Tree Evaluation and Robustness Testing

C'est nous

September 1, 2017

Abstract

100 mots

Contents

1 Motivation

1.1 Applications of Phylogenies

Molecular phylogenetics is a lively field of research with a number of practical applications. Reconstructing large phylogenies, such as the bird [ref] or mammal phylogeny [ref], is of intrinsic interest to evolutionary biologists, but those phylogenies also *the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically* (Felsenstein, 2004). They arise frequently in comparative genomics, conservation issues (Bordewich et al., 2008), functional prediction of genes (Eisen, 1998) and more generally are at the heart of Phylogenetic Comparative Methods (Revell et al., 2008; Pennell and Harmon, 2013). Most, if not all, applications of phylogenetics have in common that they rely on accurate phylogenetic estimates and it is crucial to validate the tree as different trees can lead to vastly different conclusions (*e.g.* one ancestral emergence of a phenotypic trait versus many independent ones [ref]).

With the advent of molecular data and increased formalism of the field (Gascuel, 2005), modern phylogenetic reconstruction is now essentially a statistical inference problem. Many popular reconstruction softwares such as PhyML (Guindon and Gascuel, 2003), RAxML (Stamatakis, 2006), FastTree (Price et al., 2010) or MrBayes (Ronquist and Huelsenbeck, 2003) produce a statistical estimate of the tree and we also frame validation in a statistical framework. We first discuss the different sources of inaccuracies in the reconstructed tree (Section 2) and distinguish between natural variability ($\simeq$ variance) and modeling errors ($\simeq$ bias). We then briefly describe and discuss popular support values (Section 3) aimed at validating a tree. The variability, when expressed as a forest of trees, can be summarized in order to produce robust tree estimates (Section 4). Finally, we review promising developments in the field of robustness validation (Section 5)

1.2 Validating the Tree

Many inference methods return a single focal tree on and validation is most often concerned with it. A tree is a complex object that encodes the evolutionary relationship of a set of species. There are two families of validation methods based on (i) the scale at which validation is performed and (ii) whether the tree is considered alone or with respect to other trees.

- The first family is **local** and grounded on the observation that a tree is uniquely determined by its branches. A tree can thus be validated by computing a *support value* for each of its branch. Support values tell us which parts of the tree are *reliable*, in a yet to be defined way.

- The second family is **global** and considers the focal tree as a whole. It compares it to a set of alternatives using statistical tests. The tests tell us whether the tree is strictly better (*i.e.* a better fit to the molecular data) than the alternatives.

The two approaches have a different focus but are complimentary. In particular, testing the focal tree against alternatives derived from it can be used to compute support values.

1.3 Robust Estimate

Most softwares return the best tree for a given criteria (*e.g.* likelihood for PhyML and RAxML). Support values tell us whether the tree is reliable and tests tell us whether it's much better than second best or other alternatives.

In the presence of outlier data, the best tree may be very sensitive to a few data points: slight changes in the molecular data may dramatically change the best tree, with deep clades moving from one position in the tree to another (Bar-Hen et al., 2008). Since molecular data are inherently noisy, it is interesting to produce **robust** trees that nearly, but not completely, optimize the criteria while being resilient to small changes in the molecular data.

A straighforward way to build a robust estimate is to start from a forest of *good* trees and summarize them in some way to build a **consensus** tree. The forest can consist of trees that are only slightly worse than the best tree (*e.g* bayesian consensus) or that are inferred from slightly perturbed data (*e.g.* bootstrap consensus).

2 Sources of Error

Genome analysis indicate that a substantial fraction of genes yield phylogenies that are in strong conflict with one another Broadly speaking, these conflicts derive from two sources: (i) The inability of traditional phylogenetic reconstruction methods to deal with the complexity of molecular evolution on the largest scale (methodological sources) and/or (ii) Real biological events such as lateral gene transfer (LGT) of whole genes between genomes or transfer of subgene fragments within and between genomes (biological sources).

Methodological factors affecting phylogenetic reconstruction include the choice of optimality criterion, limited data availability, taxon sampling and specific assumptions in the modelling of sequence evolution. Biological processes such as the action of natural selection or genetic drift may cause the history of the genes under analysis to obscure the history of the taxa. The large number of potential explanations for the presence of incongruence in molecular phylogenetic analyses makes decisions on how to handle conflict in larger sets of molecular data difficult Rokas et al. (2003b)

2.1 Sampling Errors

Biologial source of errors are very diverse and among other we may cite contamination, frameshift events, incorrect annotations, erroneous chimerical sequences, wrong orthology assessment, horizontal gene transfer, gene conversion, incomplete lineage sorting or hybridization, etc. (see Philippe et al. (2017) for example)

Due to the limitations of ancient sequencing technologies, sequencing errors were frequent and sequence quality was quite variable in these early datasets. High-throughput sequencing has greatly improved sequence quality, mainly through the large coverage of each nucleotide, but has simultaneously flooded researchers with an amount of data that is impossible to handle by hand. The most frequent errors observed are sequencing errors (especially for transcriptomic data) and annotation errors (especially for genomic data). Errors due to the inclusion of non-orthologous sequences in phylogenomics can have drastic consequences on the final results (Laurin-Lemay et al. (2012); Philippe et al.

(2011b)). Contamination is not the only source of non-orthology, paralogy being a common underhand source of issues, given the high frequency of gene/genome duplication, gene conversion and gene loss. Sequence contamination can occurr at the sampling step or at the laboratory or sequencing steps (cross-contamination).

The importance of a correct alignment in phylogenetic inference has long been pointed out (Morrison and Ellis (1997); Ogden and Rosenberg (2006); Talavera and Castresana (2007); Wong et al. (2008)). Yet, due to the lack of a tractable model of sequence evolution in the presence of insertion and deletion events (indels), the criteria optimized by alignment software are mostly ad hoc and based on the simplistic assumptions that homologous characters should be similar and that indels are rare events.

