 edge-lengths in the consensus tree for τ_0 and τ_i rather than the consensus tree for τ_0 and $\hat{\tau}$.

The Bonferroni approach seeks to determine whether τ_0 is in the confidence set by checking whether any of the tests

$$H_0 : \tau = \tau_0 \quad \text{against} \quad H_A : \tau = \tau_i \quad \text{where} \quad i \in A(T_0) \tag{2.3}$$

rejects. If not, τ_0 is in the confidence set. Assuming approximations C1 and C2 are true, then the i^{th} test rejects at the α -level if

$$P_i = P[\chi^2_{df} > 2(l_{\tau_i} - l_{\tau_0})] < \alpha \tag{2.4}$$

Thus at least one of the tests rejects at the α -level if

$$\alpha > \min\ P_i$$

$$= \min\ P[\chi^2_{df} > 2(l_{\tau_i} - l_{\tau_0})]$$

$$= P[\chi^2_{df} > \max_i 2(l_{\tau_i} - l_{\tau_0})]$$

$$= P[\chi^2_{df} > 2(l_{\hat{\tau}} - l_{\tau_0})]$$

The difficulty with this approach is that there are multiple comparisons. If each of the tests has a chance approximately α of rejecting, then the probability that at least one rejects is greater than α. The usual Bonferroni correction replaces min P_i with $n(A) \min P_i$ and rejects when $n(A) \min\ P_i < \alpha$. The resulting corrected test has type I error probability at most α. The Bonferroni correction that we use treats the tests as independent and replaces min P_i with

$$1 - (1 - \min\ P_i)^{n(A)} \tag{2.5}$$

where $n(A)$ is the number of tests. When min P_i is small, which is the main case of interest, this expression is approximately the same as $n(A) \min\ P_i$. The value in (2.5) has the advantage that it is always in (0,1).

To establish that little difference arises from using the independence correction in (2.5) rather than the actual Bonferroni correction, consider the result of an $\alpha = 0.05$ level test for various choices of $n(A)$. The Bonferroni test rejects when $n(A) \min P_i < 0.05$ or equivalently when $\min P_i$ is less than the Bonferroni threshold $0.05/n(A)$. Similarly, the independence approach rejects if (2.5) is less than 0.05 which can be shown to be equivalent to $\min P_i$ being less than the independence threshold $1 - 0.95^{1/n(A)}$. Figure 2.4 gives the two thresholds for various choices of $n(A)$.

We would only get a different test result when $\min P_i$ is between the Independence and Bonferroni thresholds.

The Bonferroni approach makes two approximations. The first, C1, if used in isolation would cause the approach to be slightly anti-conservative. In reality any tree is possible, so the p-values should in principle be calculated from the set of all trees in place of

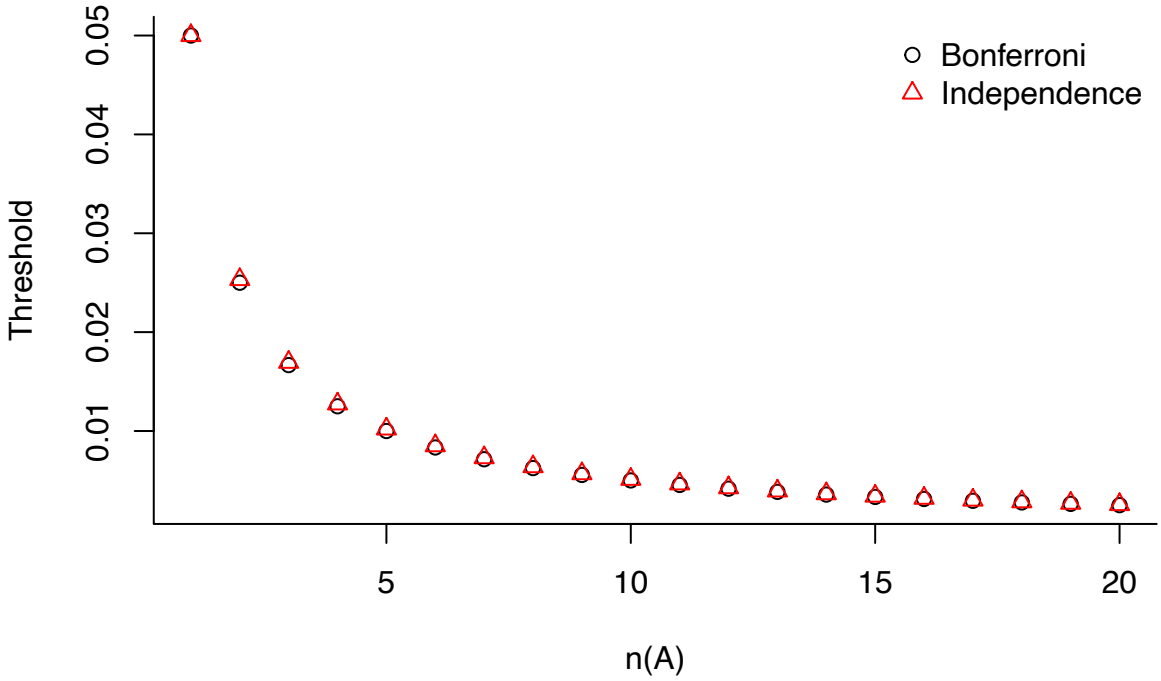

Figure 2.4: Thresholds for a 0.05 level test

$A(T_0)$. If that were to be done with the chi-square approximation it would lead to an approach that was too conservative.

We did not need the second approximation C2. We could explicitly account for differing degrees of freedom by calculating min P_i where the *df* are different for different tests. In practice this would require explicitly calculating l_{τ_i} for all $i \in A(T_0)$. For larger trees, this is unfeasible. We wanted an approach that only required $l_{\hat{\tau}}$ and the log likelihood, l_{τ_0} for the tree being considered for inclusion in a confidence set.

The second approximation, C2, treats the *df* as the number of edge-lengths in T_0 that were set to 0. In reality for some of the tests, the *df* would be smaller and the *p*-value would be smaller. So some of the P_i used in calculating min P_i are too large. Thus min P_i might be larger than it would have been had the differing degrees of freedom been used. Consequently, C2 results in making the testing procedure more conservative. We don't expect it to be an overly conservative adjustment, however, because the ML tree uses the correct *df*. Because the ML tree gives the largest log likelihood, it will still frequently give the smallest P_i even had the P_i been calculated with the correct degrees of freedom.

In summary, the use of chi-square tests, condition C2 and Bonferroni corrections all result in a conservative confidence set construction procedure. On balance, these are the most important issues, so generally we expect the Bonferroni method will produce confidence sets with coverage greater than $1 - \alpha$.

2.2 AU Correction

The AU test effectively solves (1.16). Equivalently, it solves

$$d\sqrt{r} + c/\sqrt{r} = \Phi^{-1}(1 - BP_r) \tag{2.6}$$

where Φ^{-1} is the inverse of the standard normal distribution function.

Two approaches are given in (Shimodaira 2002) to solve (2.6). Consider a set $r_1, \ldots, r_k$ of multipliers. Let y_i be $\Phi^{-1}(1-BP_{r_i})$, $x_{i1} = \sqrt{r_i}$ and $x_{i2} = 1/\sqrt{r_i}$. The first approach solves for c and d using least squares applied to (2.6) for $r = r_1, \ldots, r_k$. The other approach treats $rnBP_r$ as an independent Binomial, where the probability of success for each trial is defined by (1.16), and uses maximum likelihood to solve for c and d; using optimization since no explicit solution is available.

Because $y_i = \Phi^{-1}(1 - BP_{r_i})$, when $BP_{r_i} \approx 0$, small variations in BP_{r_i} can cause large variations in y_i, making the approach unstable. The quantity BP_r is not directly available from the software we use, so we use $BP = 0$ as a proxy for $BP_r = 0$. Because $BP = 0$ suggests the tree is not plausible, the AU correction is 0 whenever $BP = 0$ and otherwise coincides with AU. Note that one might consider more aggressive AU corrections that are 0 when BP is less than some small threshold.

Figure 2.5 compares the p-values spread between AU and AU Correction for 1000 simulations from a six-taxon tree. It shows AU Correction has a stretched down spread of p-values, meaning lower values than AU's.

2.3 Confidence Intervals for Coverage

We will use simulation to approximate the true coverage of any given confidence set construction procedure by the average percent of true trees that are in the confidence

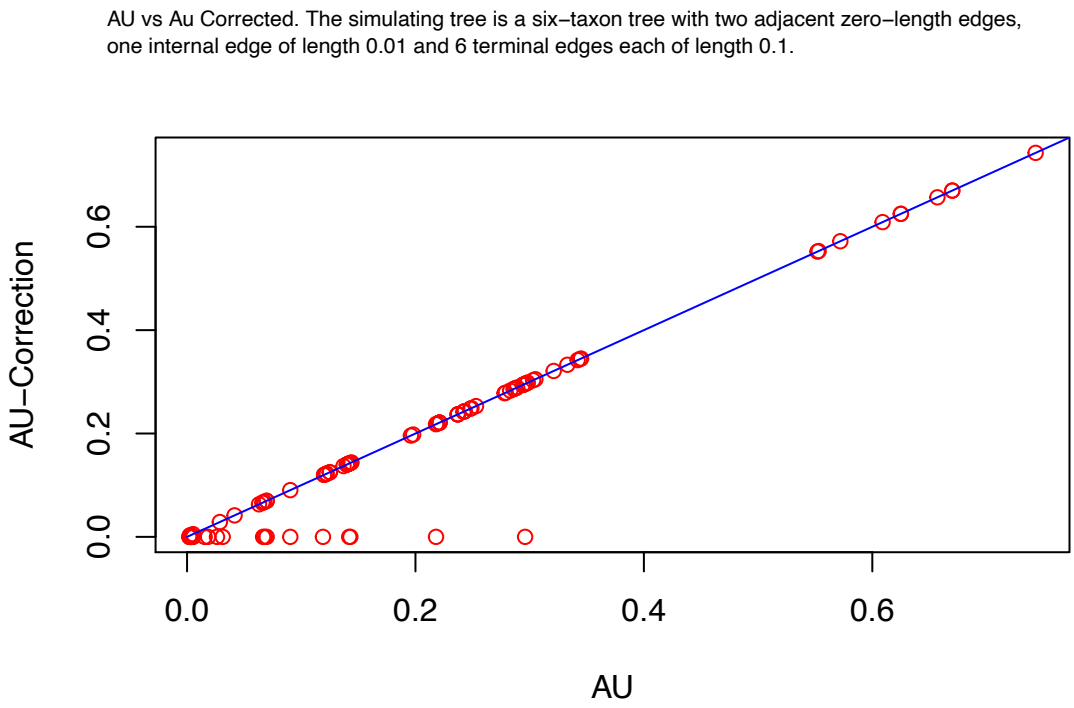

Figure 2.5: AU vs AU Correction p-values

sets over all simulations. We need to adjust for the possibility that there are often multiple true trees due to some edge-lengths being 0 in the simulating model. We ran either 100 or 1000 simulations. Thus it was important to have confidence intervals (CI) for the true coverage.

Let C be the average observed coverage over simulations, then

$$C = \sum_{j=1}^{K} \sum_{b=1}^{B} \delta_{jb}/(KB)$$

where

K = Number of *true* trees,

B = Number of simulations

$$\delta_{jb} = \begin{cases} 1 & \text{if tree } j \in C_b \\ 0 & \text{otherwise} \end{cases}$$

C_b = Confidence set for simulation b

The average observed coverage, C, is an approximation to the true coverage.

Complication: δ_{jb} & $\delta_{jb'}$ are independent, but δ_{jb} & $\delta_{j'b}$ are correlated.

Define

$$p_j = \text{long-run proportion of } C_b \text{ with } j \in C_b \tag{2.7}$$

$$p_{ij} = \text{long-run proportion of } C_b \text{ with both } i, j \in C_b \tag{2.8}$$

Then the covariance is

$$\text{Cov}[\delta_{jb}, \delta_{ib}] = p_{ij} - p_i p_j \tag{2.9}$$

The quantities p_i, p_j can be estimated by observed proportions.

Since δ_{jb} is Bernoulli:

$$\text{Var}[\delta_{jb}] = p_j(1 - p_j) \tag{2.10}$$

$$\text{Var}\left[\sum_{j=1}^{K} \delta_{jb}\right] = \sum_{j=1}^{K} \text{Var}[\delta_{jb}] + 2\sum_{i<j} \text{Cov}[\delta_{jb}, \delta_{ib}] \tag{2.11}$$

$$\text{Var}[C] = \text{Var}\left[\sum_{j=1}^{K} \delta_{jb}\right]/(K^2 B) \tag{2.12}$$

Rather than obtaining a confidence interval for C directly, we obtain a confidence interval for the logistic transformation of C and then transform back to get a CI for C. This ensures that the bounds are in (0,1).

$$g(C) = \log\left[\frac{C}{1-C}\right] = \log[C] - \log[1-C]$$

$$g'(C) = \frac{1}{C} + \frac{1}{1-C} \tag{2.13}$$

Δ -method:

$$\mathrm{Var}[g(C)] = g'(C)^2 \cdot \mathrm{Var}[C] \tag{2.14}$$

Let $[L, U]$ be the lower and upper limits of the 95% CI for $g(C)$, then:

$$[L, U] = g(C) \pm 1.96\sqrt{\mathrm{Var}[g(C)]} \tag{2.15}$$

Then the 95% CI for C is:

$$[g^{-1}(L), g^{-1}(U)] \tag{2.16}$$

2.4 Simulation Settings

We considered two main sets of simulations. One with six taxa and another with eight. For any fixed tree setting, we generated data sets with sequence length 1000. The substitution model used was the HKY model with transition-transversion ratio 2 and frequencies of nucleotides $\pi_A = 0.1$, $\pi_C = 0.2$, $\pi_G = 0.3$ and $\pi_T = 0.4$. Simulations included rates-across-sites variation from an 8-category gamma rates across sites model with $\alpha = 1$. We generated 1000 data sets for each six-taxon simulation setting and 100 data sets for each eight-taxon simulation setting. Internal edge-lengths varied across settings but terminal edge-lengths were always set to 0.1.

To obtain the p-values for each of the tests described above, an open-source phylogenetic reconstruction package, IQ-TREE was used (L. Nguyen et al. 2014). In the words of the product description on its website, IQ-TREE takes as input a multiple sequence alignment and will reconstruct an evolutionary tree that is best explained by the input data. We used the latest version of IQ-TREE available at the time of analysis. All version numbers were greater than or equal to 1.6.8. IQ-TREE returns the Maximum likelihood tree and p-values for the tests KH, SH and AU. More info on IQ-TREE can be found here: http://www.iqtree.org

In our simulation applications of IQ-TREE, for each simulation, taxa and length of internal-edge likelihoods and p-values were obtained for a fitted GTR model with 4 gamma rate categories. The following IQ-TREE command was run:

```
iqtree -s {sequences input file} -z {tree file for the taxa}
 -nt 8 -m GTR+G -n 0  -zb 10000 -au -wsl -mem 10GB -redo > {output file}
```

where the sequence input file looks like:

```
 6 1000
0          ATCGTTGGCCCGGCGTCGGTTGATGTGAGTTCGCG...
1          ATCGTTGGCCCTGCGCCGGCTGATGTGGATTCGCG...
...
```

and represents the edge nodes for the tree, so for example, for the 6-taxon the file contained 6 alignments (0-5) for each of the nodes.

The tree file for the taxa looks like:

```
(3,(2,5),(1,(0,4)));
((2,5),(3,1),(0,4));
(5,(3,2),(1,(0,4)));
(2,(3,5),(1,(0,4)));


...
```

and goes to 105 possible trees for a 6-taxon, and 10395 tree possibilities for the 8-taxon.

The software outputs data from which a p-value dataframe is extracted and saved as a file.

Setting a fixed seed for the data sampling, running the same test n times, and varying the length of internal edges, calculating for each case the p-value for each of the tests, a confidence set was calculated.

The term *set* and *simulation* are used interchangeably throughout this paper as a single simulation produces a set of trees.

2.5 Six-taxon Simulation Settings

The maximum number of internal edge lengths for a 6-taxon tree is 3. We ran $n = 1,000$ simulations for trees with 0, 1, 2 or 3 of the internal edge-lengths set to 0. For each setting we also considered three choices of edge-lengths for the other internal edge-lengths (0.1, 0.01 and 0.001).

The formula for the number of possible trees depends on the number of taxa. For an unrooted tree, $N_u = (2m-5)(2m-7)...1$, where m is the number of taxa. So for six taxa there are 105 possible tree structures to examine in each simulation. For the case of two edge-lengths set to 0, there are two possible cases; one of which the two zero-length internal edges are adjacent, and one when they're not.