2.2 Modeling Errors

Every model is only a rough approximation to the reality of molecular evolution. At first, the assumption of independent and identical evolutionary forces across sites is certainly not true in reality. What happens at one site will depend quite critically on where that site lies in a protein or structural RNA. One approach that has been developed is to allow the rate of evolution to vary across the sites (Goldman and Yang (1994)). In essence the rate at the site is introduced into the model as a random effect from a given distribution (often a gamma distribution). Felsenstein and Churchill (1996) extended this approach to allow the rate to depend on the rate at the neighbouring sites along the sequence using a hidden Markov model. This model is expected to work well if most autocorrelation of sites occurs at neighbouring positions in the linear sequence. However, the situation in real molecules is more complex than this with serial dependence not expected as the main type of dependence..

The rate matrices used for amino acid substitutions (JTT or PAM)are derived by averaging over patterns observed in thousands of sites. However it is clear (Halpern and Bruno (1998); Parisi and Echave (2001); Susko et al. (2002)) that amino acid frequencies at sites strongly deviate from the frequencies expected under the JTT matrix. Lartillot and Philippe (Lartillot and Philippe (2004)) implemented a Bayesian mixture model which allows the inference of site specific rate matrices. These studies indicate that relaxing the assumption that all sites evolve according to the same rate matrices is of key importance for accurate phylogenetic inference from proteins. However, it seems likely that serious, difficult to diagnose, problems stemming from over-parameterization could result from the extremely parameter-rich models (see Rannala (2002) for a discussion of over-parameterization of phylogenetic models in the context of Bayesian inference).

Most models assume stationarity, homogeneity and reversibility of the stochastic process of sequence change, even though it has been clear for many years that these assumptions are violated when considering the evolutionary process over very long time scales. For instance, the set of nucleotide or amino acid positions free to vary, and the evolutionary rates at which they vary, and the frequencies of nucleotides, codons and amino acids can periodically change over the tree, due to functional or structural alterations in the molecule or differing selective forces in the organisms' genomes.

Other cases where the homogeneity, stationarity and/or time-reversibility of substitution models are violated include situations where the underlying state frequencies (whether they be nucleotides, amino acids or codons) in the genes of different organisms vary over the tree. If distantly related lineages begin to display similar state frequencies, these lineages can be artefactually grouped together (see for example Foster (2004); Jermiin et al. (2004)).

Clearly the existence of lateral gene transfer (LGT) is a major deviation form the standard model of vertical descent. Approaches are being developed to incorporate LGT into evolutionary models of genomes (see Zhaxybayeva et al. (2004) for example).

3 Robustness

As discussed in Section 1, support values are a popular way to validate a focal tree. We present here the most popular ones before describing other methods to validate or enhance a tree.

3.1 Support Values

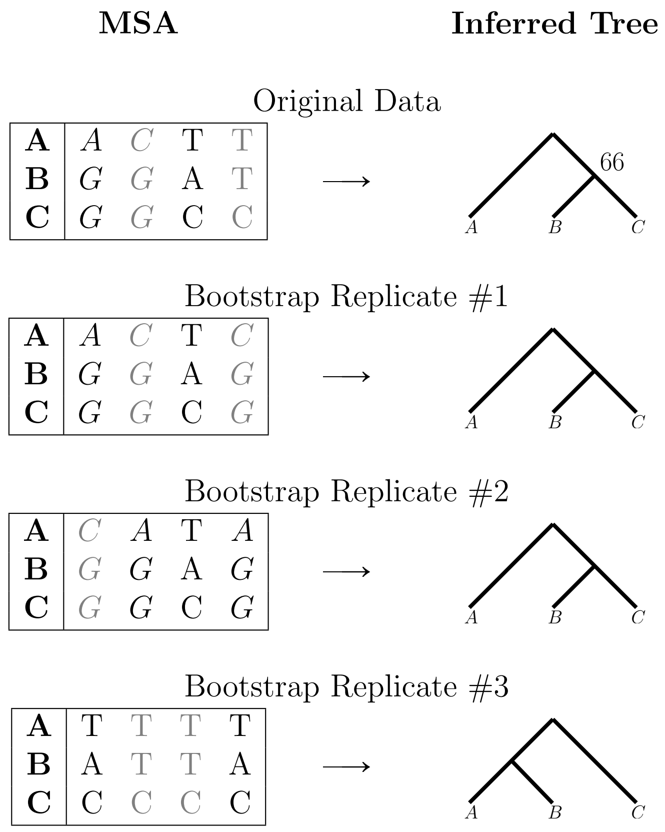
3.1.1 Bootstrap

Bootstrap values Felsenstein (1985) are probably the most popular and easiest to understand support values. Bootstrap involves resampling from one's molecular data with to create fictional datasets, called *bootstrap replicates*, of the same size. Specifically, the molecular data is typically organized as a multiple sequence alignment (MSA) of s species $\times$ n characters. Since most models assume independent characters, we generate a replicate by sampling n characters, with replacement, from the original MSA and do this B times. Note that in each replicate, some characters are sampled more than once and some left out entirely. The B replicates are used to estimate a forest of B bootstrap trees (one per replicate). Finally the bootstrap value (BP) of a branch of the original tree is its frequency of occurrence in the forest. The process is illustrated in Figure 1.

Intuitively, the variation obtained by resampling n sites from the original data should be the same as the variation obtained by sampling n new characters. Bootstrap values capture, among other, the *sampling* variability induced short MSA. When n increases, so does BP in general and it is quite common to achieve very high values for all branches when working on genome-scale alignments (Rokas et al., 2003a).

BP provides a guide for the amount of support a branch has: branches with high BP occur more often and are more reliable than those with low BP. Although it might be tempting to interpret BP as the probability that a branch is present in the (unknown) true tree, this is not the case in general. Zharkikh and Li (1992) showed in a simple case that BP is biased and underestimates that probability. Using simulation studies, Hillis and Bull (1993) showed that BP values as small as 70% could reflect highly supported branches. Many studies (Felsenstein and Kishino, 1993; Efron et al., 1996; Susko, 2008, 2010) examined the theoretical properties of bootstrap values and concluded that they are indeed biased. This bias is partly induced by the peculiar geometry of tree space (see Billera et al. (2001); Susko (2010) and Section 5).

The final limitation of bootstrap values, shared with other support values based on resampling techniques, is that they are quite computationally expensive to compute: the budget required to compute B bootstrap trees is B-fold higher than the one required to compute the original tree. Clever implementations can substantially reduce that cost (Stamatakis, 2014) but it remains prohibitive for very large trees.



Original Data

A	*A*	*C*	T	T
B	*G*	*G*	A	T
C	*G*	*G*	C	C

Bootstrap Replicate #1

A	*A*	*C*	T	*C*
B	*G*	*G*	A	*G*
C	*G*	*G*	C	*G*

Bootstrap Replicate #2

A	*C*	*A*	T	*A*
B	*G*	*G*	A	*G*
C	*G*	*G*	C	*G*

Bootstrap Replicate #3

A	T	T	T	T
B	A	T	T	A
C	C	C	C	C

Figure 1: Principle of the bootstrap for phylogenies. Each character is identified by its color and style. Characters are sampled with replacement to produce bootstrap replicates, which are then used to infer phylogenies. The split $A|BC$ appears in 2 out of 3 bootstrap trees and therefore has a bootstrap value of $BP = 2/3$ or 66%.

3.1.2 Posterior Probabilities

Posterior Probabilities (PP) are mostly used in a Bayesian framework and similar in spirit bootstrap values. The main difference lies in the forest of trees used to compute support values. Bayesian procedures estimate the posterior distribution of trees. In practice, the distribution is too complex to fully explore and software produce a Markov Chain Markov Chains (MCMC) sample the posterior distribution (Yang and Rannala, 1997). The PP of a branch is computed, just like BP, as the the probability of occurrence of that branch in the MCMC sample. MCMC trees constitute a set of highly likely trees (best, second best, etc) for the original dataset. PP are easier to interpret than BP as they approximate directly the probability that a branch is present in the true tree, given the original data. Furthermore, since MCMC trees are a natural byproduct of the estimation procedure, there is almost no overhead in computing PP.

Unfortunately, PP are not immune to bias. Empirical studies found that PP are generally higher than BP (Anisimova et al., 2011) and sometimes even overconfident. The “star-tree paradox” (Yang, 2007) is the most extreme example of spurious support. Yang (2007) showed that when the actual tree is a 3 species star-like, with no real inner branch, and that sequence length goes to ∞, one branch randomly chosen among the 3 potential but erroneous inner branches, has its PP that goes to 100% whereas one would expect all potential branches to have PP around 33%.

Intuitively, PP are higher than BP because they cover fewer sources of variability. Unlike bootstrap trees, MCMC trees all originate from the same dataset. PP are quite good at capturing the lack of phylogenetic signal in the original MSA but not the impact of a few influential characters. For example, outlier characters with a strong effect on tree inference will affect all MCMC trees consistently. By contrast, they will be included in some replicates but left out from others leading to more variation among bootstrap trees than among MCMC trees. Finally, in genome-scale context where inaccuracies are more likely to arise from modeling errors than from sampling variability, PP are uniformly high and as uninformative as BP (Philippe et al., 2011a; Kumar et al., 2012)

3.1.3 Likelihood-based Support Values

Both BP and PP quantify the agreement between a focal tree and forest of trees. Likelihood-based supports are fast alternatives that bypass the forest and deal exclusively with the focal tree (Anisimova and Gascuel, 2006).

For any inner branch, there are NNI configurations around that branch: the focal one T_1 and two alternatives T_2, T_3 (see see Figure 2). If we note $\ell_i = \log Pr(D \| T_i)$ the likelihood of the data under tree i and assume that T_1 is the maximum-likelihood tree, we have $\ell_1 \geqslant \max(\ell_2, \ell_3)$. Likelihood-based supports values essentially test whether $\delta = \ell_1 - \max(\ell_2, \ell_3)$ is significantly larger than 0.

The most popular support values are:

- the approximate Likelihood Ratio Tests (aLRT) values which evaluates the statistics δ and compares it to $0.5\chi^2_0 + 0.5\chi^2_1$ to compute a p-value. The p-value is then converted into a support value between 1/8 and 1. A branch with high δ will have high support.

- the SH-corrected aLRT (SH-aLRT) values are based on the same idea but use the

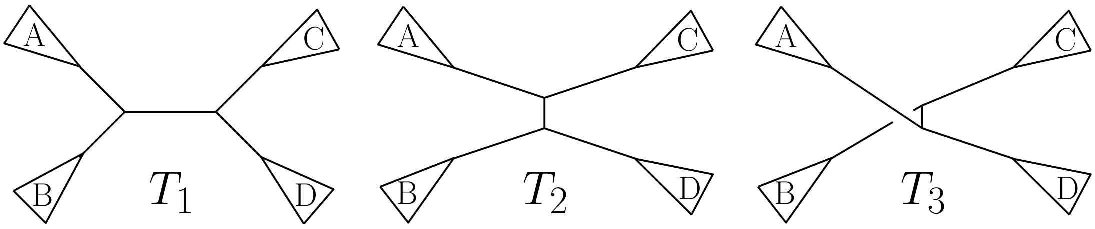


Figure 2: The maximum likelihood tree (T_1, left) and its two NNI-alternatives (T_2 middle and T_3 right) corresponding to different resolutions of the inner branch. Subtrees are sketched as triangles.

non-parametric Shimodaira and Hasegawa (1999) procedure to compute the p-value of δ.