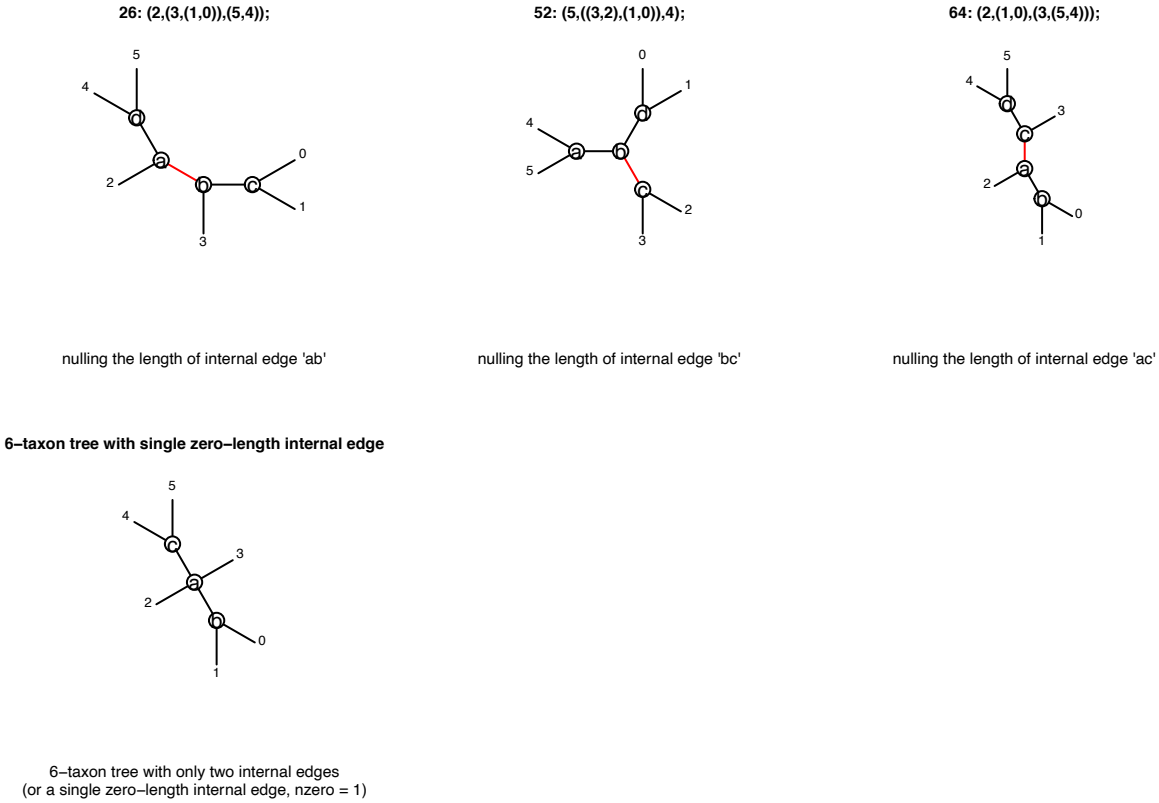

Figure 2.6: The three true six-taxon trees corresponding to the six taxon tree with a single zero-length edge considered in simulations

Settings with zero-length edges can become challenging for ML because otherwise-different structure trees become identical; In Figure 2.6, the six-taxon trees 26, 52 and 64 all become the tree on the bottom when nulling one of each of their internal edge

lengths (in red).

2.6 Eight-taxon Simulation Settings

For eight-taxon trees the maximum number of internal edge lengths is 4. Only $n = 100$ simulations were performed as it is computationally harder to obtain ML estimates for all 10,395 possible trees, and the internal edge lengths compared in the eight-taxon setting were 0.1, 0.05 and 0.01 for the same reason. Terminal edge-lengths were set to 0.1.

Tree types tested:

Table 2.1: Eight-Taxon Tree Types Tested

Name	R Script Reference	Description
8taxon-0	8cat0	with none of the edge-lengths set to 0
8taxon-1	8cat1	with one of the edge-lengths set to 0
8taxon-2	8cat2sep	for two of the edge-lengths set to 0, where the zero length internal edges *are not* adjacent
8taxon-2-adj	8cat2adj	for two of the edge-lengths set to 0, where the zero length internal edges *are* adjacent
8taxon-star	8star	where all of the edge-lengths are set to 0

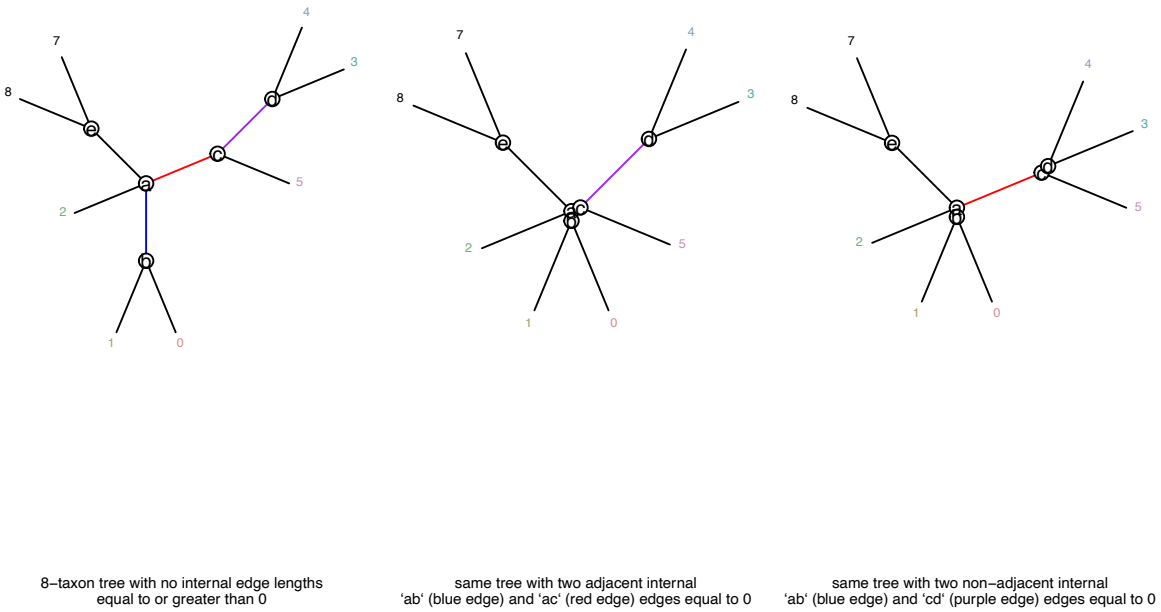

Figure 2.7: Three of the eight-taxon trees considered in simulation

Figure 2.7 shows an example of an eight-taxon tree with no zero length internal edges on the left, followed by a tree with two adjacent edge-lengths set to 0 in the middle. The tree on the right has two non-adjacent edge-lengths set to 0.

2.7 Nomenclature For Simulations

Throughout the thesis I will be referring to the tree types I've analyzed either by full name or description or by shorthand abbreviation/code that I was working with while writing the script. Table 2.2 summarizes each tree type and its abbreviation.

Table 2.2: Tree Type Abbreviations and Descriptions

Name	True Trees	Description
6taxon-0	1	Six taxon tree with all positive edge-lenghs
6taxon-1	3	Six taxon tree with one zero-length edge
6taxon-2	9	Six taxon tree with two non-adjacent zero-length edges
6taxon-2-adj	15	Six taxon tree with two adjacent zero-length edges
6taxon-star	105	Six taxon star tree
8taxon-0	1	Eight taxon tree with all positive edge-lengths
8taxon-1	3	Eight taxon tree with one zero-length edge
8taxon-2	9	Eight taxon tree with two non-adjacent zero-length edges
8taxon-2-adj	15	Eight taxon tree with two non-adjacent zero-length edges
8taxon-star	10395	Eight taxon star tree

For convenience, The tree plots below will show an example of each type of tree analyzed in the thesis for both the six and eight-taxon cases:

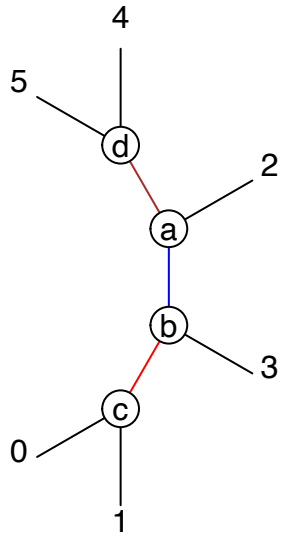

Figure 2.8: A well resolved 6-taxon tree, with all internal-edge-lenghts > 0

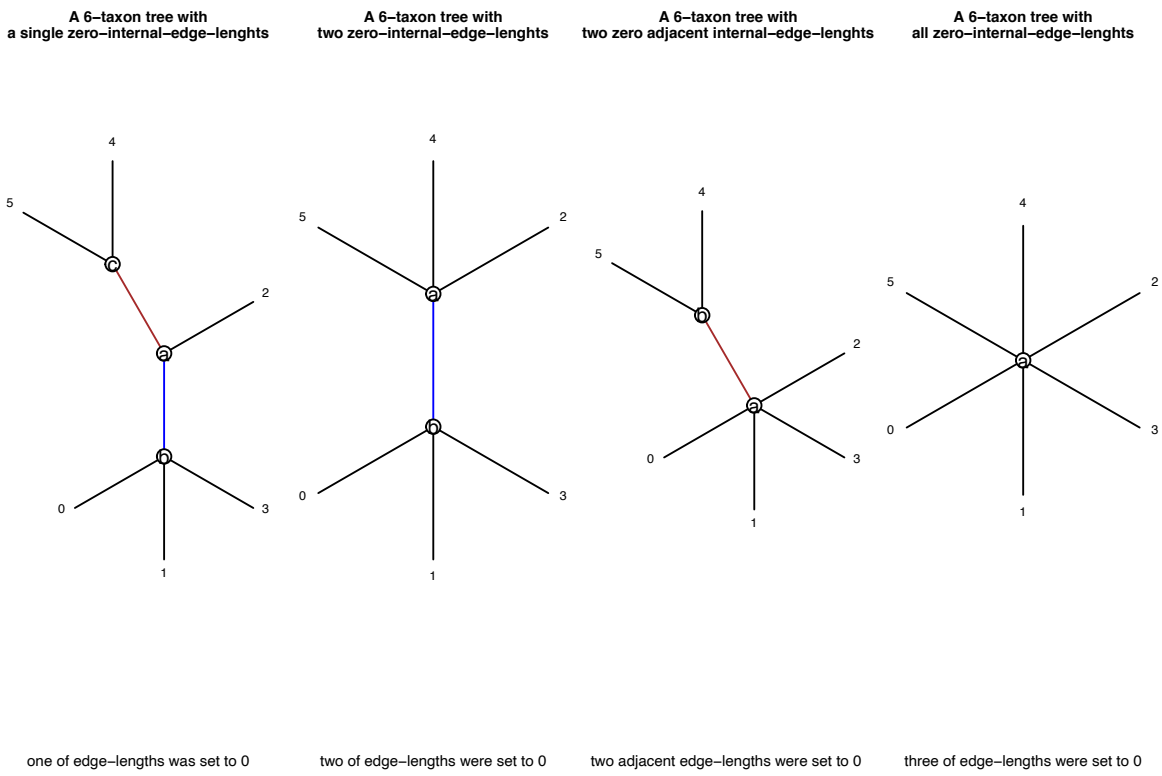

Figure 2.9: Six-taxon tree types

and the 8-taxon case:

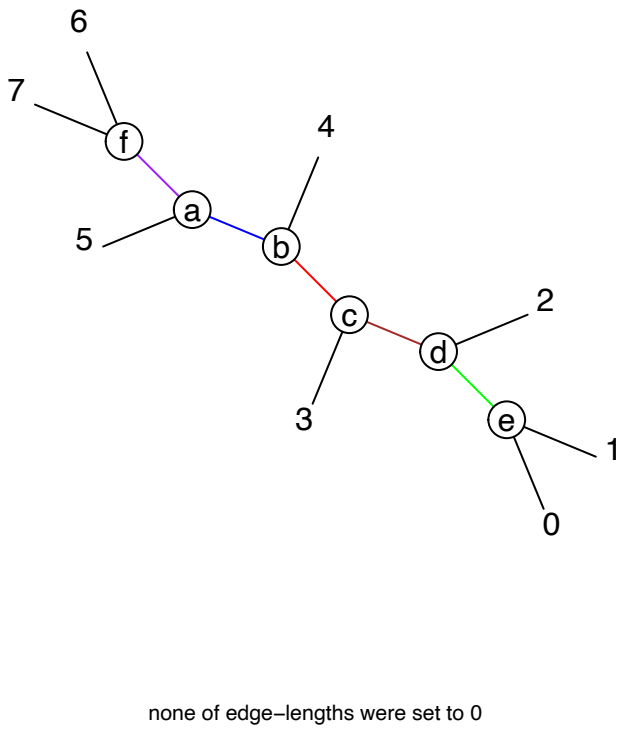

Figure 2.10: A well resolved 8-taxon tree, with all internal-edge-lenghts > 0

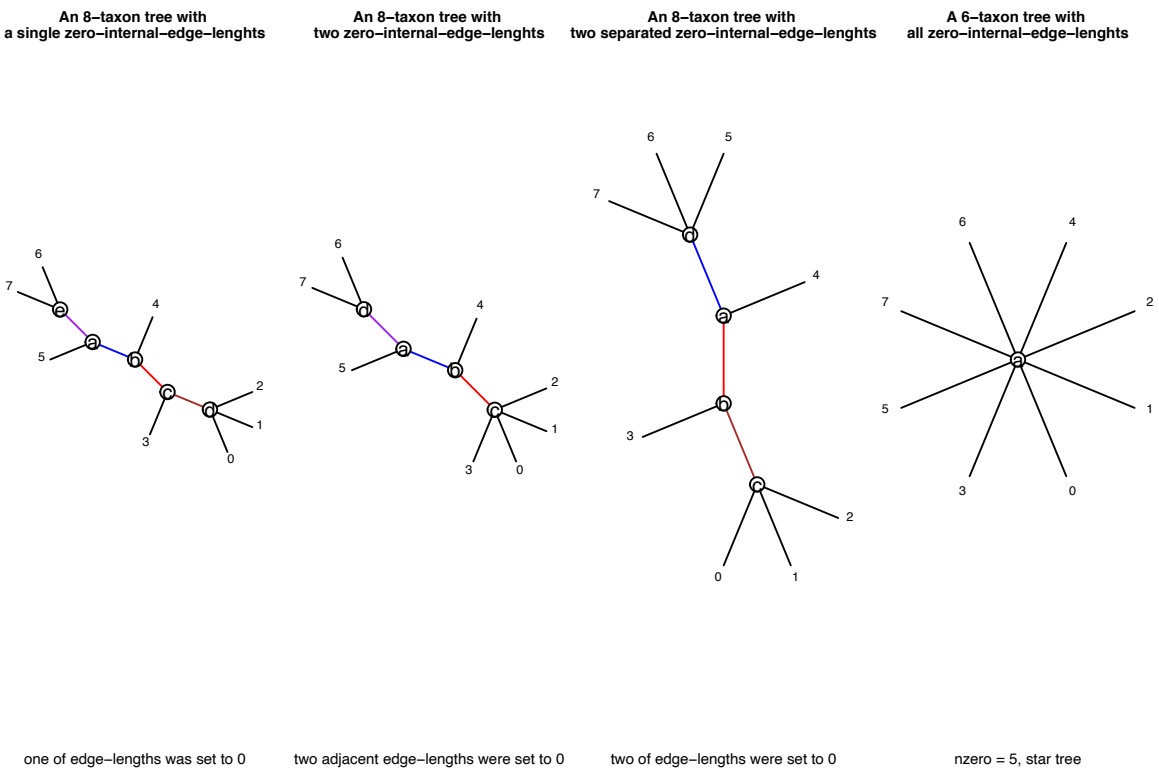

Figure 2.11: Eight-taxon tree types

2.8 Performance Comparison Metrics

The metrics used to describe and evaluate the tests are set size and coverage.

Mean Set Size

This metric counts the number of trees in a confidence set. For example for the well-resolved tree in simulation #1, for SH where internal edge length is 0.1 we have the set (26, 52, 62, 64, 65, 86, 102) so the set size here would be the count of elements in the set, 7. Ideally, we would expect to see the set size as close as possible to the expected true set size (in the case above, it's a set of a single tree, size of 1). The values in the data tables shown in each result below will show the mean set size $\pm 1.96 \times \frac{\text{Standard Deviation}}{\sqrt{n}}$.

Mean Tree Coverage

Coverage of a simulation is the count of true trees predicted over the total number of true trees for the tree type. For example, if a six-taxon simulating tree has 3 true trees, and the AU test in a particular simulation only predicted 2 of those three trees, then the coverage for this instance is 2/3. If the simulation-tree coverage of each true tree in the set is a boolean, than the mean over simulations of that boolean is the mean-tree coverage. Taking the mean of all mean-tree for each test gives us the coverage for that test. This calculation is equivalent to counting the true trees present in each set (simulation) for a test, and dividing by the number of expected true trees to produce a simulation-coverage (for example if in simulation #1 we expected 3 trees and counted only 2 true trees in the set produced by the simulation then the simulation-coverage is 0.67). Equivalently, it is the coverage (over true trees) of the coverage proportions for those true trees.

This metric shows us how many of the true trees were predicted correctly by the test. For a $(1-\alpha) \times 100\%$ confidence set procedure, the coverage should be approximately $1-\alpha$. Erring on the side of being larger than $1-\alpha$ (a conservative procedure or test) is better than being too small.

Exact Trees

Lastly, a count of simulations that produced the exact number of true trees in the confidence set, without any excess trees, over the total number of simulations was used in the comparison.

2.8.1 Datasets

In summary, three types of datasets will be examined; 6-taxon, 8-taxon data sets and several real data sets.

Several cases of tree types will be examined, each with a different number of internal edges set at length zero. When some internal edge lengths are set to zero, the resulting tree is equivalent to several tree topologies which is expected to make selection bias more challenging for a test.