- Finally approximate bayes (aBayes) is an approximation of the posterior probability of tree T_i computed as:

$$Pr(T_i|D) = \frac{Pr(T_i)Pr(D|T_i)}{\sum_{j=1}^{3} Pr(T_j)Pr(D|T_j)}$$

with a flat prior $Pr(T_1) = Pr(T_2) = Pr(T_3)$

All likelihood-based supports (aBayes, aLRT, SH-aLRT) amount to testing if T_1 is significantly better than T_2 and T_3. By focusing on one branch at the time rather than questioning the whole tree, likelihood-based supports are less conservative than BP and PP. They can also reuse likelihood computed while estimating the focal tree and are therefore much faster to compute than standard BP. Finally, they proved to be accurate in simulations studies (Anisimova et al., 2011). They are the default support values in PhyML (Guindon and Gascuel, 2003).

3.2 Outliers in the Data

The aforementioned support values aggregate all variations in the data set and are unable to distinguish between genuine and spurious variations due to outliers. The nature of resampling techniques is to use the empirical distribution as a surrogate for the true distribution. However, the empirical distribution may be polluted by outliers, defined here as "entry in the data set that are anomalous with respect to the behavior seen in the majority of the other entries in the data set" (Barnett and Lewis, 1994). This is a common occurrence in multi-locus studies where some characters can evolve according to one a tree, and others according to another tree Degnan and Rosenberg (2009). In that case, a single phylogeny is not a good fit to all the characters and Swofford et al. (1996) argued that it should be interesting to pinpoint where the phylogeny is not a good fit of the molecular data. Restricting the analyses to congruent characters usually leads to higher support values (Bar-Hen et al., 2008).

Many diagnostic approaches have been developed to identify outlier characters. Many studies (Rodríguez-Ezpeleta et al., 2007; Burleigh and Mathews, 2004) advocate removing fast-evolving characters which are a well-known cause of misleading phylogenetic signal and long branch attraction (LBA) where distantly related taxa are grouped together in the tree due to parallel or convergent evolution. Lopez et al. (1999) also suggest to investigate and remove characters with high rate variations (*i.e.* fast-evolving in some parts of the tree, slow-evolving in others). However both methods rely on good topologies to estimate rates, leading to a circularity problem.

Bar-Hen et al. (2008) adapted instead influence functions (Hampel, 1974) to phylogenetics in order to assess the impact of a single site on the likelihood. The main idea consists in removing one character at a time, to create *jackknife* replicates, and to infer a tree on each replicate. Jackknife trees are used to find influential characters, whose removal most affect the tree likelihood. Bar-Hen et al. (2008) report that influential sites have a strong impact on the topology and correspond mostly to fast evolving sites. All approaches found that removing outliers leads to more stable phylogenies but none is available as a routine in popular softwares.

3.3 Taxon Sampling

In phylogenomics studies, it is common to have conflicting trees with support values higher than 95% for all inner branches (Rydin and Källersjö, 2002). This correspond to setups where the estimated tree has a very small variance and differences between trees result from bias and modeling errors. In particular, Swofford et al. (1996) argues that adequate taxon sampling is one of the primary factors for accurate phylogenetic estimates, on par with enough sequence data. For example, dense taxon sampling can reduce the impact of LBA by splitting long branches (Felsenstein, 1978). Similarly, Holland et al. (2003) and Shavit et al. (2007) showed that the inclusion of an outgroup to the analysis may disrupt the ingroup phylogeny. When there are only a few taxa, but many characters, phylogenetic analysis can produce high support values (BP, PP, etc.) for incorrect or misleading phylogenies (Rokas et al., 2003a; Rokas and Carroll, 2005; Heath et al., 2008).

Analysis of sensitivity to taxon inclusion should be a part of careful and thorough phylogenetic analysis (Heath et al., 2008). Mariadassou et al. (2012) defined a the Taxon Influence Index (TII) to assess the influence of each on the phylogeny. Using any inference method, we define T^* to be the tree inferred from the complete MSA. Let T_k be a smaller tree, inferred from the alignment deprived of taxon k and T_k^* the tree obtained by pruning taxon k from T^*. The TII is the distance between trees T_k and T_k^*, such that

$$TII(k) = d(T_k, T_k^*)$$

They found that most taxa have small TII(k) and little influence on the topology whereas a few are highly influential *rogue taxa* and alter the phylogeny in clades even loosely related to their placement in the tree. Aberer et al. (2013) use a different approach, they start from a forest of trees (*e.g.* bootstrap trees) and search for a small set of taxa whose pruning increases the agreement between trees in the forest. The method is implemented in the webservice RogueNarok. Both methods find that pruning rogue taxa improves accuracy and results in more stable phylogenies with higher support values.

4 Consensus Methods

Boostrap, jackknife and bayesian estimation naturally produce a forest of trees with the same species set. But different trees can also be estimated using different methods or

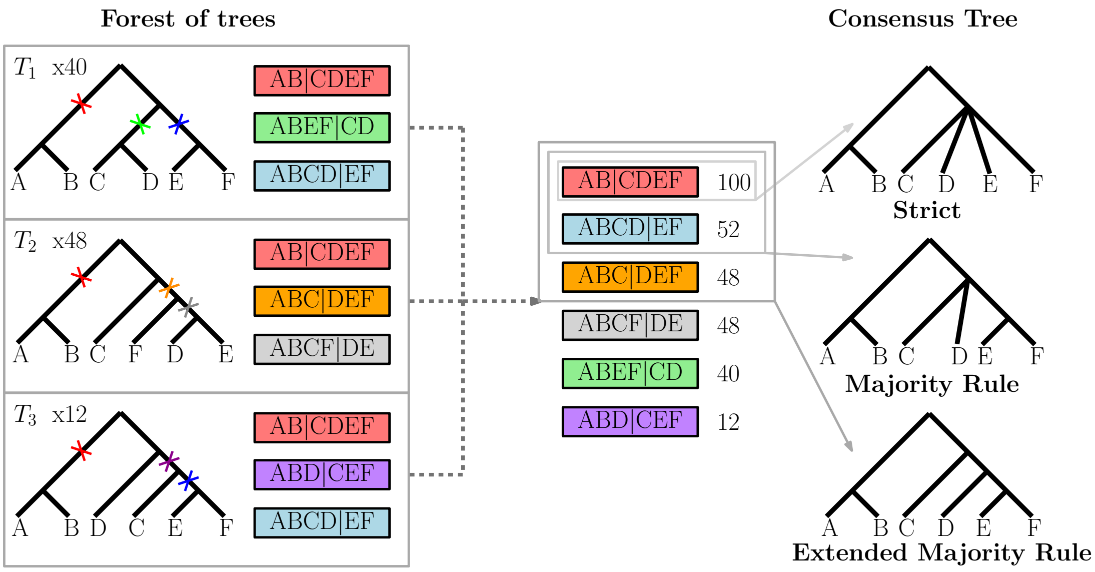


Figure 3: Left: a forest of 100 trees, corresponding to 3 topologies. Middle: Set of all bipartitions, with their frequencies found in the forest. Right: Different consensus made up by increasing large sets of partitions.

different sources of data. One way to summarize the forest is to *project* it on a focal tree to compute support values (see section 3). Alternatively, one can bypass the focal tree and combine all trees in the forest to get a single tree. That is the purpose of consensus trees methods.

4.1 Consensus Trees

Consensus trees are trees that summarize a forest of trees with the same species set. We present here only the *strict* consensus, the *majority rule* consensus and the *extended majority rule* consensus but there are many other consensus (see Bryant (2003) for an extensive survey). The different notions are best understood on an example. Consider the forest featured in Figure 3 with 40 copies of trees T_1, 48 of tree T_2 and 12 of tree T_3. Each tree is completely defined by the bipartitions it induces[1]. For example, T_1 induces the partitions $AB|CDEF$, $ABCD|EF$ and $ABEF|CD$[2], in addition to all trivial partitions $A|BCDEF$, $B|ACDEF$, etc not shown in the figure. Our three consensus methods scan the forest to build a list of all partitions occuring in the forest with their frequency of occurrence (middle column of Figure 3). They then select a subset of partitions according to some rules and build a consensus tree from that subset only.

4.1.1 Strict Consensus

The *strict consensus* tree (Rohlf, 1982) only uses partitions that appear in **all** trees, *i.e.* with a 100% occurence frequency. The strict consensus is fully compatible with all trees in the forest. However, it is less resolved than any tree and usually too strict. In our example, T_1 and T_2 and T_3 only differ in the position of D: if we removed D from all trees,

[1] or clades for for rooted trees

[2] or clades AB, CD, EF and $CDEF$ if considered as rooted

their topologies would be identical. We could therefore expect a branch separating EF from the ABC. However, the set $CDEF$ is completely unresolved in the strict consensus.

4.1.2 Majority-rule Consensus

The *majority-rule* consensus tree (Margush and McMorris, 1981) relaxes the condition that a bipartition must appear in all trees to be included in the consensus. Instead, it must only appear in *most* trees, *i.e.* have a occurence frequency higher than 50%. Although not obvious, all such partitions are pairwise-compatible and can be used to build a proper tree (Buneman, 1971). The majority-rule tree is more resolved than the strict consensus one but is not compatible with all trees in the forest. In our example, the partition $ABCD|EF$ seen in the majority-rule consensus is in conflict with the partition $ABCF|DE$ present in T_2.

4.1.3 Extended Majority-rule Consensus

The *extended majority-rule* consensus (Felsenstein, 2005), also called greedy consensus, relaxes the occurence frequency condition even further. The consensus is build by sequentially adding one partition at a time, in decreasing order of occurence and only if compatible with previously included partitions, until the tree is fully resolved or no more partitions can be added. Since all partitions with frequency higher than 50% are compatible, they must included in the consensus and the greedy consensus is thus a refinement of the majority-rule consensus. In our example, after including partitions $ABCD|EF$ and $AB|CDEF$, we can add either $ABC|DEF$ or $ABCF|DE$. The latter is in conflict $ABCD|EF$ and thus not included whereas the former is compatible with both partitions and thus included. After addition of $ABC|DEF$, the greedy consensus is fully resolved. Note that the greedy consensus is different from all trees in the forest.

4.1.4 Branch Lengths in Consensus Trees

The strict, majority-rule and extended majority-rule consensus trees only use tree toplogies and produce consensus topologies. One way to add branch lengths to that consensus is to take, for each branch, the average branch length over trees where that branch is present. This is the approach used in MrBayes (Ronquist and Huelsenbeck, 2003) when building a consensus tree from posterior trees.

4.2 Distance-based Consensus

The previous consensus methods are easy to understand and implement but can appear ad-hoc. They however have some theoretical grounding. Barthélemy and McMorris (1986) showed that the majority-rule consensus is the *center* of the forest in the sense that it minimizes the sum of Robinson-Foulds distances (Robinson and Foulds, 1979) to all trees in the forest and is thus the forest *median tree*. Alternatively, one could mininmize total squared distance to all trees in the forest to find the *mean tree*.

The Robinson-Foulds is but one of many distances between trees (see St. John (2017) for an excellent review) and one can define consensus similarly to the majority-rule consensus as the mean or median of the forest for some distance. Although seducing, this approach suffers from two shortcomings that severely limit its use. First, it is not obvious that the mean, or median, tree is well-defined or unique for some distances (Billera et al., 2001). Second, even when the mean is unique, computing it can be quite difficult and approximations must be used in pratice (Miller et al., 2015a). Computing the mean, variance and more general analyses of forest of trees based on tree distances is an active research field with many potential applications.

5 Extensions

The geometric model of Billera et al. (2001) allows one to compare phylogenetic trees, with the same leaf set of cardinality m, in a quantitative way. This space has a natural metric, giving a way of measuring distance between phylogenetic trees and providing some procedures for averaging or combining several trees whose leaves are identical. This geometry also shows which trees appear within a fixed distance of a given tree and enables construction of convex hulls of a set of trees. It also provides a justification for disregarding portions of a collection of trees that agree, thus simplifying the space in which comparisons are to be made.