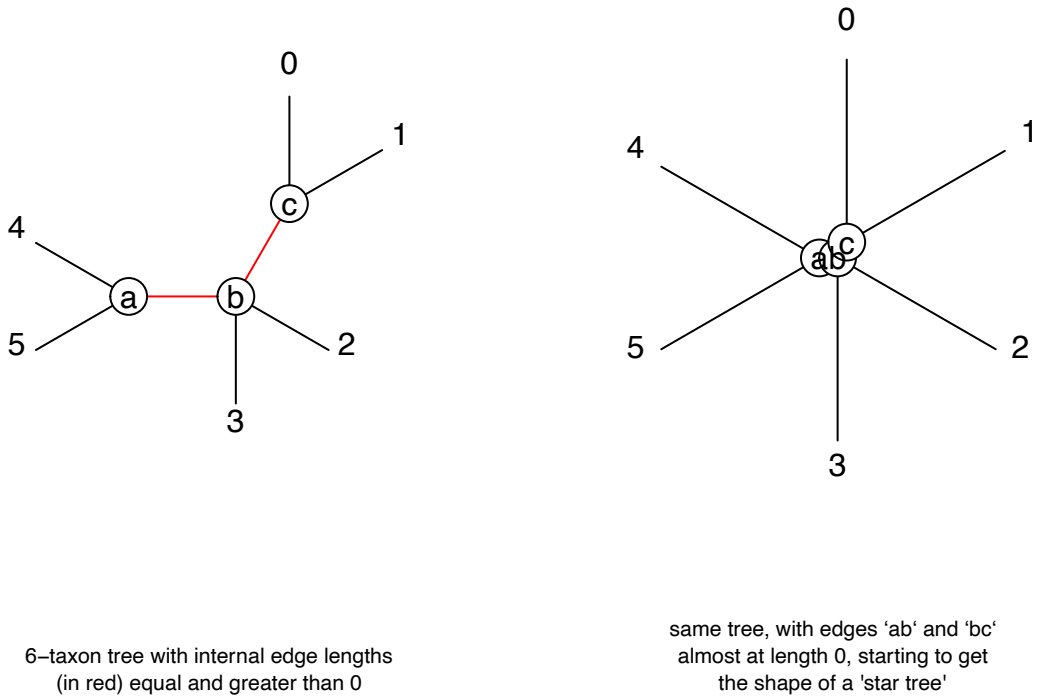

Figure 2.12: Six-taxon with two zero-length edges Example

The trees plotted in Figure 2.12 demonstrate the case how a particular tree is consistent with the star tree. In the star tree case, all trees are true making selection bias particularly challenging.

Chapter 3

Simulation Results

Throughout this chapter I will be using abbreviations at times for the tests as follows:

Table 3.1: Test Abbrevaitions

Abbreviation	Test
KH	KH
SH	SH
AU	AU
AU-corr	AU with Bootstrap correction
chisq	Chi-sqaure
bonf	Bonferroni

3.1 Six-taxon

3.1.1 Six taxon tree with all positive edge-lenghs (6taxon-0)

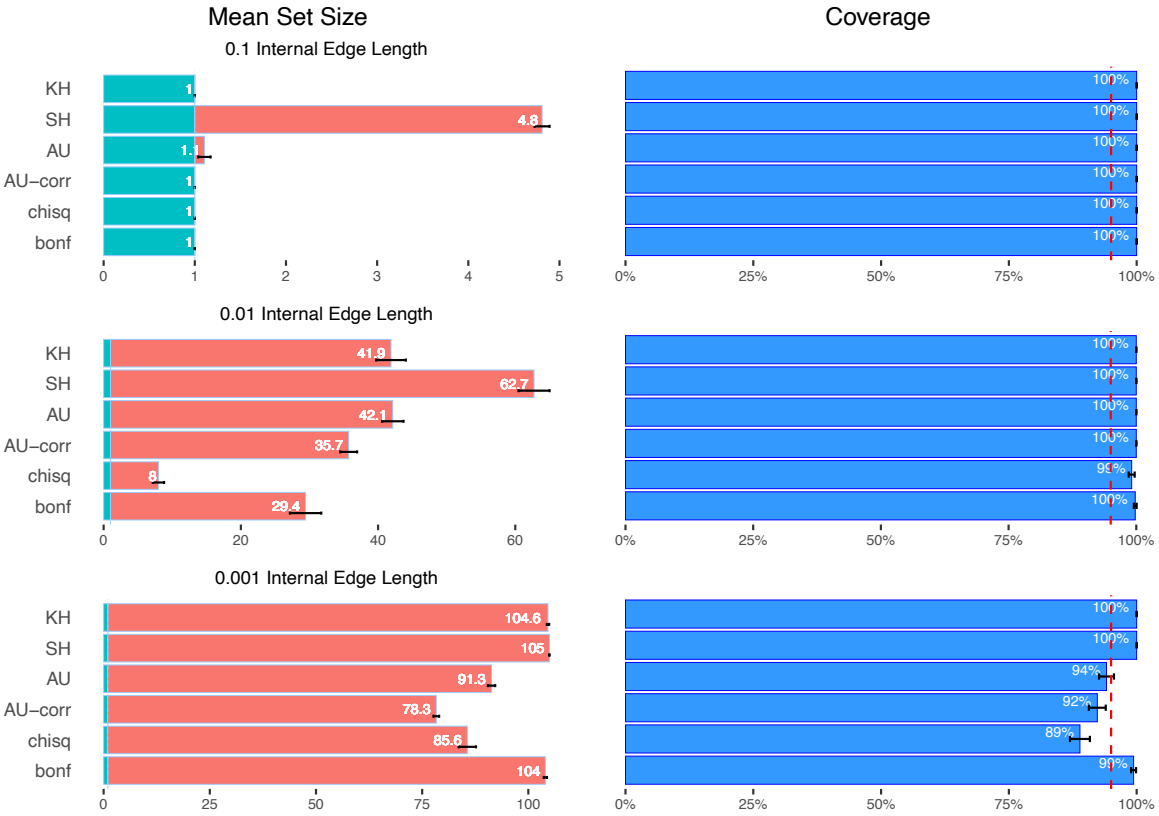

Figure 3.1: Results for six taxon simulations with no zero edge lengths

Table 3.2: Results for six taxon simulations with no zero edge-lengths

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	1.000 ± 0.000	1.000	0.000	100.00%	100.000%	100.000%	1000
SH	4.809 ± 0.082	1.000	3.809	100.00%	100.000%	100.000%	5
AU	1.106 ± 0.067	1.000	0.106	100.00%	100.000%	100.000%	976
AU-corr	1.000 ± 0.000	1.000	0.000	100.00%	100.000%	100.000%	1000
chisq	1.000 ± 0.000	1.000	0.000	100.00%	100.000%	100.000%	1000
bonf	1.000 ± 0.000	1.000	0.000	100.00%	100.000%	100.000%	1000
0.01 Internal Edge Length							
KH	41.886 ± 2.166	1.000	40.886	100.00%	100.000%	100.000%	0
SH	62.725 ± 2.270	1.000	61.725	100.00%	100.000%	100.000%	0
AU	42.138 ± 1.569	1.000	41.138	100.00%	100.000%	100.000%	0
AU-corr	35.717 ± 1.230	1.000	34.717	100.00%	100.000%	100.000%	0
chisq	7.983 ± 0.828	0.991	6.992	99.10%	98.515%	99.685%	161
bonf	29.432 ± 2.283	0.998	28.434	99.80%	99.523%	100.077%	65
0.001 Internal Edge Length							
KH	104.623 ± 0.281	1.000	103.623	100.00%	100.000%	100.000%	0
SH	104.963 ± 0.073	1.000	103.963	100.00%	100.000%	100.000%	0
AU	91.337 ± 0.860	0.941	90.396	94.10%	92.640%	95.560%	0
AU-corr	78.326 ± 0.668	0.923	77.403	92.30%	90.648%	93.952%	0
chisq	85.632 ± 2.045	0.889	84.743	88.90%	86.953%	90.847%	0
bonf	104.008 ± 0.379	0.994	103.014	99.40%	98.921%	99.879%	0

For the configuration where none of the edge-lengths were set to 0, we have a single correct tree. This is the simplest tree to identify.

Mean Set Size demonstrates the mean set sizes of 95% confidence sets of trees for each of the tests listed in the x-axis, broken down by the three internal edge lengths shown in the legend on the right. The maximum value in this case would be 105 for all possible trees for a 6-taxon. The larger the edge length, the easier it is for the test to distinguish between trees, and since only one tree is true in the case where none of the edge-lengths is set to 0, we expect to have a set size of 1. Most tests performed well with an internal edge-length of 0.1 (the largest value tested) having a mean of set size = 1, with the exception of AU with a value of 1.1, and SH with a value of 4.8. SH's performance is relatively conservative with roughly 5x the set size of the other tests.

The black lines at the top of each bar are the error bars that show the 95% spread of the values for each case.

Focusing on the differences between internal edge-lengths for each of the trees, we see an increase in set size as mentioned above, however, the rate of increase is not the same for every test. Moreover, we see that for internal edge-length set to 0.01 for instance, chisq managed to keep its mean set size at a low 8.0, the smallest one in the test group, while AU-corr had a value of 35.717 $\sim\times 4.5$ as big, while for internal edge-length set to 0.001 AU-corr has the lowest mean set size value of 78.326 and chisq's value shoots up to 85.632.

Coverage shows the proportion of times the true tree set (in this case only tree #64) was present in the set. This value was calculated for each simulation and then the mean value was taken per test, with 95% confidence intervals for the mean being indicated as well. The maximum value for this plot is 1 (as there is only one true tree). Even though we set the probability to be at 0.95 (so we expect to see the true tree 95% of the times), we still see that when internal edge-lengths are 0.1 the mean value for coverage is 1 across the board. With the high set size we viewed in the Mean Set Size plot set on the left, it appears that all tests predicted the true tree in every set. The values are high for lower internal edge-length's with chi-square having the

lowest coverage rate when internal edge-lengths are 0.001 with a value of 0.889.

Exact Trees shows a large contrast between SH and the rest of the tests for internal edge length 0.1, where SH only has 5 simulations containing exactly the one true tree out of 1,000, whereas the rest of the tests either had the exact set every single simulation or very close to it (AU). Varying the internal edge length to a smaller value shows the tests struggle with this metric, and for a value set to 0.001 none of them had a single simulation that contained only the true tree.

3.1.2 Six taxon tree with one zero-length edge (6taxon-1)

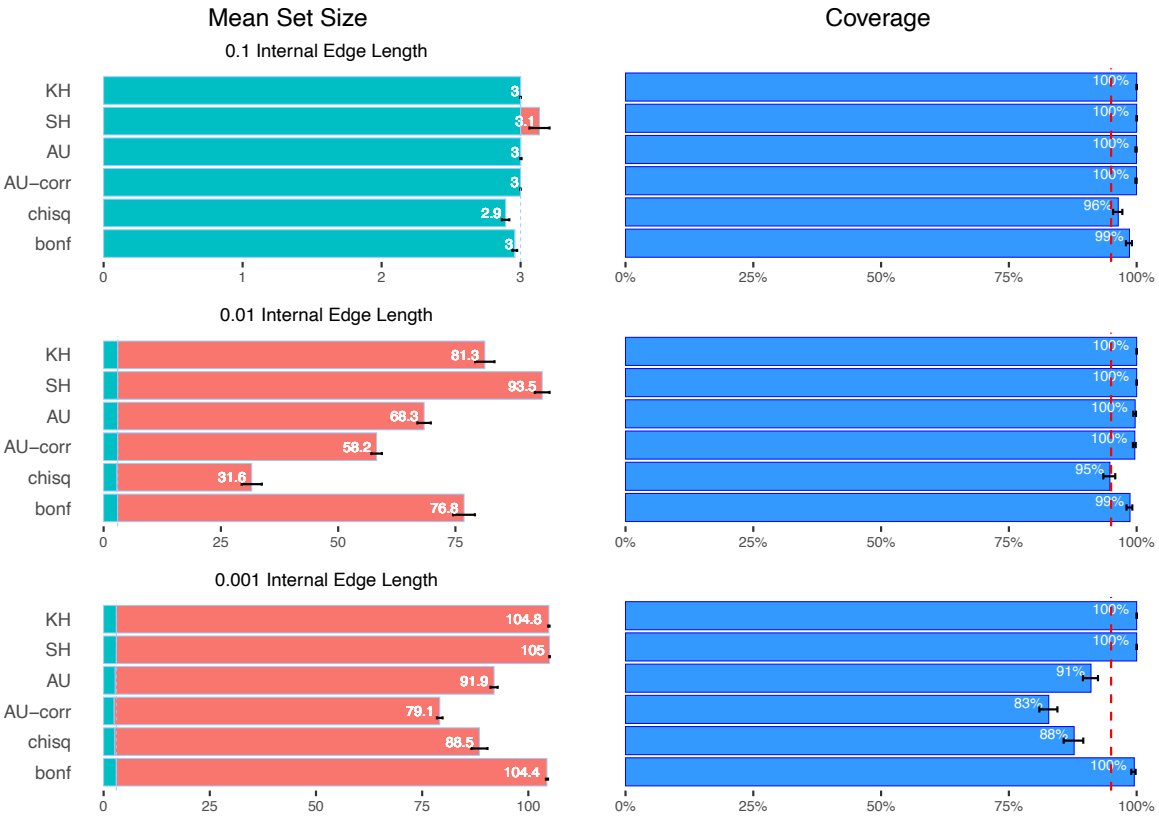

Figure 3.2: Results for six taxon simulations with a single zero-length edge

Table 3.3: Results for six taxon simulations with a single zero-length edge

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	3.000 ± 0.000	3.000	0.000	100.00%	100.000%	100.000%	1000
SH	3.136 ± 0.072	3.000	0.136	100.00%	100.000%	100.000%	980
AU	3.002 ± 0.005	2.999	0.003	99.97%	99.764%	99.995%	997
AU-corr	2.999 ± 0.002	2.999	0.000	99.97%	99.764%	99.995%	999
chisq	2.892 ± 0.027	2.892	0.000	96.40%	95.382%	97.200%	939
bonf	2.958 ± 0.017	2.958	0.000	98.60%	97.899%	99.069%	976
0.01 Internal Edge Length							
KH	81.276 ± 2.059	3.000	78.276	100.00%	100.000%	100.000%	1
SH	93.492 ± 1.581	3.000	90.492	100.00%	100.000%	100.000%	0
AU	68.329 ± 1.432	2.991	65.338	99.70%	99.345%	99.863%	0
AU-corr	58.204 ± 1.119	2.989	55.215	99.63%	99.272%	99.816%	0
chisq	31.573 ± 2.145	2.843	28.730	94.77%	93.480%	95.811%	123
bonf	76.841 ± 2.319	2.961	73.880	98.70%	98.014%	99.151%	26
0.001 Internal Edge Length							
KH	104.837 ± 0.192	3.000	101.837	100.00%	100.000%	100.000%	0
SH	105.000 ± 0.000	3.000	102.000	100.00%	100.000%	100.000%	0
AU	91.928 ± 0.854	2.732	89.196	91.07%	89.484%	92.432%	0
AU-corr	79.141 ± 0.667	2.484	76.657	82.80%	80.960%	84.496%	0
chisq	88.474 ± 1.900	2.633	85.841	87.77%	85.720%	89.556%	0
bonf	104.385 ± 0.246	2.986	101.399	99.53%	98.949%	99.793%	0

This is the setting where we have a single internal edge length that is equal to zero.

There are three trees that are compatible with the generating tree. Consequently, the expected set size is 3.

Mean Set Size Again, all tests performed well under internal edge-length set at 0.1 having a mean of set size of ~3, SH still slightly higher than the rest at 3.136 but this time the difference is minor.

Focusing on the differences between internal edge-length again we see the same story as in none of edge-lengths were set to 0, with the same phenomena where chisq performs pretty well even at 0.01, but grows above AU for internal edge-length 0.001.

Coverage Even though these are 95% confidence sets (so we expect to see the true tree 95% of the times), we still see that for internal edge-length set to 0.1 the mean value for coverage is roughly ~100% across the board, so the set of three trees shown above was correctly predicted most of the time, and it's only at internal edge-length 0.001 that we start seeing the mean decrease as low as 83% for AU Corr.

3.1.3 Six taxon tree with two non-adjacent zero-length edges (6taxon-2)

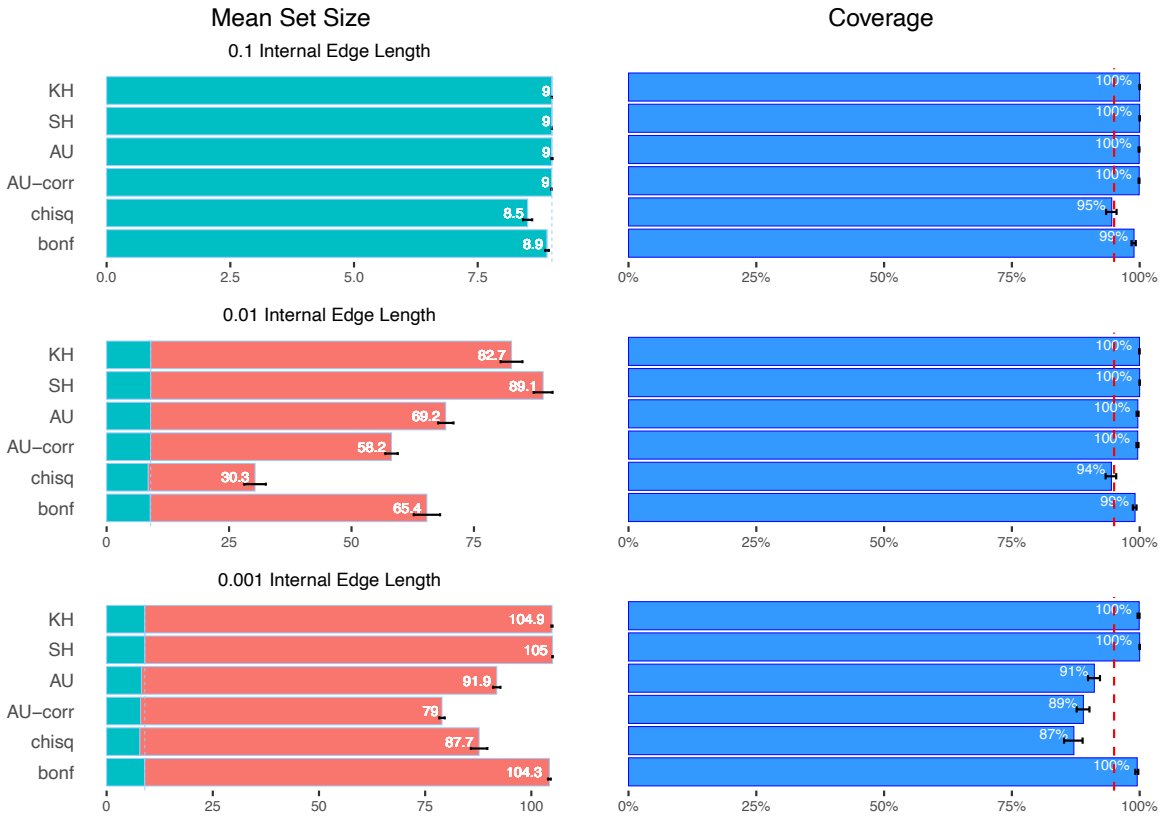

Figure 3.3: Results for six taxon simulations with two non-adjacent zero-length edges

Table 3.4: Results for six taxon simulations with two non-adjacent zero-length edges

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	9.000 ± 0.000	9.000	0.000	100.00%	100.000%	100.000%	1000
SH	9.000 ± 0.000	9.000	0.000	100.00%	100.000%	100.000%	1000
AU	9.001 ± 0.012	8.992	0.009	99.91%	99.793%	99.962%	990
AU-corr	8.992 ± 0.007	8.992	0.000	99.91%	99.793%	99.962%	994
chisq	8.509 ± 0.092	8.509	0.000	94.54%	93.428%	95.480%	868
bonf	8.901 ± 0.036	8.901	0.000	98.90%	98.417%	99.237%	959
0.01 Internal Edge Length							
KH	82.675 ± 2.229	8.997	73.678	99.97%	99.897%	99.989%	79
SH	89.121 ± 1.933	9.000	80.121	100.00%	100.000%	100.000%	48
AU	69.244 ± 1.571	8.968	60.276	99.64%	99.405%	99.788%	19
AU-corr	58.168 ± 1.273	8.966	49.202	99.62%	99.378%	99.771%	19
chisq	30.304 ± 2.201	8.502	21.802	94.47%	93.339%	95.413%	504
bonf	65.397 ± 2.688	8.919	56.478	99.10%	98.663%	99.395%	273
0.001 Internal Edge Length							
KH	104.881 ± 0.157	8.992	95.889	99.91%	99.599%	99.980%	0
SH	104.998 ± 0.004	9.000	95.998	100.00%	100.000%	100.000%	0
AU	91.867 ± 0.851	8.202	83.665	91.13%	89.895%	92.233%	0
AU-corr	78.961 ± 0.644	8.009	70.952	88.99%	87.677%	90.176%	0
chisq	87.706 ± 1.935	7.842	79.864	87.13%	85.212%	88.837%	8
bonf	104.251 ± 0.321	8.957	95.294	99.52%	99.149%	99.732%	0

For the case of two internal edge lengths equal to zero, we have two possibilities: the

0-length branches are adjacent, or not adjacent. In this sub-section we consider first the **non-adjacent** case, with 9 trees compatible with the true tree.

Mean Set Size At an internal edge length of 0.1 SH did not show any excess of trees in this configuration as opposed to the two previous cases, however, SH did find all 105 trees correct in its set for internal edge length of 0.001 as in the previous two configurations.

Coverage Chi-square's confidence interval coverage contained the 95% confidence level for internal edge lengths 0.1 and 0.01, but was well below the level for internal edge length set to 0.001. For the harder-to-distinguish case of internal edge length of 0.001 it seems like KH, SH and Bonferroni were very conservative, all predicting 100% of the true trees coverage, and that the AU's and chisq were below.

3.1.4 Six taxon tree with two adjacent zero-length edges (6taxon-2-adj)

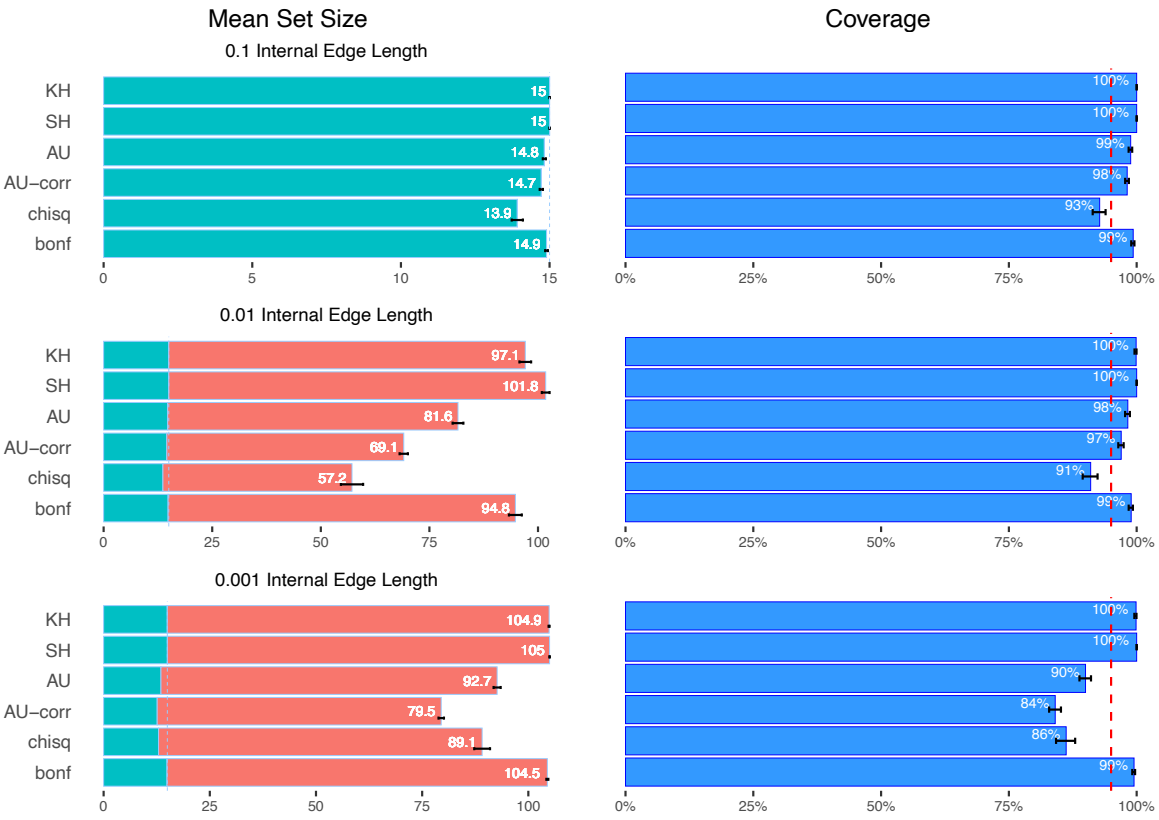

Figure 3.4: Results for six taxon simulations with two adjacent zero-length edges

Table 3.5: Results for six taxon simulations with two adjacent zero-length edges

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	15.000 ± 0.000	15.000	0.000	100.00%	100.000%	100.000%	1000
SH	15.000 ± 0.000	15.000	0.000	100.00%	100.000%	100.000%	1000
AU	14.824 ± 0.052	14.824	0.000	98.83%	98.428%	99.125%	930
AU-corr	14.722 ± 0.054	14.722	0.000	98.15%	97.751%	98.474%	832
chisq	13.914 ± 0.189	13.914	0.000	92.76%	91.397%	93.921%	828
bonf	14.898 ± 0.045	14.898	0.000	99.32%	98.942%	99.564%	962
0.01 Internal Edge Length							
KH	97.069 ± 1.345	14.984	82.085	99.89%	99.579%	99.973%	13
SH	101.773 ± 0.891	15.000	86.773	100.00%	100.000%	100.000%	6
AU	81.580 ± 1.239	14.740	66.840	98.27%	97.740%	98.672%	4
AU-corr	69.092 ± 0.957	14.546	54.546	96.97%	96.408%	97.452%	5
chisq	57.180 ± 2.556	13.650	43.530	91.00%	89.439%	92.350%	228
bonf	94.790 ± 1.482	14.839	79.951	98.93%	98.438%	99.263%	39
0.001 Internal Edge Length							
KH	104.901 ± 0.134	14.982	89.919	99.88%	99.546%	99.968%	0
SH	105.000 ± 0.000	15.000	90.000	100.00%	100.000%	100.000%	0
AU	92.651 ± 0.833	13.499	79.152	89.99%	88.805%	91.068%	0
AU-corr	79.488 ± 0.640	12.610	66.878	84.07%	82.876%	85.190%	0
chisq	89.119 ± 1.858	12.931	76.188	86.21%	84.231%	87.970%	4
bonf	104.477 ± 0.233	14.923	89.554	99.49%	99.142%	99.693%	0

By contrast with the previous example, the two zero-length edges are adjacent. There are thus 15 trees that are compatible with the true generating tree.

Mean Set Size This is the first case where we can see that other than the conservative KH and SH, all other tests predicted a set size of less than 15 (number of true trees for this configuration) for internal edge length 0.1.

Coverage This brings our coverage down to lower than 100% for the case of internal edge length 0.1. But for any test other than the chi-square test, coverage is above the 95% confidence level. SH & KH both predict 100% of the coverage, for all three varying internal edge lengths, although the expected value would have been closer to the 95% confidence level. You can also see that the mean wrong trees for these tests is the highest compared to the other tests. This again shows how conservative these tests are.

3.1.5 Six taxon star tree (6taxon-star)

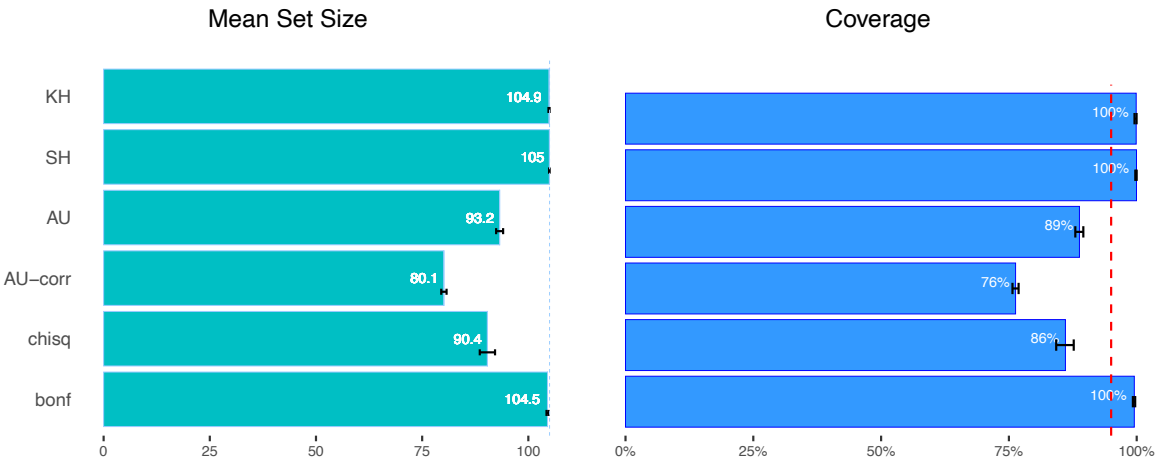

Figure 3.5: Results for six taxon simulations with three zero-length edges

Table 3.6: Results for six taxon simulations with three zero-length edges

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
KH	104.894 ± 0.153	104.894	0	99.90%	99.574%	99.976%	994
SH	104.960 ± 0.078	104.960	0	99.96%	99.730%	99.995%	999
AU	93.242 ± 0.821	93.242	0	88.80%	87.997%	89.560%	251
AU-corr	80.119 ± 0.628	80.119	0	76.30%	75.701%	76.896%	0
chisq	90.361 ± 1.799	90.361	0	86.06%	84.256%	87.684%	652
bonf	104.507 ± 0.229	104.507	0	99.53%	99.254%	99.705%	900

The set of true trees is all 105 possible trees, this is a star tree where all terminal edges stem from the root.

Mean Set Size With star tree being the hardest configuration to deal with, The AU tests and chisq were struggling to recognize that all trees should be in their confidence sets, while KH, SH and Bonferroni were consistently showing a mean set size of the size of all possible trees, 105.

Coverage As with the mean set size, AU-correction had the most trouble distinguish the trees and only predicted ~75% of the trees as true trees. Chi-Square was slightly better at 86%.

3.2 Eight-taxon

3.2.1 Eight taxon tree with all positive edge-lengths (8taxon-0)

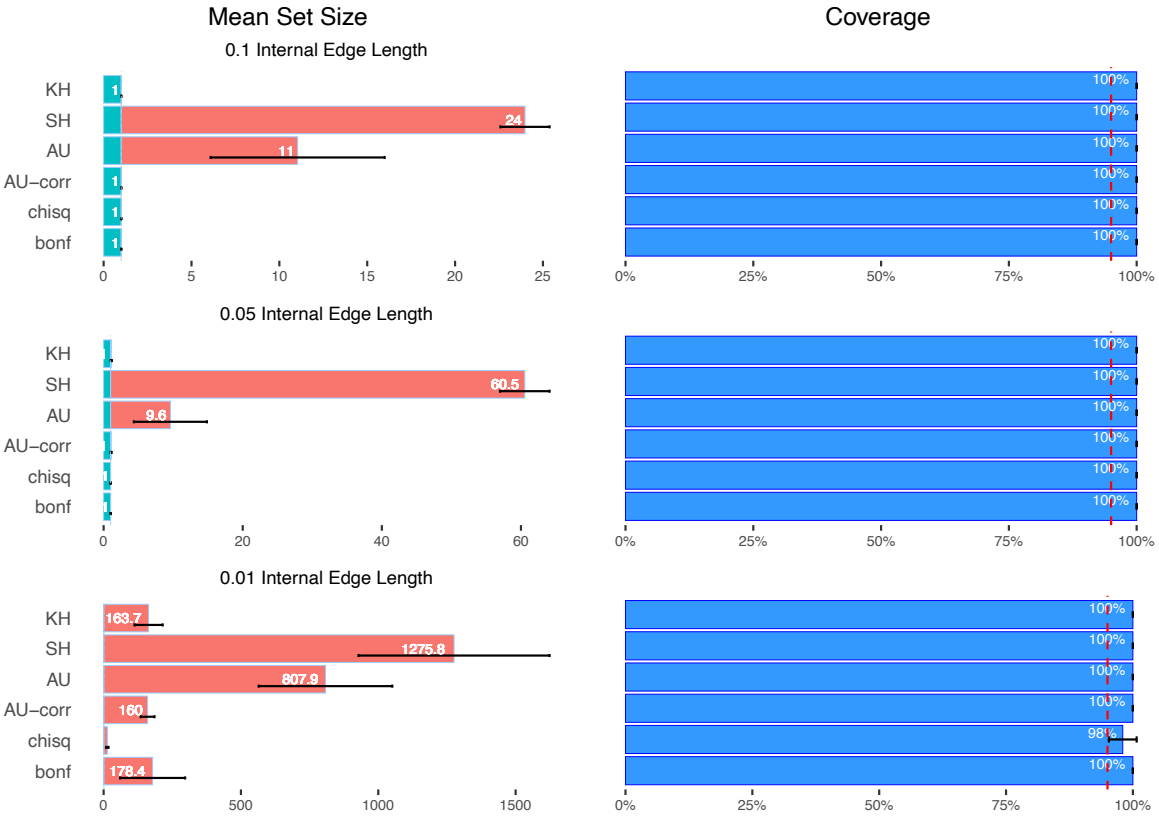

Figure 3.6: Eight-taxon simulations with a single true tree

Table 3.7: Eight-taxon simulations with a single true tree

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	1.000 ± 0.000	1.00	0.00	100.00%	100.000%	100.000%	100
SH	23.980 ± 1.398	1.00	22.98	100.00%	100.000%	100.000%	0
AU	11.040 ± 4.950	1.00	10.04	100.00%	100.000%	100.000%	53
AU-corr	1.000 ± 0.000	1.00	0.00	100.00%	100.000%	100.000%	100
chisq	1.000 ± 0.000	1.00	0.00	100.00%	100.000%	100.000%	100
bonf	1.000 ± 0.000	1.00	0.00	100.00%	100.000%	100.000%	100
0.05 Internal Edge Length							
KH	1.100 ± 0.081	1.00	0.10	100.00%	100.000%	100.000%	94
SH	60.550 ± 3.562	1.00	59.55	100.00%	100.000%	100.000%	0
AU	9.600 ± 5.249	1.00	8.60	100.00%	100.000%	100.000%	66
AU-corr	1.080 ± 0.072	1.00	0.08	100.00%	100.000%	100.000%	95
chisq	1.000 ± 0.000	1.00	0.00	100.00%	100.000%	100.000%	100
bonf	1.000 ± 0.000	1.00	0.00	100.00%	100.000%	100.000%	100
0.01 Internal Edge Length							
KH	163.720 ± 51.305	1.00	162.72	100.00%	100.000%	100.000%	0
SH	1275.850 ± 347.877	1.00	1274.85	100.00%	100.000%	100.000%	0
AU	807.870 ± 243.268	1.00	806.87	100.00%	100.000%	100.000%	0
AU-corr	159.990 ± 25.233	1.00	158.99	100.00%	100.000%	100.000%	0
chisq	13.530 ± 4.232	0.98	12.55	98.00%	95.256%	100.744%	12
bonf	178.420 ± 118.411	1.00	177.42	100.00%	100.000%	100.000%	4

With no zero-length internal edge lengths, there is a single tree.

Mean Set Size Comparing to the six-taxon case with a configuration of a single true tree, we see a somewhat similar picture here; SH is still being very conservative at internal edge length of 0.1, predicting more wrong trees than the rest, and this time AU-correction is behaving similarly too, whereas in the six-taxon case its number of wrong trees was ~10% larger than the true set size, this time it's ~1000%. The variance for the set size spreads much wider here than in the six-taxa since only 100 simulations were run for the eight-taxa, versus 1,000 in the former case. We see that for larger internal edge length only AU's variance is large compared to the others, although for smaller internal edge lengths SH has large variance as well.

The large variance for AU can be explained by comparing to the relatively small variance of AU-corr. The two methods differ only when $BP = 0$. For reasons discussed in Section 2.2, the AU $p-$ values become highly unstable in this case.

Coverage For both internal edge lengths of 0.1 and 0.01 we see that all tests were able to predict close to 100% of the coverage, hence a little too conservative for our 95% confidence level. At internal edge length of 0.01, the lowest value of internal edge length for the eight taxa that we tested, only chisq had a very slight decrease in coverage to 98%.