5.1 Tree Space definition

. The distance $d(T_i, T_j)$ between two trees T_i and T_j account for differences with respect to both their tree topologies (branching structure) and branch lengths. The space is constructed by representing each of the $(2m-3)!!$ possible tree topologies by a single non-negative Euclidean orthant of dimension $m-3$ (the largest possible number of internal branches). The orthants are then “glued together” along appropriate axes. Specifically, nearest neighbor interchange (NNI) topologies lie in adjacent non-negative orthants along the boundary corresponding to the collapse of the relevant NNI edge.

For two trees with different topologies, the BHV distance is the length of the shortest path between them that remains in the treespace. The length of any path can be computed by calculating the Euclidean distance of the path restricted to each orthant that it passes though, and summing these lengths. The shortest path is called a geodesic, and will pass from one orthant to the next orthant through lower-dimensional boundaries corresponding to trees with fewer splits. Since the space is nonpositively curved, the geodesics are unique.

In Euclidean space, the Fréchet mean is the point minimizing the sum of the squared distances to the sample points, and is equivalent to the coordinate-wise average of the sample points. The mean tree is not necessarily a refinement of the majority-rule consensus tree. Fréchet variance is the tree that minimize the sum of squared distances. This variance is unique because treespace is non-positively curved. The Fréchet variance of a set of trees quantifies how spread out a set of trees is from their mean. See Miller et al. (2015b); Brown and Owen (2017) for details.

5.2 Use of BHV distance

Barden and Le (2017) proved a Central Limit Theorem on the BHV treespace, showing that the distribution of the sample means converges to a certain Gaussian distribution. It is useful for detecting splits of weak and strong support and in tree-valued hypothesis testing.

A key tool of Barden et al. (2014)is the log map that permits to map trees from their metric space to Euclidean space, where it is possible to model a tree estimate $\hat{T}$ as a noisy realization of the true tree T. Once the model parameters are estimated, Euclidean multivariate analysis techniques to reduce the dimension of the trees can be used. This allows to visualize tree estimates, along with their uncertainties.

For example it is possible to use the variance covariance matrix to estimate the principal directions of variability via principal components analysis. The axes of the $\mathbb{R}^m$ ellipsoid indicate the relative directions of precision, and the ellipsoid can be shrunk to be wholly contained in the same orthant as $\hat{T}_n$. This gives an unambiguous indication of the relative confidence in the edges of the estimated tree. Note that the procedure is unambiguous about the trees contained in the confidence set for a given confidence level α (see Willis (2016)).

Recently, de Vienne et al. (2012) developed a statistical non-parametric method to detect outlier trees from the set of gene trees. They first convert gene trees into vectors in a multidimensional Euclidean space and then apply multiple co-inertia analysis (MCOA)—an extension of principal coordinate analysis—directly to these vectorized gene trees. Their method, Phylo-MCOA, also detects outlier species, those whose position varies widely from tree to tree. Included in our results are simulation studies comparing our non-parametric method with Phylo-MCOA.

Weyenberg et al. (2014) proposes a non-parametric estimator of the distribution that generated the sample trees $T_1, \ldots, T_n$. This estimator can be viewed as a refined version of histogram-based estimation of a density. The kernel function, is a non-negative function defined on pairs of trees, which measures how similar two trees are. Kernel density estimation use the fact that points close to sample points tend to have higher likelihood than distant outlier points. The ultimate goal is to detect outlier trees, T_j, which are not actually drawn from the true distribution.

5.3 Need for a Strong Theoretical Framework

With the democratization of high-throughput sequencing, acquiring new genome data has ceased to be the limiting factor in phylogenetic, except maybe for organisms that are difficult to sample from the environment. In contrast, phylogeneticists are now faced with a rise in data errors, which stems from a flood of increasingly bogus genomic data that has become intractable by hand. Given the size of current and upcoming phylogenomic datasets (thousands of genes for hundreds — and soon thousands — of species), computational requirements are emerging as the new limiting factor. As multiple possible improvements are theoretically able to limit systematic error (and artefacts) during phylogenetic inference, we need to determine the set of properties that the ideal model of sequence of evolution should combine. An alternative approach is to use robustness methods to downweight or discard portions of the data where model misspecification occurs potentially mitigating the estimation biases these data may incur. This latter, in combination with improved models, may ultimately allow the development of robust and efficient analytical tools that are also computationally tractable.

References

Andre J. Aberer, Denis Krompass, and Alexandros Stamatakis. Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, 62(1):162–166, 2013. doi: 10.1093/sysbio/sys078. URL `+http://dx.doi.org/10.1093/sysbio/sys078`.

Maria Anisimova and Olivier Gascuel. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*, 55(4):539–552, Aug 2006. doi: 10.1080/10635150600755453. URL `http://dx.doi.org/10.1080/10635150600755453`.

Maria Anisimova, Manuel Gil, Jean-François Dufayard, Christophe Dessimoz, and Olivier Gascuel. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol*, 60(5):685–699, Oct 2011. doi: 10.1093/sysbio/syr041. URL `http://dx.doi.org/10.1093/sysbio/syr041`.

Avner Bar-Hen, Mahendra Mariadassou, Marie-Anne Poursat, and Philippe Vandenkoornhuyse. Influence function for robust phylogenetic reconstructions. *Mol Biol Evol*, 25(5):869–873, May 2008. doi: 10.1093/molbev/msn030. URL `http://dx.doi.org/10.1093/molbev/msn030`.

Dennis Barden and Huiling Le. The logarithm map, its limits and frechet means in orthant spaces. *arXiv preprint arXiv:1703.07081*, 2017.

Dennis Barden, Huiling Le, and Megan Owen. Limiting behaviour of fréchet means in the space of phylogenetic trees. *Annals of the Institute of Statistical Mathematics*, pages 1–31, 2014.

V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1994.

Jean-Pierre Barthélemy and F. R. McMorris. The median procedure for n-trees. *Journal of Classification*, 3(2):329–334, Sep 1986. ISSN 1432-1343. doi: 10.1007/BF01894194. URL `https://doi.org/10.1007/BF01894194`.

Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Math.*, 27:733–767, 2001. URL `http://www.math.cornell.edu/~billera/papers/treespace.pdf`.

Magnus Bordewich, Allen G Rodrigo, and Charles Semple. Selecting taxa to save or sequence: desirable criteria and a greedy solution. *Syst Biol*, 57(6):825–834, Dec 2008. doi: 10.1080/10635150802552831. URL `http://dx.doi.org/10.1080/10635150802552831`.

Daniel G Brown and Megan Owen. Mean and variance of phylogenetic trees. *arXiv preprint arXiv:1708.00294*, 2017.

D. Bryant. A classification of consensus methods for phylogenetics. *Bioconsensus*, 61:163–183, 2003.

Peter Buneman. *Mathematics the the Archeological and Historical Sciences*, chapter The Recovery of Trees from Measures of Dissimilarity, pages 387–395. Edinburgh University Press, 1971. ISBN 9780852242131.

J. Gordon Burleigh and Sarah Mathews. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany*, 91(10):1599–1613, 2004. doi: 10.3732/ajb.91.10.1599. URL `http://www.amjbot.org/content/91/10/1599.abstract`.

Damien M de Vienne, Sébastien Ollier, and Gabriela Aguileta. Phylo-mcoa: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Molecular biology and evolution*, 29(6):1587–1598, 2012.

James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*, 24(6):332–340, Jun 2009. doi: 10.1016/j.tree.2009.01.009. URL `http://dx.doi.org/10.1016/j.tree.2009.01.009`.

B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*, 93(14):7085–7090, Jul 1996. URL `http://www.pnas.org/cgi/content/full/93/23/13429`.

J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3):163–167, Mar 1998.

J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978. URL `http://www.molecularevolution.org/si/resources/references/files/Felsenstein_1978.pdf`.

J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791, July 1985. doi: 10.2307/2408678. URL `http://links.jstor.org/sici?sici=0014-3820(198507)39:4%3C783:CLOPAA%3E2.0.CO;2-L`.

J. Felsenstein. Phylip (phylogeny inference package) version 3.6. Distributed by the author, 2005.

J. Felsenstein and H. Kishino. Is there something wrong with the bootstrap on phylogenies? a reply to hillis and bull. *Systematic Biology*, 42(2):193–200, June 1993. doi: 10.2307/2992541. URL `http://links.jstor.org/sici?sici=1063-5157(199306)42%3A2%3C193%3AITSWWT%3E2.0.CO%3B2-Y`.

Joseph Felsenstein. *Inferring Phylogenies.* Sinauer Associates, September 2004.

Joseph Felsenstein and Gary A Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, 13(1):93–104, 1996.

Peter G Foster. Modeling compositional heterogeneity. *Systematic Biology*, 53(3):485–495, 2004.

O. Gascuel. *Mathematics of evolution and phylogeny.* Oxford University Press, 2005.

Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736, 1994.

Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, Oct 2003. doi: 10.1080/10635150390235520. URL http://www.informaworld.com/smpp/ftinterface~content=a713850337~fulltext=713240930.

Aaron L Halpern and William J Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15 (7):910–917, 1998.

F. R. Hampel. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393, 1974.

Tracy A Heath, Shannon M Hedtke, and David M Hillis. Taxon sampling and the accuracy of phylogenetic analyses. *J Mol Evol*, 46:239–257, 2008. URL http://www.plantsystematics.com/qikan/epaper/hb_zhaiyao.asp.

David M. Hillis and James J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2):182–192, June 1993. doi: 10.2307/2992540.

B. R. Holland, D. Penny, and M. D. Hendy. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock–a simulation study. *Syst Biol*, 52(2):229–238, Apr 2003. doi: 10.1080/10635150390192771. URL http://www.informaworld.com/smpp/content~content=a713850188~db=all~order=page.

Lars S Jermiin, Simon YW Ho, Faisal Ababneh, John Robinson, and Anthony WD Larkum. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, 53(4):638–643, 2004.

Sudhir Kumar, Alan J Filipski, Fabia U Battistuzzi, Sergei L Kosakovsky Pond, and Koichiro Tamura. Statistics and truth in phylogenomics. *Mol Biol Evol*, 29(2):457–472, Feb 2012. doi: 10.1093/molbev/msr202. URL http://dx.doi.org/10.1093/molbev/msr202.

Nicolas Lartillot and Hervé Philippe. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21 (6):1095–1109, 2004.

Simon Laurin-Lemay, Henner Brinkmann, and Hervé Philippe. Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, 22 (15):R593–R594, 2012.

P. Lopez, P. Forterre, and H. Philippe. The root of the tree of life in the light of the covarion model. *J Mol Evol*, 49(4):496–508, Oct 1999. URL http://www.springerlink.com/content/hwle29fxv74ra6xy/.

T. Margush and F. R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, 43(2):239–244, Mar 1981. ISSN 1522-9602. doi: 10.1007/BF02459446. URL https://doi.org/10.1007/BF02459446.

Mahendra Mariadassou, Avner Bar-Hen, and Hirohisa Kishino. Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference. *Syst Biol*, 61(2):337–345, Mar 2012. doi: 10.1093/sysbio/syr129. URL `http://dx.doi.org/10.1093/sysbio/syr129`.

Ezra Miller, Megan Owen, and J. Scott Provan. Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, 68:51 – 91, 2015a. ISSN 0196-8858. doi: https://doi.org/10.1016/j.aam.2015.04.002. URL `http://www.sciencedirect.com/science/article/pii/S0196885815000470`.

Ezra Miller, Megan Owen, and J Scott Provan. Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, 68:51–91, 2015b.

David A Morrison and John T Ellis. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18s rdnas of apicomplexa. *Molecular Biology and Evolution*, 14(4):428–441, 1997.

T Heath Ogden and Michael S Rosenberg. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic biology*, 55(2):314–328, 2006.

Gustavo Parisi and Julián Echave. Structural constraints and emergence of sequence patterns in protein evolution. *Molecular Biology and Evolution*, 18(5):750–756, 2001.