3.2.2 Eight taxon tree with one zero-length edge (8taxon-1)

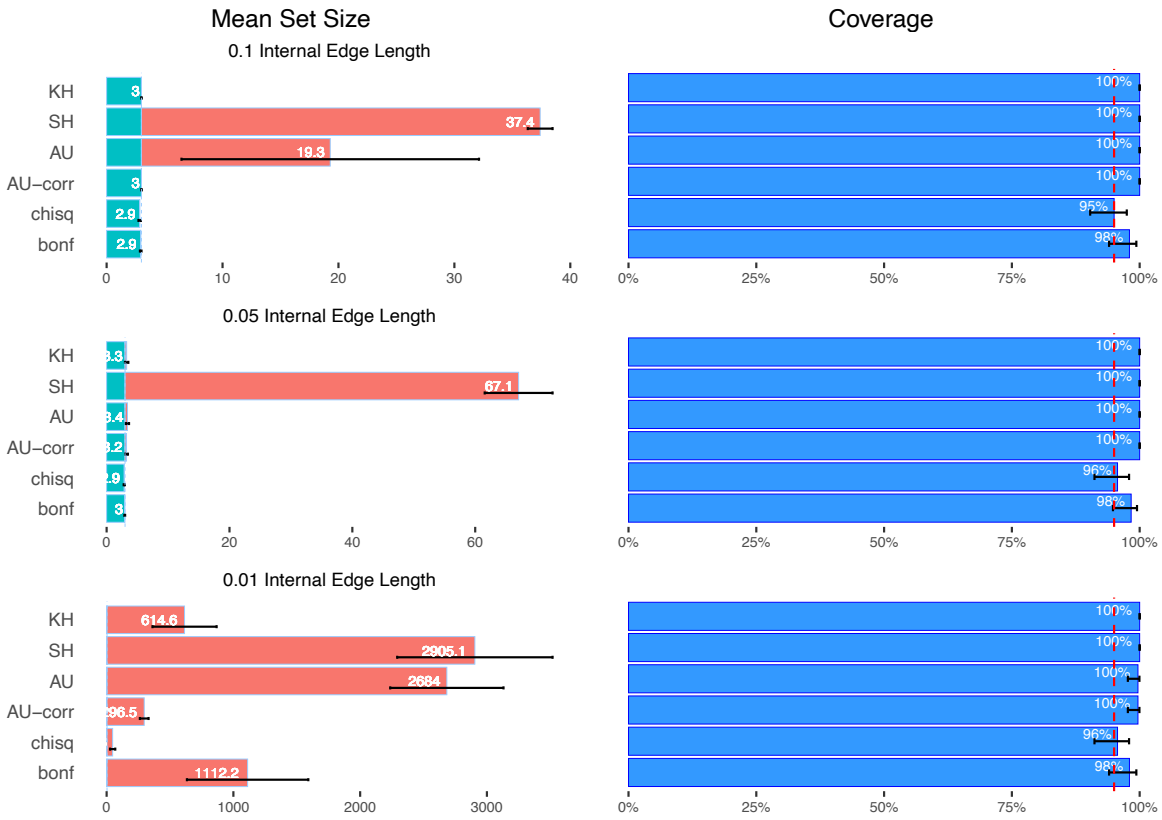

Figure 3.7: Eight-taxon simulations with a single zero-length edge

Table 3.8: Eight-taxon simulations with a single zero-length edge

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	3.000 ± 0.000	3.00	0.00	100.00%	100.000%	100.000%	100
SH	37.400 ± 1.070	3.00	34.40	100.00%	100.000%	100.000%	0
AU	19.290 ± 12.834	3.00	16.29	100.00%	100.000%	100.000%	60
AU-corr	3.000 ± 0.000	3.00	0.00	100.00%	100.000%	100.000%	100
chisq	2.850 ± 0.102	2.85	0.00	95.00%	90.318%	97.481%	92
bonf	2.940 ± 0.067	2.94	0.00	98.00%	94.017%	99.350%	97
0.05 Internal Edge Length							
KH	3.260 ± 0.272	3.00	0.26	100.00%	100.000%	100.000%	94
SH	67.070 ± 5.525	3.00	64.07	100.00%	100.000%	100.000%	0
AU	3.390 ± 0.255	3.00	0.39	100.00%	100.000%	100.000%	83
AU-corr	3.240 ± 0.221	3.00	0.24	100.00%	100.000%	100.000%	92
chisq	2.870 ± 0.095	2.87	0.00	95.67%	91.163%	97.927%	93
bonf	2.950 ± 0.058	2.95	0.00	98.33%	94.776%	99.481%	97
0.01 Internal Edge Length							
KH	614.650 ± 252.991	3.00	611.65	100.00%	100.000%	100.000%	0
SH	2905.130 ± 612.882	3.00	2902.13	100.00%	100.000%	100.000%	0
AU	2684.050 ± 446.997	2.99	2681.06	99.67%	97.688%	99.953%	0
AU-corr	296.510 ± 34.297	2.99	293.52	99.67%	97.688%	99.953%	0
chisq	47.330 ± 20.691	2.87	44.46	95.67%	91.163%	97.927%	12
bonf	1112.220 ± 478.899	2.94	1109.28	98.00%	94.017%	99.350%	1

With a single zero-length internal edge. There are three true trees.

Mean Set Size Again comparing to the six-taxon case where only a single edge length is set to zero, we see a similar comparison as with the previous case, where SH's wrong tree count has increased significantly, as well as AU's for the case of 0.1 internal edge length. At 0.01 internal edge length chisq seems to have the least number of wrong trees.

Coverage As well as having a low count of wrong trees, the coverage of the chi-square test is closest to the stated level (95%) of the confidence set.

3.2.3 Eight taxon tree with two non-adjacent zero-length edges (8taxon-2)

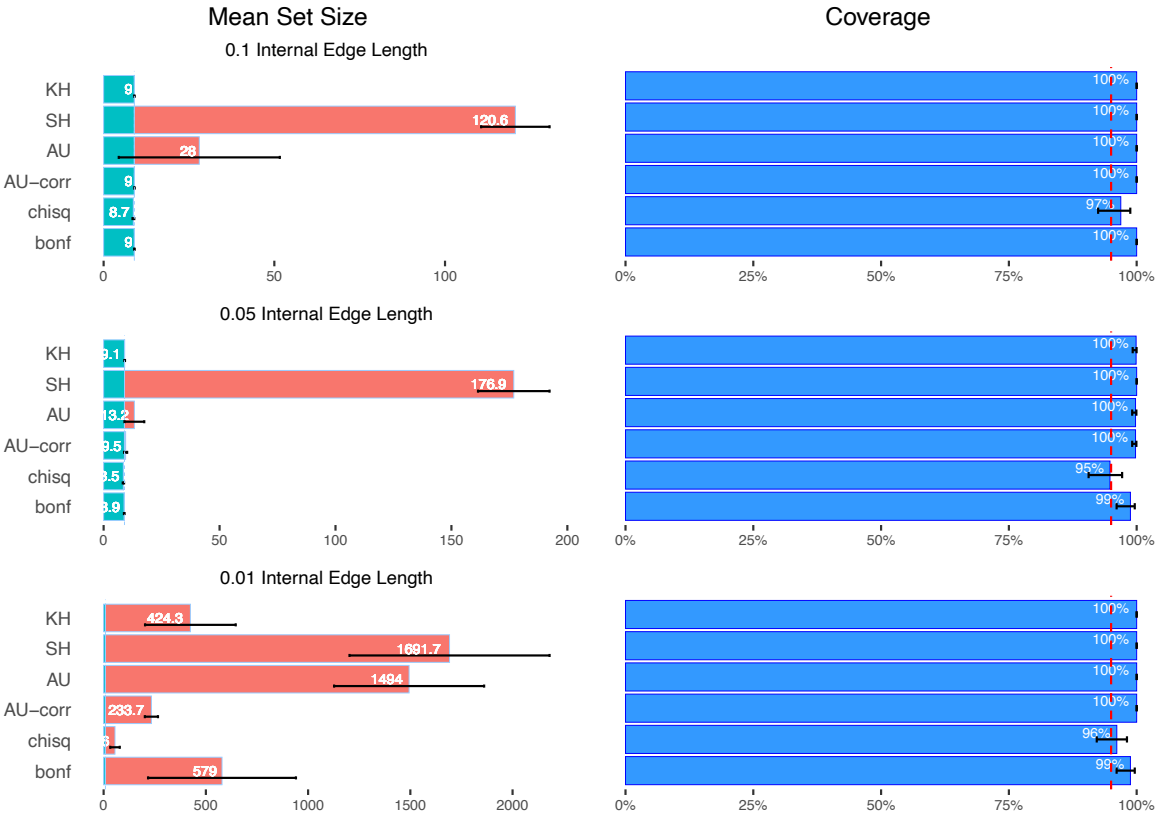

Figure 3.8: Eight-taxon simulations with two non-adjacent zero-length edges

Table 3.9: Eight-taxon simulations with two non-adjacent zero-length edges

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	9.000 ± 0.000	9.00	0.00	100.00%	100.000%	100.000%	100
SH	120.600 ± 10.024	9.00	111.60	100.00%	100.000%	100.000%	0
AU	28.020 ± 23.591	9.00	19.02	100.00%	100.000%	100.000%	64
AU-corr	9.000 ± 0.000	9.00	0.00	100.00%	100.000%	100.000%	100
chisq	8.720 ± 0.254	8.72	0.00	96.89%	92.465%	98.751%	94
bonf	9.000 ± 0.000	9.00	0.00	100.00%	100.000%	100.000%	100
0.05 Internal Edge Length							
KH	9.080 ± 0.127	8.99	0.09	99.89%	99.222%	99.984%	96
SH	176.900 ± 15.431	9.00	167.90	100.00%	100.000%	100.000%	0
AU	13.250 ± 4.267	8.98	4.27	99.78%	99.127%	99.944%	76
AU-corr	9.480 ± 0.657	8.98	0.50	99.78%	99.127%	99.944%	95
chisq	8.530 ± 0.283	8.53	0.00	94.78%	90.600%	97.157%	88
bonf	8.890 ± 0.130	8.89	0.00	98.78%	96.078%	99.626%	96
0.01 Internal Edge Length							
KH	424.300 ± 221.759	9.00	415.30	100.00%	100.000%	100.000%	0
SH	1691.730 ± 489.247	9.00	1682.73	100.00%	100.000%	100.000%	0
AU	1493.980 ± 366.928	9.00	1484.98	100.00%	100.000%	100.000%	0
AU-corr	233.730 ± 32.118	9.00	224.73	100.00%	100.000%	100.000%	0
chisq	55.610 ± 23.221	8.65	46.96	96.11%	92.191%	98.104%	25
bonf	578.980 ± 361.691	8.89	570.09	98.78%	96.078%	99.626%	4

With two **non-adjacent** zero-length internal edge. There are 9 true trees.

Mean Set Size The trend continues for SH and AU having a substantial number of wrong trees in their confidence set with SH having as much as 91% of its set wrong, and AU 68% for internal edge length of 0.1. Chisq is the only test with a set size less than the number of true trees (9).

Coverage Very close to the configuration of a single zero-length edge, the tests all performed similarly when varying the internal edge length, with chisq being larger than the 95% confidence interval but the closest to it.

3.2.4 Eight taxon tree with two adjacent zero-length edges (8taxon-2-adj)

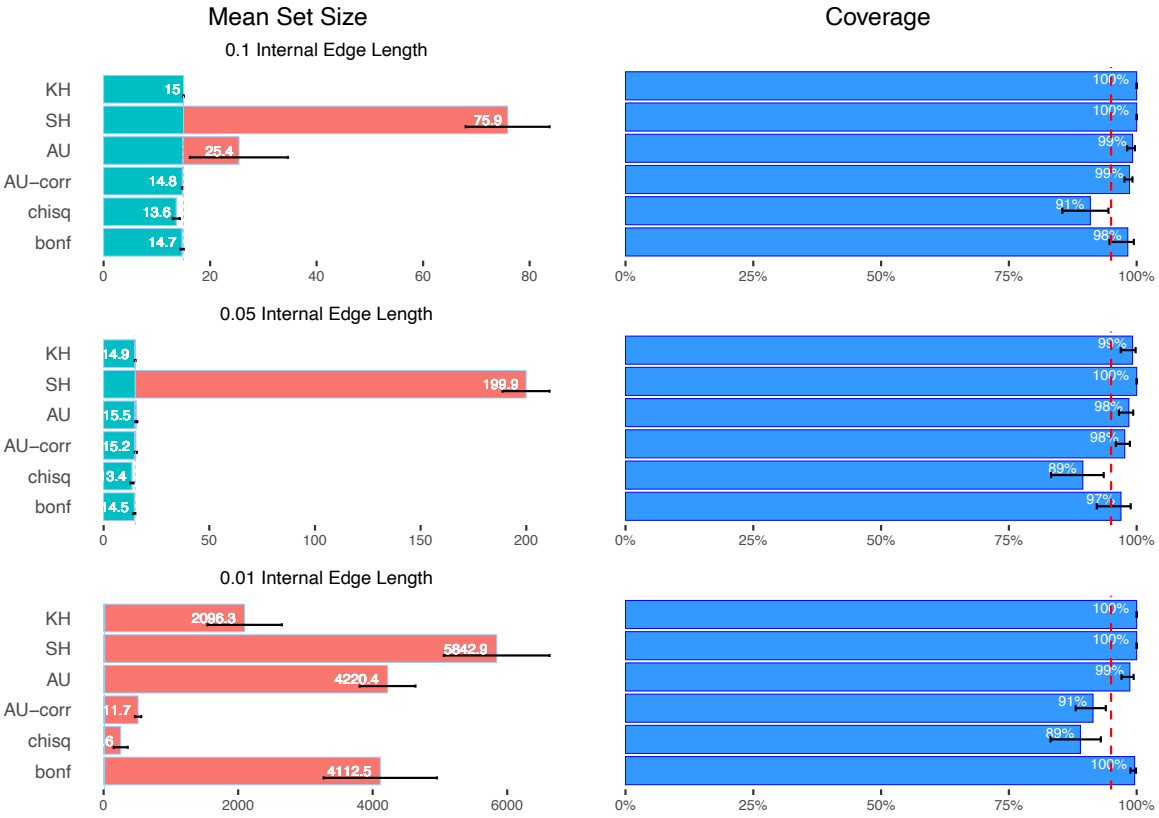

Figure 3.9: Eight-taxon simulations with two adjacent zero-length edges

Table 3.10: Eight-taxon simulations with two adjacent zero-length edges

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
0.1 Internal Edge Length							
KH	15.000 ± 0.000	15.00	0.00	100.00%	100.000%	100.000%	100
SH	75.870 ± 7.875	15.00	60.87	100.00%	100.000%	100.000%	5
AU	25.420 ± 9.210	14.88	10.54	99.20%	98.151%	99.656%	74
AU-corr	14.790 ± 0.112	14.79	0.00	98.60%	97.621%	99.180%	84
chisq	13.640 ± 0.665	13.64	0.00	90.93%	85.454%	94.482%	81
bonf	14.740 ± 0.297	14.74	0.00	98.27%	94.691%	99.448%	95
0.05 Internal Edge Length							
KH	14.930 ± 0.181	14.88	0.05	99.20%	96.885%	99.798%	96
SH	199.900 ± 11.121	15.00	184.90	100.00%	100.000%	100.000%	0
AU	15.470 ± 0.452	14.77	0.70	98.47%	96.526%	99.331%	78
AU-corr	15.240 ± 0.450	14.65	0.59	97.67%	95.925%	98.674%	71
chisq	13.420 ± 0.760	13.42	0.00	89.47%	83.263%	93.549%	80
bonf	14.540 ± 0.441	14.54	0.00	96.93%	92.199%	98.831%	93
0.01 Internal Edge Length							
KH	2096.310 ± 553.728	15.00	2081.31	100.00%	100.000%	100.000%	0
SH	5842.910 ± 786.055	15.00	5827.91	100.00%	100.000%	100.000%	0
AU	4220.440 ± 414.809	14.80	4205.64	98.67%	97.026%	99.408%	0
AU-corr	511.660 ± 46.647	13.72	497.94	91.47%	88.070%	93.962%	0
chisq	253.580 ± 104.404	13.35	240.23	89.00%	83.154%	92.988%	5
bonf	4112.470 ± 842.260	14.94	4097.53	99.60%	98.785%	99.869%	0

With two **adjacent** zero-length internal edge. There are 15 true trees.

Mean Set Size For both internal edge lengths of 0.1 and 0.05 we see a very similar pattern to that of the previous configuration, Eight-taxon simulations with two non-adjacent zero-length edges. However, at internal edge length of 0.01 the number of wrong trees in the confidence set increases unproportionately to the ratio of the true trees between the two configuration, 15/9. We see the wrong trees for KH, for example, increase by ~5 folds, chisq by over 7 folds, bringing the percentage of true trees in the set to less than 1% of the trees for all tests except for AU-Corr and Chisq. This shows that it is harder to distinguish the adjacent compared to the non-adjacent zero-length internal edges.

Coverage For the 0.01 internal edge length we see that both AU-Corr and Chisq, mentioned above with smaller set set sizes, also have below confidence-level coverage results with Chisq predicting as low as 89% of the true trees.

3.2.5 Eight taxon star tree (8taxon-star)

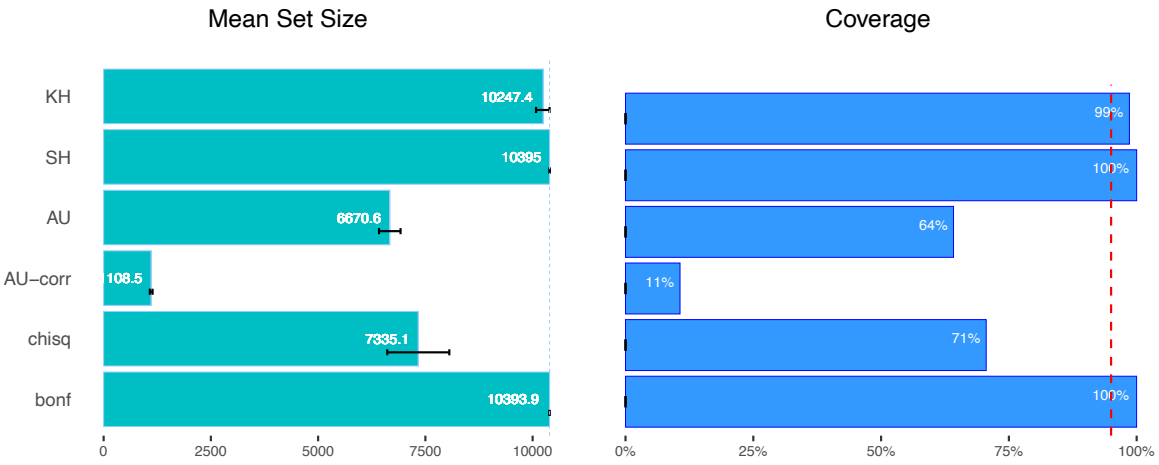

Figure 3.10: Eight-taxon simulations from a star tree

Table 3.11: Eight-taxon simulations from a star tree

Test	Mean Set Size	Mean Right Trees	Mean Wrong Trees	Coverage	Lower 95% CI	Upper 95% CI	Exact Trees
KH	10247.420 ± 168.467	10247.42	0	98.58%	0.000%	0.000%	94
SH	10395.000 ± 0.000	10395.00	0	100.00%	0.000%	0.000%	100
AU	6670.570 ± 250.385	6670.57	0	64.17%	0.000%	0.000%	0
AU-corr	1108.540 ± 31.678	1108.54	0	10.66%	0.000%	0.000%	0
chisq	7335.100 ± 723.405	7335.10	0	70.56%	0.000%	0.000%	25
bonf	10393.850 ± 0.944	10393.85	0	99.99%	0.000%	0.000%	86

With all internal edges are of zero length, and the true set is all 10395 possible trees.

Coverage The AU tests struggle again with this configuration, with AU-correction this time predicting only 11% of the true trees. The original AU is much higher but still relatively low with a coverage of ~65% along with chi-square at a similar rate.

The following plots compare the coverage of all tests at an alpha-level of 0.05 dissected by tree type (a combination of tree configuration and taxa) and internal edge length. The first set of six plots is for all but the star tree case. The results have been considered in Sections 3.1- 3.2. The plots below are included to more clearly illustrate some of the points above. We can see that the variance between tree types makes the biggest difference for chisq whereas KH and SH hardly have any difference in coverage between the tree type. Another observation to emphasize is that the AU and KH tests were not designed to be conservative, so it is surprising that they tend to be. By contrast, the SH and Bonferroni approaches were expected to be conservative. As expected, at the lower internal edge lengths the tests generally perform worse, and only the very conservative tests, SH & KH had consistently high coverage as well as Bonferroni that was slightly lower, which was a surprise as KH was not designed to be conservative.

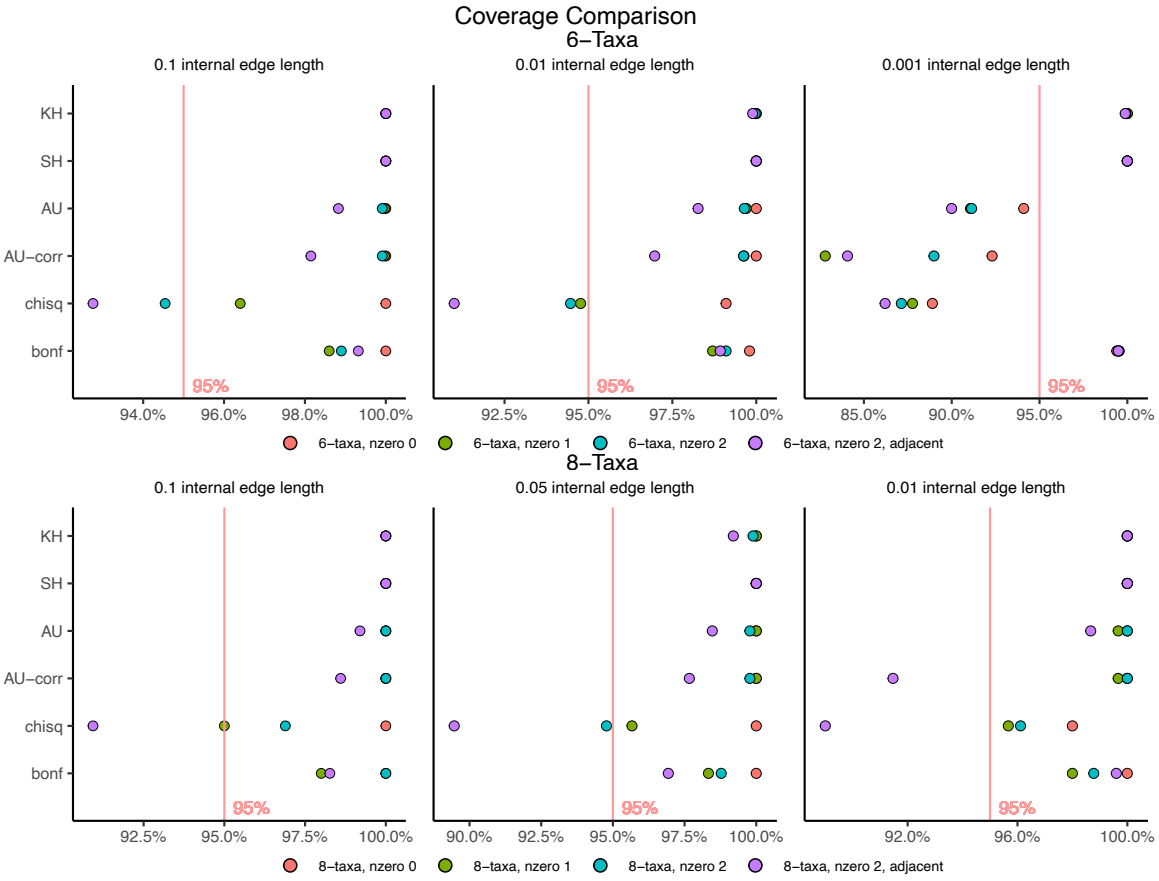

Figure 3.11: Coverage Comparison

Table 3.12: Coverage Comparison

Test	0.1 IEL	0.01 (6tx) / 0.05 (8tx) IEL	0.001 (6tx) / 0.01 (8tx) IEL
6-taxa, nzero 0, 6-taxa, nzero 1, 6-taxa, nzero 2, 6-taxa, nzero 2, adjacent			
6 Taxa			
KH	1.000, 1.000, 1.000, 1.000	1.000, 1.000, 1.000, 0.999	1.000, 1.000, 0.999, 0.999
SH	1.000, 1.000, 1.000, 1.000	1.000, 1.000, 1.000, 1.000	1.000, 1.000, 1.000, 1.000
AU	1.000, 1.000, 0.999, 0.988	1.000, 0.997, 0.996, 0.983	0.941, 0.911, 0.911, 0.900
chisq	1.000, 0.964, 0.945, 0.928	0.991, 0.948, 0.945, 0.910	0.889, 0.878, 0.871, 0.862
bonf	1.000, 0.986, 0.989, 0.993	0.998, 0.987, 0.991, 0.989	0.994, 0.995, 0.995, 0.995
AU-corr	1.000, 1.000, 0.999, 0.981	1.000, 0.996, 0.996, 0.970	0.923, 0.828, 0.890, 0.841
8 Taxa			
KH	1.000, 1.000, 1.000, 1.000	1.000, 1.000, 0.992, 0.999	1.000, 1.000, 1.000, 1.000
SH	1.000, 1.000, 1.000, 1.000	1.000, 1.000, 1.000, 1.000	1.000, 1.000, 1.000, 1.000
AU	1.000, 1.000, 0.992, 1.000	1.000, 1.000, 0.985, 0.998	1.000, 0.997, 0.987, 1.000
chisq	1.000, 0.950, 0.909, 0.969	1.000, 0.957, 0.895, 0.948	0.980, 0.957, 0.890, 0.961
bonf	1.000, 0.980, 0.983, 1.000	1.000, 0.983, 0.969, 0.988	1.000, 0.980, 0.996, 0.988
AU-corr	1.000, 1.000, 0.986, 1.000	1.000, 1.000, 0.977, 0.998	1.000, 0.997, 0.915, 1.000

[a] Where IEL here is internal edge length

and for the extreme case of the star tree:

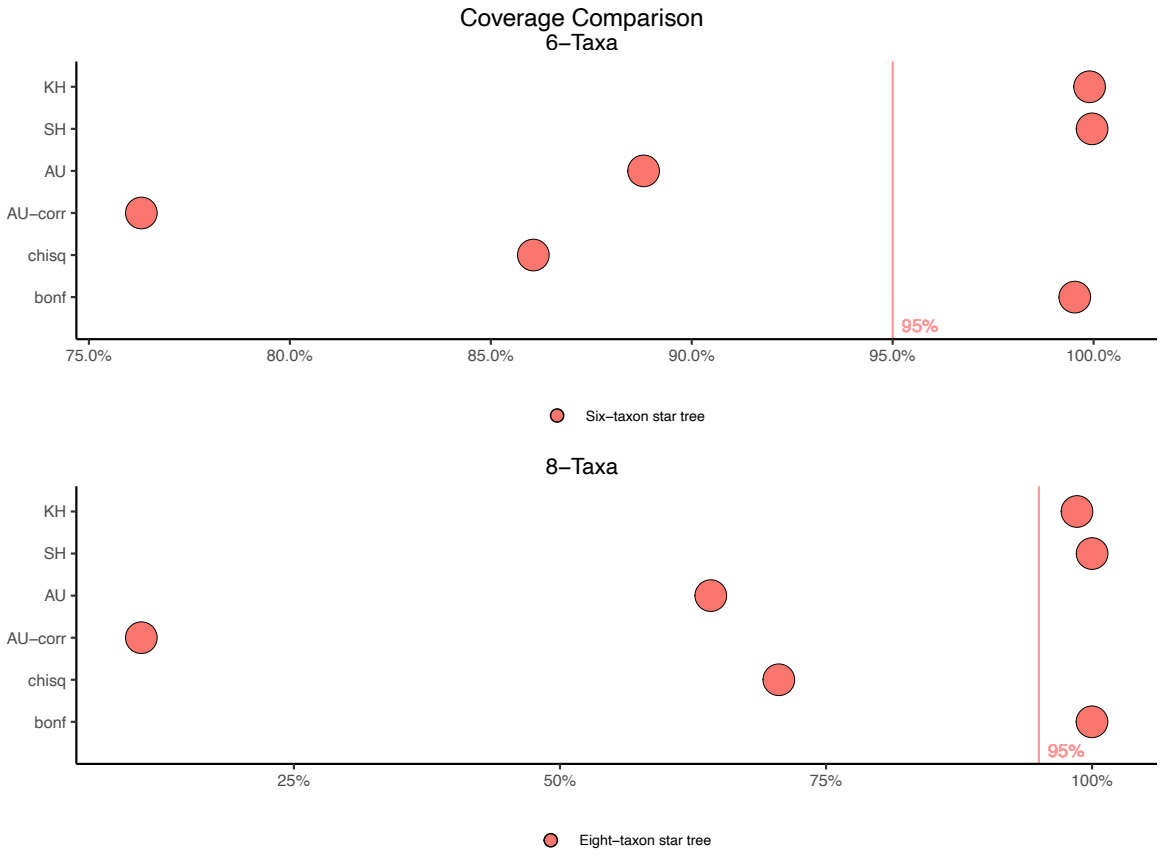

Figure 3.12: Coverage Comparison - Star tree

3.3 Confidence Level

The analysis up to this point has looked only at a confidence level of 0.95. This section examines the difference of setting the confidence level at 0.9 and 0.99 and comparing to 0.95. A side by side comparison of the results for 6-taxon simulations with 1 or 2 edge-lengths set to 0 with the varying confidence level is plotted in Figure 3.13

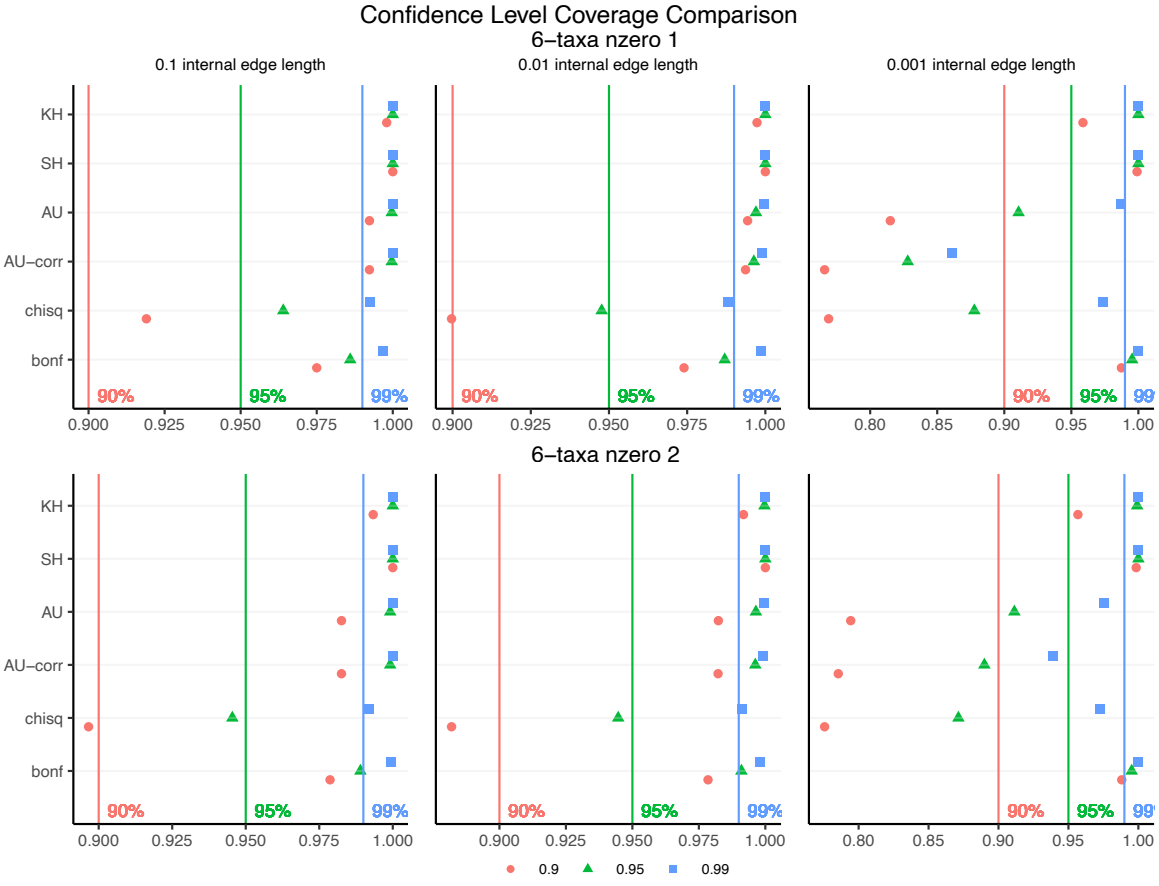

Figure 3.13: Varying Confidence Levels - 6-taxon

Table 3.13: 6-Taxon Side by Side Comparison with Varying Confidence Level

Test	internal edge length 0.1			internal edge length 0.01			internal edge length 0.001		
Confidence Level	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
6 taxa nzero 1									
KH	0.998	1.000	1.000	0.997	1.000	1.000	0.959	1.000	1.000
SH	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
AU	0.992	1.000	1.000	0.994	0.997	1.000	0.815	0.911	0.987
AU-corr	0.992	1.000	1.000	0.994	0.996	0.999	0.766	0.828	0.861
chisq	0.919	0.964	0.993	0.900	0.948	0.988	0.769	0.878	0.973
bonf	0.975	0.986	0.997	0.974	0.987	0.999	0.987	0.995	1.000
6 taxa nzero 2									
KH	0.993	1.000	1.000	0.992	1.000	1.000	0.957	0.999	1.000
SH	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000
AU	0.983	0.999	1.000	0.982	0.996	0.999	0.794	0.911	0.976
AU-corr	0.983	0.999	1.000	0.982	0.996	0.999	0.785	0.890	0.939
chisq	0.897	0.945	0.992	0.882	0.945	0.991	0.776	0.871	0.973
bonf	0.979	0.989	0.999	0.978	0.991	0.998	0.988	0.995	0.999

Repeating the same for 8-taxon:

Table 3.14: 8-Taxon Side by Side Comparison with Varying Confidence Level

Test	internal edge length 0.1			internal edge length 0.05			internal edge length 0.01		
Confidence Level	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
8 taxa nzero 1 (8cat1)									
KH	1.000	1.000	1.000	1.000	1.000	1.000	0.987	1.000	1.000
SH	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AU	0.987	1.000	1.000	1.000	1.000	1.000	0.993	0.997	1.000
AU-corr	0.987	1.000	1.000	1.000	1.000	1.000	0.993	0.997	1.000
chisq	0.900	0.950	0.980	0.927	0.957	1.000	0.897	0.957	0.987
bonf	0.970	0.980	0.993	0.980	0.983	1.000	0.967	0.980	0.993
8 taxa nzero 2 (8cat2sep)									
KH	0.994	1.000	1.000	0.987	0.999	1.000	0.991	1.000	1.000
SH	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AU	0.990	1.000	1.000	0.970	0.998	1.000	0.999	1.000	1.000
AU-corr	0.990	1.000	1.000	0.970	0.998	1.000	0.999	1.000	1.000
chisq	0.930	0.969	0.996	0.898	0.948	0.992	0.937	0.961	0.991
bonf	0.984	1.000	1.000	0.980	0.988	0.993	0.984	0.988	1.000

With respect to the difference between confidence levels, we can clearly see that in higher confidence levels the tests are more conservative, and in fact, most tests only start showing smaller confidence sets for level 0.9. As far as the difference between the tests goes, for the most part we see that the coverage is greater than the confidence level, where the main exception to that is Chi-square where it was consistently slightly below. It can be argued that Chi-Square's coverage is more inline with the level that the confidence regions were set to. Notice how its coverage is closer to the vertical threshold in every case. At the smaller internal edge length 0.001 for the 6-taxon case we also notice the AU Corr's coverage is below the confidence level

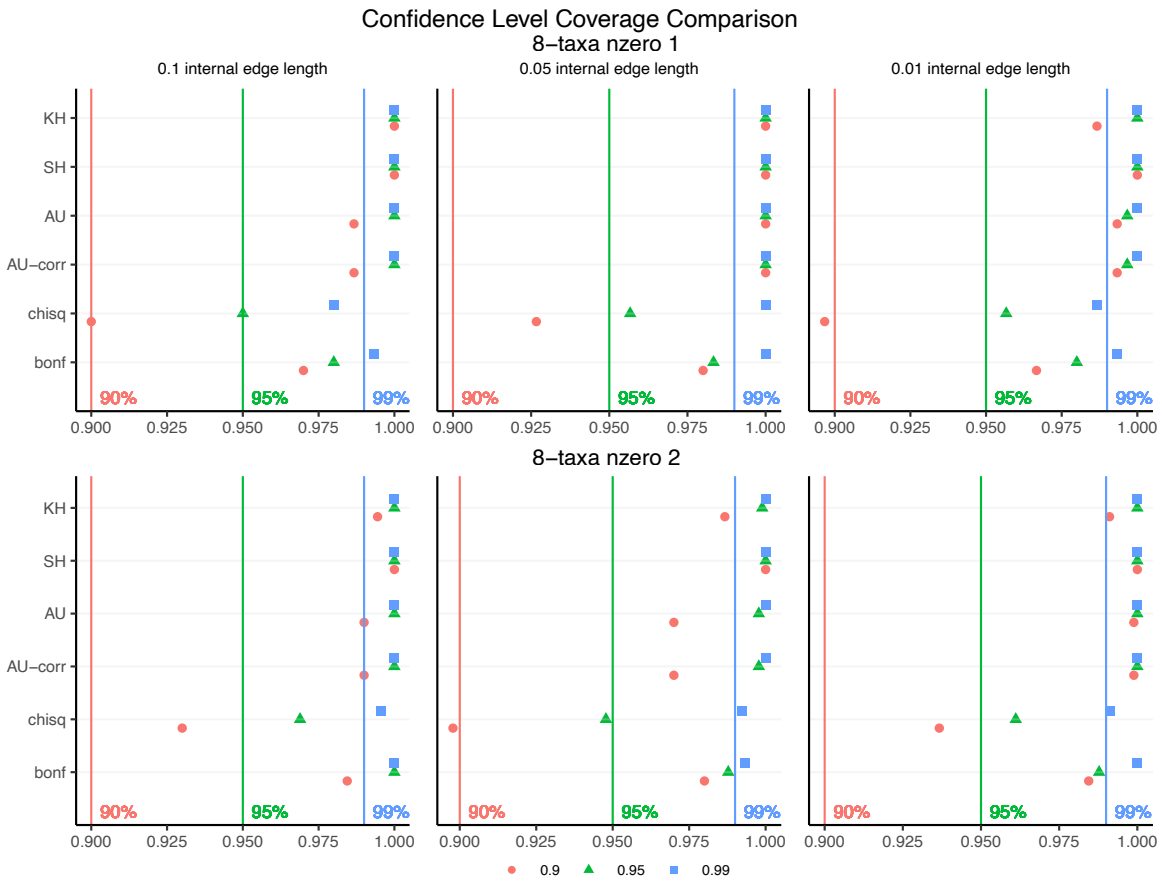

Figure 3.14: Varying Confidence Levels - 8-taxon

Chapter 4

Real Data Analysis

To illustrate how the tests compare in more complicated settings we consider several real data sets. The table below summarizes the properties of the data sets considered.

Name	Number of Taxa	Number of Sites	Type	Model Used *
HIV	6	2,000	nucleotide	GTR+G
Mammal	6	3,414	amino acid	mtRev+F+G
Amborella	24	15,688	amino acid	JTT+F+G
Rqua	148	195	amino acid	LG+C60+F+G

- The column 'Model Used' will be discussed later.

4.1 HIV

The HIV data is a set of six homologous sequences, each 2000 base pairs long, from the *gag* and *pol* genes for isolates of HIV-1 subtypes A, B, D and E: A1 (Q23), A2 (U455), B (BRU), D (NDK), E1 (90CF11697) and E2 (93TH057). The data set was first used in a testing context in Goldman, Anderson, and Rodrigo (2000). Again, here the analysis will consider confidence sets coming from analyses of all 105 possible 6-taxon trees.

Confidence sets of trees were extracted for each of the four species using each of the tests discussed in this thesis. For HIV that is 6-taxon for example, the file with possible trees contained 105 lines.

Table 4.2: HIV p-values

Tree	Likelihood Ratio	KH	SH	AU	AU Corr	Chi-square	Bonferroni	Agreement	Newick Format
26	0.000	0.800	1.000	0.837	0.837	1.000	1.000	6	(A1, (A2, (E2, E1)), (B, D));
64	3.964	0.200	0.869	0.262	0.262	0.005	0.015	4	(A1, (E2, E1), (A2, (B, D)));
52	4.266	0.177	0.867	0.258	0.258	0.003	0.010	4	(B, ((A2, A1), (E2, E1)), D);
99	21.621	0.012	0.398	0.005	0.005	0.000	0.000	1	(E2, (E1, (A2, A1)), (B, D));
18	21.665	0.011	0.397	0.002	0.002	0.000	0.000	1	(B, (E1, (E2, (A2, A1))), D);
79	25.856	0.005	0.316	0.003	0.003	0.000	0.000	1	(A1, (E2, (A2, E1)), (B, D));
43	26.739	0.002	0.302	0.001	0.001	0.000	0.000	1	(A1, (E1, (A2, E2)), (B, D));
102	31.702	0.003	0.226	0.003	0.003	0.000	0.000	1	(E2, (A1, E1), (A2, (B, D)));
86	32.037	0.002	0.220	0.002	0.000	0.000	0.000	1	(E1, (E2, A1), (A2, (B, D)));
95	38.148	0.002	0.147	0.003	0.003	0.000	0.000	1	(E2, (A1, (A2, E1)), (B, D));
74	38.171	0.002	0.146	0.001	0.000	0.000	0.000	1	(B, ((A2, E1), (E2, A1)), D);
104	38.942	0.001	0.136	0.002	0.000	0.000	0.000	1	(E2, (A2, (A1, E1)), (B, D));
100	38.999	0.001	0.135	0.004	0.000	0.000	0.000	1	(B, ((A2, E2), (A1, E1)), D);
37	39.432	0.001	0.131	0.000	0.000	0.000	0.000	1	(B, (E1, (A1, (A2, E2))), D);
91	39.432	0.001	0.131	0.000	0.000	0.000	0.000	1	(B, (E1, (A2, (E2, A1))), D);
25	153.829	0.000	0.000	0.000	0.000	0.000	0.000	0	(B, (A2, (E2, E1)), (A1, D));
27	153.829	0.000	0.000	0.000	0.000	0.000	0.000	0	((A1, B), (A2, (E2, E1)), D);
65	161.800	0.000	0.000	0.000	0.000	0.000	0.000	0	(A1, (E2, E1), (D, (A2, B)));

[a] Data is sorted by descending Tests Count, and then ascending Likelihood Ratio
[b] The table above has any probability greater than 0.05 with a dark green background, and a light green for probabilities greater than 0.01
[c] The Test Count on the right is the count of tests with a probability greater than 0.05

Please note that the **first value for KH should in fact be always equal to 1** by definition, as KH uses a t-test to compare every tree to the ML tree. The row in the table represents the tree with the least likelihood ratio, the ML tree. The program used to calculate the *p*-values, IQ-TREE, treats the ML tree as an observation and the result is not well defined, and that is the reason we see a value lower than 1.

We can see that for the trees, as the likelihood ratio increases, the *p*-values for the tests decrease. Another thing to point out is that the likelihood ratio jump from the third line, tree 52 to the fourth line, tree 99 is a substantial one, going from 4.266 to 21.621, and again going from tree 91 to tree 25 where the likelihood ratio went from 39.432 to 153.829.

We can also see some of the things we've seen so far; SH correction compared to KH is a lot more conservative, and the most conservative compared to the other tests. Bonferroni correction is a bit more conservative than Chi-square (probabilities for trees 64 and 52 are slightly larger).

We see that all tests are in agreement on tree #26, and all but chisq and Bonferroni

are in agreement on trees #52 and #64 as well. An interesting observation here is that Bonferroni gave a smaller confidence set than AU-corr which is the opposite of what we observed in the simulations. We can also observe that as in our simulations, SH behaves in the most conservative way, with a confidence set size of 15 trees.

Plotting trees 26, 52 and 64 to view similarities:

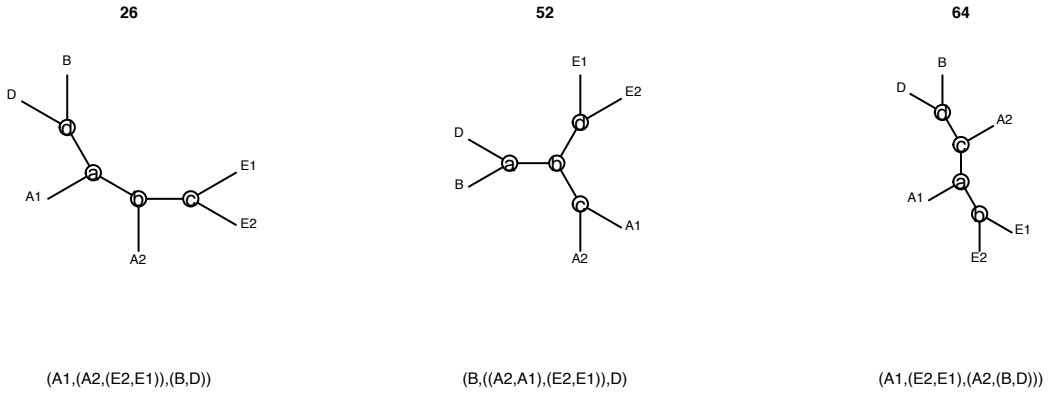

Figure 4.1: HIV - Trees In Agreement

4.2 Mammal

The mammalian mitochondrial data has been considered in Goldman, Anderson, and Rodrigo (2000) and Shimodaira (2002). It was also considered in Susko (2014), but in a context where only two trees were considered, where in this thesis the confidence sets come from analyses of all 105 possible 6-taxon trees.

The results from the tests:

where Hm = human, Sl = seal, Cw = cow, Rb = rabbit, Op = opossum, Ms = mouse

We observe a similar table to the HIV's here, except we can clearly see that KH's p-values are not perfectly inversely correlated to the likelihood ratio, and neither are the AU tests.

Table 4.3: Mammal p-values

Tree	Likelihood Ratio	KH	SH	AU	AU Corr	Chi-square	Bonferroni	Agreement	Newick Format
86	0.000	0.527	1.000	0.674	0.674	1.000	1.000	6	(Hm, (Sl, Cw), (Rb, (Op, Ms)));
91	0.597	0.473	0.965	0.631	0.631	0.275	0.618	6	(Op, (Hm, (Rb, (Sl, Cw))), Ms);
74	8.694	0.103	0.841	0.082	0.082	0.000	0.000	4	(Op, ((Rb, Hm), (Sl, Cw)), Ms);
90	18.020	0.095	0.622	0.052	0.052	0.000	0.000	4	((Op, Hm), (Rb, (Sl, Cw)), Ms);
85	20.476	0.067	0.562	0.121	0.121	0.000	0.000	4	(Op, Hm, ((Rb, Ms), (Sl, Cw)));
80	24.821	0.033	0.453	0.058	0.058	0.000	0.000	3	(Sl, Cw, (Op, (Hm, (Rb, Ms))));
32	21.068	0.058	0.547	0.013	0.013	0.000	0.000	2	(Op, (Rb, (Sl, Cw)), (Hm, Ms));
87	20.957	0.028	0.547	0.047	0.047	0.000	0.000	1	(Hm, (Sl, Cw), (Op, (Rb, Ms)));
8	28.529	0.021	0.377	0.019	0.019	0.000	0.000	1	(Sl, Cw, (Op, (Rb, (Hm, Ms))));
67	29.498	0.001	0.355	0.001	0.000	0.000	0.000	1	(Hm, (Sl, Cw), (Ms, (Rb, Op)));
68	33.287	0.007	0.285	0.021	0.021	0.000	0.000	1	(Sl, Cw, (Op, (Ms, (Rb, Hm))));
73	34.181	0.005	0.277	0.006	0.000	0.000	0.000	1	((Sl, Cw), (Rb, (Op, Hm)), Ms);
12	34.538	0.003	0.266	0.009	0.000	0.000	0.000	1	(Sl, Cw, ((Hm, Ms), (Rb, Op)));
75	40.040	0.001	0.179	0.008	0.000	0.000	0.000	1	((Sl, Cw), (Op, (Rb, Hm)), Ms);
66	42.188	0.000	0.152	0.017	0.000	0.000	0.000	1	((Sl, Cw), (Hm, (Rb, Op)), Ms);
102	55.111	0.001	0.043	0.001	0.000	0.000	0.000	0	(Sl, (Cw, Hm), (Rb, (Op, Ms)));
64	57.230	0.000	0.035	0.000	0.000	0.000	0.000	0	(Cw, (Sl, Hm), (Rb, (Op, Ms)));
92	77.135	0.000	0.002	0.002	0.000	0.000	0.000	0	(Sl, (Cw, Hm), (Op, (Rb, Ms)));

[a] Data is sorted by descending Tests Count, and then ascending Likelihood Ratio
[b] The table above has any probability greater than 0.05 with a dark green background, and a light green for probabilities greater than 0.01
[c] The Test Count on the right is the count of tests with a probability greater than 0.05

4.3 Amborella

The data set has been considered in Leebens-Mack, Soltis, and Soltis (2005), Lartillot, Brinkmann, and Philippe (2007), Wang, Susko, and Roger (2019) and Susko, Lincker, and Roger (2018). The amborella data includes a large number of taxa (24) and sites (15,688). Because there are so many taxa, the trees considered for inclusion in the confidence set are not all trees. Rather they are all of the trees estimated in 100 bootstrap samples.

In this example there is more contrast between the tests; at a 0.95 confidence level Chi-Square and Bonferroni only predict a single tree #3 whereas KH predicts 6, SH predicts 35, and both the AU tests predict 13.

Table 4.5 summarizes the tests agreement on the data in Table 4.4 where each element in the matrix represents a count of trees in agreement in the two tests (row header and column header), and the diagonal is the number of total trees in the confidence set for the test.

Figures 4.2 and 4.3 consider the first two trees with the smallest likelihood ratio from Table 4.4.

In the tree in Figure 4.3, *Amborella* (an evergreen shrub) branches at the base of the

Table 4.4: Amborella p-values

Tree	Likelihood Ratio	KH	SH	AU	AU Corr	Chi-square	Bonferroni	Agreement
3	0.000	0.607	1.000	0.878	0.878	1.000	1.000	6
36	6.015	0.393	0.960	0.640	0.640	0.002	0.012	4
1	10.289	0.246	0.936	0.422	0.422	0.000	0.000	4
24	15.082	0.132	0.855	0.206	0.206	0.000	0.000	4
29	31.772	0.101	0.644	0.071	0.071	0.000	0.000	4
28	47.210	0.095	0.382	0.100	0.100	0.000	0.000	4
20	25.152	0.036	0.713	0.110	0.110	0.000	0.000	3
10	28.432	0.032	0.664	0.074	0.074	0.000	0.000	3
18	28.437	0.014	0.650	0.052	0.052	0.000	0.000	3
9	33.080	0.015	0.579	0.051	0.051	0.000	0.000	3
11	35.303	0.042	0.539	0.067	0.067	0.000	0.000	3
8	38.859	0.035	0.483	0.054	0.054	0.000	0.000	3
22	41.264	0.034	0.437	0.050	0.050	0.000	0.000	3
2	25.799	0.026	0.715	0.005	0.005	0.000	0.000	1
35	33.738	0.023	0.548	0.046	0.046	0.000	0.000	1
6	39.188	0.022	0.475	0.034	0.034	0.000	0.000	1
14	43.572	0.017	0.412	0.016	0.016	0.000	0.000	1
17	44.900	0.002	0.393	0.007	0.007	0.000	0.000	1
7	50.535	0.005	0.323	0.007	0.000	0.000	0.000	1
12	54.375	0.003	0.277	0.011	0.011	0.000	0.000	1
16	54.428	0.001	0.278	0.000	0.000	0.000	0.000	1
32	56.864	0.003	0.260	0.010	0.010	0.000	0.000	1
4	58.964	0.002	0.228	0.010	0.000	0.000	0.000	1
21	59.596	0.002	0.243	0.006	0.000	0.000	0.000	1
23	61.517	0.001	0.219	0.024	0.000	0.000	0.000	1
41	62.434	0.003	0.221	0.012	0.012	0.000	0.000	1
30	64.028	0.006	0.197	0.009	0.000	0.000	0.000	1
42	70.721	0.003	0.174	0.025	0.000	0.000	0.000	1
33	72.536	0.002	0.142	0.008	0.000	0.000	0.000	1
26	74.587	0.002	0.145	0.005	0.000	0.000	0.000	1
19	79.265	0.000	0.105	0.010	0.000	0.000	0.000	1
38	80.725	0.020	0.106	0.003	0.003	0.000	0.000	1
25	82.786	0.000	0.101	0.010	0.000	0.000	0.000	1
40	83.915	0.000	0.089	0.018	0.000	0.000	0.000	1
31	91.724	0.000	0.076	0.007	0.000	0.000	0.000	1
5	128.141	0.002	0.019	0.000	0.000	0.000	0.000	0
34	138.854	0.001	0.008	0.001	0.000	0.000	0.000	0
13	142.383	0.001	0.007	0.000	0.000	0.000	0.000	0

[a] Data is sorted by descending Tests Count, and then ascending Likelihood Ratio
[b] The table above has any probability greater than 0.05 with a dark green background, and a light green for probabilities greater than 0.01
[c] The Test Count on the right is the count of tests with a probability greater than 0.05

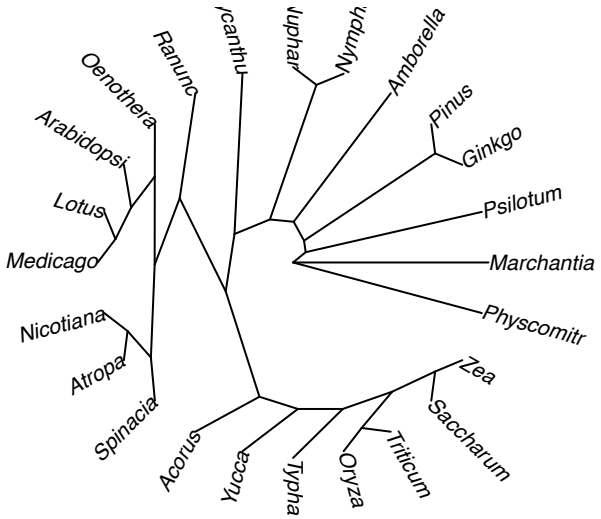

Figure 4.2: Amborella - Trees with Smallest Likelihood Raio, Tree 3

Table 4.5: Amborella Tests Agreement

	KH	SH	AU	AU Corr	Chi-square	Bonferroni
KH	6	6	6	6	1	1
SH	6	35	12	12	1	1
AU	6	12	12	12	1	1
AU Corr	6	12	12	12	1	1
chisq	1	1	1	1	1	1
bonf	1	1	1	1	1	1

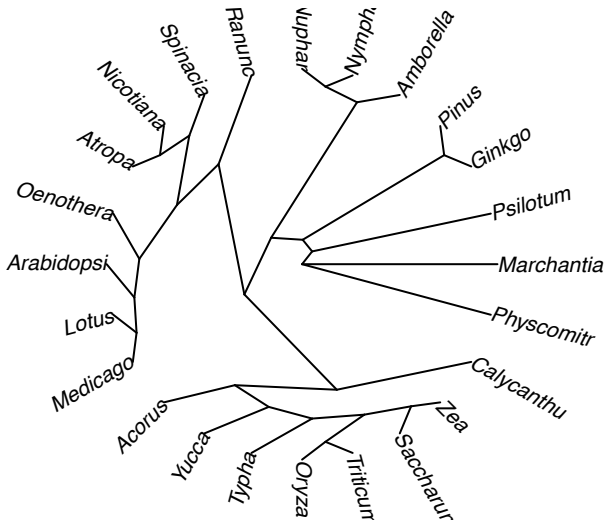

Figure 4.3: Amborella - Trees with Smallest Likelihood Ratio, Tree 36

angiosperms (flowering plants including *Spinacia* (spinach) to the *Nuphar*, *Nymphia* clade (water lilies). By contrast in Figure 4.3 *Amborella* branches with the water lilies. Which of these placements is correct has been contentious (cf. (Drew et al. 2014); (Goremykin et al. 2015)). We see here that for many of the tests, the differences can be explained by sampling variation whereas the chi-square and Bonferroni approaches suggest that the tree in Figure 4.2 is correct.

Table 4.6: Rqua p-values

Tree	Likelihood Ratio	KH	SH	AU	AU Corr	Chi-square	Bonferroni	Agreement
1	0.000	0.519	1.000	0.858	0.858	1.000	1.000	6
133	0.936	0.481	1.000	0.691	0.691	0.931	0.995	6
76	2.404	0.425	0.998	0.697	0.697	0.904	0.991	6
27	2.578	0.405	0.999	0.583	0.583	0.821	0.968	6
117	2.707	0.429	0.997	0.594	0.594	0.988	1.000	6
155	2.727	0.417	0.999	0.681	0.681	0.941	0.997	6
161	2.775	0.420	0.999	0.651	0.651	0.902	0.990	6
20	2.826	0.419	0.998	0.691	0.691	0.933	0.995	6
192	2.943	0.423	1.000	0.708	0.708	0.969	0.999	6
167	2.962	0.416	0.998	0.668	0.668	0.981	1.000	6
171	3.191	0.424	0.999	0.630	0.630	0.956	0.998	6
121	3.299	0.412	0.999	0.575	0.575	0.679	0.967	6

[a] Data is sorted by descending Tests Count, and then ascending Likelihood Ratio
[b] The table above has any probability greater than 0.05 with a dark green background, and a light green for probabilities greater than 0.01
[c] The Test Count on the right is the count of tests with a probability greater than 0.05

4.4 Rqua

The Rqua data set was considered in Stairs et al. (2018). This is unusual data. There are 148 taxa and 195 sites. Since the total number of possible trees is very large only 216 trees were considered. The were obtained through a combination of bootstrapping but also include trees created from apriori hypothesized topological relationships. Additional information is available in Stairs et al. (2018). So it is a setting that is very different from the usual case where there are a lot of sites and few taxa. It pushes the boundaries of how the methods might be used.

The full detailed probability table per tree for each of the tests is too large to display here.

Table 4.7: Rqua Tests Agreement

	KH	SH	AU	AU Corr	Chi-square	Bonferroni
KH	202	202	201	200	135	150
SH	202	209	203	202	135	150
AU	201	203	203	202	135	150
AU Corr	200	202	202	202	135	150
chisq	135	135	135	135	135	135
bonf	150	150	150	150	135	150

Table 4.7 shows a matrix of the number of overlapping trees above the confidence level between each two-test pair.

Chapter 5

Discussion

This thesis compared six different phylogenetic ML tests, pointing out the strengths and weaknesses of each under various scenarios.

- **KH** had surprisingly high coverage even though it does not adjust for selection bias at all. KH, similarly to SH performed very well in the star-tree cases, although it was not perfect. In the six-taxon case where two adjacent edge-lengths were set to 0, it was the only one with SH (at high internal edge-length) to have a mean set size of 15 (the number of true trees for that case), where the rest averaged below that. Overall based on the results in this thesis, it seems as if the SH correction to KH isn't necessary, as it has similar coverage to SH, with mean set size generally less than or equal to SH's which is better.
- **SH** seems too conservative usually, overshooting the confidence region to achieve higher coverage than KH on the expense of additional wrong trees. SH's mean set size is conservative relative to the other tests, especially in the case of 6-taxon none of the edge-lengths were set to 0, and all the 8-taxon trees (star tree case all tests are conservative just as much). The only case where SH might be considered to perform better than other tests is the extreme case of the star tree (both 6 and 8 taxa) with coverage of 100%.
- **AU** - Throughout the simulations we observed a difference between AU and AU-corr, which means that the case where BP = 0 happens frequently. Generally the AU tests performed somewhere between SH/KH and Chi-square/Bonferroni. It seems hard to interpret as the test definition is a bit convoluted and there is some question as to whether it can really be expected to give correct coverage in a phylogenetic setting. AU's mean set size is also conservative, but only in the 8-taxon tree simulations. In the 6-taxon tree it is very comparable to the other tree configurations, and only SH's set sizes stands out. Another interesting

thing is that for the most part when decreasing the internal edge-length for any given test, the set size increases (the test is more conservative as it's harder to distinguish between the branches and they look more like a star tree as we shrink the internal edge-length), however, AU's set size actually decreases with internal edge-length in the 8-taxon cases. In the 8-taxon simulations we observed a high variability in AU's mean set size. This may be due to the fact that, as shown in Section 2.2, small variations in BP can cause large variations in the $p-$value. The large variance for AU can be explained by comparing to the relatively small variance of AU-corr. The two methods differ only when $BP = 0$. For reasons discussed in Section 2.2, the AU p-values become highly unstable in this case. We believe that the instabilities arising with small BP_r provide the explanation for why AU was often conservative in the eight-taxon case. Due to the instability, small BP can lead to large p-values and small BP was more likely in the eight-taxon rather than six-taxon case due to the larger number of trees under consideration.

- **AU Corr** - AU Corrected (with a BP correction) was a little less conservative than the original AU. In 6-taxon it was consistent in being below the 95% coverage at 0.001 internal edge-length (unlike the rest), performed the worst in coverage for the 6-taxon star tree, as well as the 8-taxon star tree case. Adjusting for the difficulties with small BP discussed in Section 2.2 by simply setting the AU to 0 when $BP = 0$ may be an over-adjustment when there are a large number of competing trees.
- **Chi-square** was a bit too aggressive with low set sizes. Its coverages were almost in all cases below the 95%, but at the same time it was closer to the 95% mark than the other tests, which is more inline with the expectation. The Chi-square test's coverage was consistently below-95% in many of the cases: eight taxon tree with two non-adjacent zero-length edges, eight-taxon star-tree, six-taxon star tree, six-taxon tree with two adjacent zero-length edges and in all cases when internal edge-length was small. Chi-square had the best exact trees metric, however.
- **Bonferroni** correction to Chi-square performed very well in terms of coverage, and although more conservative than Chi-square, its set didn't introduce as

many excess trees as KH did. Bonferroni's correction is still above the 95% coverage confidence level while chisq tends to have a lower coverage when the internal edge-lengths in the simulating tree were smaller or even when we have harder cases to distinguish (higher number of internal edge lengths set to zero), which is where we expect it to be in order not to exclude a true tree in the resulting confidence set. This was consistent in both the 6-taxon and 8-taxon case. For the star-tree chisq's coverage was hovering around 85% when all trees are right in the 6-taxon while Bonferroni maintained a 100% coverage. The mean set size of Bonferroni was higher than chisq's for that reason. Curiously, both Bonferroni and chi-square gave relatively small set sizes for the real data. The exact reasons for this are not clear but it may be that the inevitable presence of model misspecification, which was absent in simulations, plays a role.

5.1 Final Thoughts

Using the the arguments in the discussion above, we conclude that SH is not a necessary adjustment to KH. Indeed, KH was surprisingly so conservative that, although it frequently included all of the true trees in its sets, that came at the cost of large Type II error (excess of trees) to the point that it sometimes included all possible trees in the confidence set.

Although the AU test was originally motivated as being higher-order correct, the argument leading to this result was later found to be problematic in phylogenetic settings (Susko 2009). Consistent with Susko (2009) we did not find AU to have approximately correct coverage, as BP is not first order correct when used in a phylogenetic setting. A major difficulty for AU arises when BP is expected to be small for the trees under test. The AU test becomes unstable in such cases and coverage and set sizes were found to be highly variable as a consequence. Small BP is expected when there are many trees being considered for inclusion in the confidence set. In such settings we do not recommend use of the AU test.

Bonferroni correction was a more conservative version of Chi-square and, as expected, tended to have excess coverage and excess trees, Chi-square seemed to have a balance between Type I and Type II error, although its Type I error could be large in the case

of a very poorly resolved tree. We conclude from the findings in this thesis that overall the chi-square would be the best to use perhaps with Bonferroni as a conservative cross-check.

Chapter 6

Bibliography

Adachi, J., and M. Hasegawa. 1996. "Model of Amino Acid Substitution in Proteins Encoded by Mitochondrial Dna." *Journal of Molecular Evolution* 42 (4): 459–68. https://doi.org/10.1007/BF02498640.

Drew, B. T., B. R. Ruhfel, S. A. Smith, M. J. Moore, B. G. Briggs, M. A. Gitzendanner, P. S. Soltis, and D. E. Soltis. 2014. "Another Look at the Root of the Angiosperms Reveals a Familiar Tale." *Systematic Biology* 63 (3): 368–82. https://doi.org/10.1093/sysbio/syt108.

Efron, B. 1982. *The Jacknife, the Bootstrap and Other Resampling Plans*. CBMS-NF regional conference series in applied mathematics, volume 38 Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611970319.

Efron, B., E. Halloran, and S. Holmes. 1996. "Bootstrap Confidence Levels for Phylogenetic Trees." *Proceedings of the National Academy of Sciences of the United States of America* 93 (14): 7085–90. http://www.jstor.org/stable/39541.

Efron, B., and R. Tibshirani. 1998. "The Problem of Regions." *The Annals of Statistics* 26 (5): 1687–1718. http://www.jstor.org/stable/120017.

Felsenstein, J. 1985. "Confidence Limits on Phylogenies: An Approach Using the Bootstrap." *Evolution* 39 (4): 783–91. http://www.jstor.org/stable/2408678.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. "Likelihood-Based Tests of Topologies in Phylogenetics." *Systematic Biology* 49 (4): 652–70. http://www.jstor.org/stable/2585286.

Goremykin, V. V., S. V. Nikiforova, D. Cavalieri, M. Pindo, and P. Lockhart. 2015. "The Root of Flowering Plants and Total Evidence." *Systematic Biology* 64 (5): 879–91.

https://doi.org/10.1093/sysbio/syv028.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. "The Rapid Generation of Mutation Data Matrices from Protein Sequences." *Bioinformatics* 8 (3): 275–82. https://doi.org/10.1093/bioinformatics/8.3.275.

Kishino, H., and M. Hasegawa. 1989. "Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologiesfrom Dna Sequencedata, and the Branching Order in Hominoidea." *Journal of Molecular Evolution* 29: 170–79.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. "Maximum Likelihood Inference of Protein Phylogeny and the Origin of Chloroplasts." *Journal of Molecular Evolution* 31 (2): 151–60. https://doi.org/10.1007/BF02109483.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. "Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model." *BMC Evolutionary Biology* 7 (1): S4. https://doi.org/10.1186/1471-2148-7-S1-S4.

Le, S. Q., and O. Gascuel. 2008. "An Improved General Amino Acid Replacement Matrix." *Molecular Biology and Evolution* 25 (7): 1307–20. https://doi.org/10.1093/molbev/msn067.

Leebens-Mack, J., D. E. Soltis, and P. S. Soltis. 2005. "Plant Reproductive Genomics at the Plant and Animal Genome Conference." *Comparative and Functional Genomics* 6 (3): 159–69. https://doi.org/https://doi.org/10.1002/cfg.469.

L. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2014. "IQ-Tree: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Miklós, I., Á. Novák, R. Satija, R. Lyngsø, and J. Hein. 2009. "Stochastic Models of Sequence Evolution Including Insertion—Deletion Events." *Statistical Methods in Medical Research* 18 (5): 453–85. https://doi.org/10.1177/0962280208099500.

R. Ota, P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. "Appropriate Likelihood Ratio Tests and Marginal Distributions for Evolutionary Tree Models with Constraints on Parameters." *Molecular Biology and Evolution* 17 (5): 798–803. https://doi.org/10.1093/oxfordjournals.molbev.a026358.

Sheldon, R. M. 1996. *Stochastic Processes*. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley. https://books.google.de/books?id=ImUPAQAAMAAJ.

Shi, X., H. Gu, E. Susko, and C. Field. 2005. "The Comparison of the Confidence Regions in Phylogeny." *Molecular Biology and Evolution* 22 (11): 2285–96. https://doi.org/10.1093/molbev/msi226.

Shimodaira, H. 2002. "An Approximately Unbiased Test of Phylogenetic Tree Selection." *Systematic Biology* 13 (3): 492–508. https://doi.org/10.1080/10635150290069913.

Shimodaira, H., and M. Hasegawa. 1999. "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference." *Molecular Biology and Evolution* 16 (8): 1114–4. https://doi.org/10.1093/oxfordjournals.molbev.a026201.

Stairs, C., L. Eme, S. Muñoz-Gómez, A. Cohen, G. Dellaire, J. Shepherd, J. Fawcett, and A. Roger. 2018. "Microbial Eukaryotes Have Adapted to Hypoxia by Horizontal Acquisitions of a Gene Involved in Rhodoquinone Biosynthesis." *eLife* 7 (April). https://doi.org/10.7554/eLife.34292.

Susko, E. 2009. "Bootstrap Support Is Not First-Order Correct." *Systematic Biology* 58 (2): 211–23. https://doi.org/10.1093/sysbio/syp016.

———. 2014. "Tests for Two Trees Using Likelihood Methods." *Molecular Biology and Evolution* 31 (4): 1029–39. https://doi.org/10.1093/molbev/msu039.

Susko, E., L. Lincker, and A. J. Roger. 2018. "Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models." *Molecular Biology and Evolution* 35 (5): 1266–83. https://doi.org/10.1093/molbev/msy026.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 2004. "Phylogenetic Inference." Edited by C. Moritz D. Hillis and B. K.Mable, 407–514.

Tavaré, S. 1986. "Some Probabilistic and Statistical Problems on the Analysis of DNA Sequences." *Lectures on Mathematics in the Life Sciences* 17: 57–86.

Wang, H., E. Susko, and A. J. Roger. 2019. "The Relative Importance of Modeling Site Pattern Heterogeneity Versus Partition-Wise Heterotachy in Phylogenomic Inference." *Systematic Biology* 68: 1003–19.

Wu, J. 2010. "Distance Method Adjustments and a Test for General Heterotachy in Phylogenetic Estimation." Doctor of Philosophy, Dalhousie University, Halifax, Nova Scotia.

Yang, Z. 1993. "Maximum-Likelihood Estimation of Phylogeny from Dna Sequences When Substitution Rates Differ over Sites." *Molecular Biology and Evolution* 10: 1396–1401.

———. 1994. "Maximum Likelihood Phylogenetic Estimation from Dna Sequences with Variable Rates over Sites: Approximate Methods." *Journal of Molecular Evolution* 39 (3): 306–14. https://doi.org/10.1007/BF00160154.