Matthew W. Pennell and Luke J. Harmon. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann N Y Acad Sci*, 1289:90–105, Jun 2013. doi: 10.1111/nyas.12157. URL `http://dx.doi.org/10.1111/nyas.12157`.

Hervé Philippe, Henner Brinkmann, Dennis V Lavrov, D. Timothy J Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*, 9(3):e1000602, Mar 2011a. doi: 10.1371/journal.pbio.1000602. URL `http://dx.doi.org/10.1371/journal.pbio.1000602`.

Hervé Philippe, Henner Brinkmann, Dennis V Lavrov, D Timothy J Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS biology*, 9(3):e1000602, 2011b.

Hervé Philippe, Damien M De Vienne, Vincent Ranwez, Béatrice Roure, Denis Baurain, and Frédéric Delsuc. Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283:1–25, 2017.

Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, 2010. doi: 10.1371/journal.pone.0009490. URL `http://dx.doi.org/10.1371/journal.pone.0009490`.

Bruce Rannala. Identifiability of parameters in mcmc bayesian inference of phylogeny. *Systematic Biology*, 51(5):754–760, 2002.

Liam J Revell, Luke J Harmon, and David C Collar. Phylogenetic signal, evolutionary process, and rate. *Syst Biol*, 57(4):591–601, Aug 2008. doi: 10.1080/10635150802302427. URL `http://dx.doi.org/10.1080/10635150802302427`.

D. F. Robinson and L. R. Foulds. *Lectures Note in mathematics*, volume 748, chapter Comparison of weighted labelled trees, pages 119–126. Springer-Verlag, Berlin, 1979.

Naiara Rodríguez-Ezpeleta, Henner Brinkmann, Béatrice Roure, Nicolas Lartillot, B. Franz Lang, Hervé Philippe, and Frank Anderson. Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, 56(3):389–399, 2007. doi: 10.1080/10635150701397643. URL `+http://dx.doi.org/10.1080/10635150701397643`.

F. James Rohlf. Consensus indices for comparing classifications. *Mathematical Biosciences*, 59(1):131 – 144, 1982. ISSN 0025-5564. doi: http://dx.doi.org/10.1016/0025-5564(82)90112-2. URL `http://www.sciencedirect.com/science/article/pii/0025556482901122`.

Antonis Rokas and Sean B Carroll. More genes or more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*, 22(5):1337–1344, May 2005. doi: 10.1093/molbev/msi121. URL `http://dx.doi.org/10.1093/molbev/msi121`.

Antonis Rokas, Barry L Williams, Nicole King, and Sean B Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003a. doi: 10.1038/nature02053. URL `http://dx.doi.org/10.1038/nature02053`.

Antonis Rokas, Barry L Williams, Nicole King, and Sean B Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798, 2003b.

Fredrik Ronquist and John P Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, Aug 2003. URL `http://bioinformatics.oxfordjournals.org/cgi/reprint/19/12/1572`.

Catarina Rydin and Mari Källersjö. Taxon sampling and seed plant phylogeny. *Cladistics*, 18(5):485 – 513, 2002. ISSN 0748-3007. doi: http://dx.doi.org/10.1016/S0748-3007(02)00104-4. URL `http://www.sciencedirect.com/science/article/pii/S0748300702001044`.

Liat Shavit, David Penny, Michael D Hendy, and Barbara R Holland. The problem of rooting rapid radiations. *Mol Biol Evol*, 24(11):2400–2411, Nov 2007. doi: 10.1093/molbev/msm178. URL `http://dx.doi.org/10.1093/molbev/msm178`.

H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol*, 16(8):1114–1116, 1999.

Katherine St. John. Review paper: The shape of phylogenetic treespace. *Systematic Biology*, 66(1):e83–e94, 2017. doi: 10.1093/sysbio/syw025. URL `+http://dx.doi.org/10.1093/sysbio/syw025`.

Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, Nov 2006. doi: 10.1093/bioinformatics/btl446. URL `http://dx.doi.org/10.1093/bioinformatics/btl446`.

Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014. doi: 10.1093/bioinformatics/btu033. URL `+http://dx.doi.org/10.1093/bioinformatics/btu033`.

Edward Susko. On the distributions of bootstrap support and posterior distributions for a star tree. *Syst Biol*, 57(4):602–612, Aug 2008. doi: 10.1080/10635150802302468. URL `http://dx.doi.org/10.1080/10635150802302468`.

Edward Susko. First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. *Mol Biol Evol*, Feb 2010. doi: 10.1093/molbev/msq048. URL `http://dx.doi.org/10.1093/molbev/msq048`.

Edward Susko, Yuji Inagaki, Chris Field, Michael E Holder, and Andrew J Roger. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Molecular biology and evolution*, 19(9):1514–1523, 2002.

David L Swofford, Gary J Olsen, Peter J Waddell, and David M Hillis. Phylogenetic inference. 1996.

Gerard Talavera and Jose Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4):564–577, 2007.

Grady Weyenberg, Peter M Huggins, Christopher L Schardl, Daniel K Howe, and Ruriko Yoshida. Kdetrees: non-parametric estimation of phylogenetic tree distributions. *Bioinformatics*, 30(16):2280–2287, 2014.

Amy Willis. Confidence sets for phylogenetic trees. *arXiv preprint arXiv:1607.08288*, 2016.

Karen M Wong, Marc A Suchard, and John P Huelsenbeck. Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476, 2008.

Z. Yang and B. Rannala. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Mol Biol Evol*, 14(7):717–724, Jul 1997. URL `http://mbe.oxfordjournals.org/cgi/reprint/14/7/717`.

Ziheng Yang. Fair-balance paradox, star-tree paradox, and bayesian phylogenetics. *Mol Biol Evol*, 24(8):1639–1655, Aug 2007. doi: 10.1093/molbev/msm081. URL `http://dx.doi.org/10.1093/molbev/msm081`.

A. Zharkikh and W. H. Li. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. i. four taxa with a molecular clock. *Mol Biol Evol*, 9(6):1119–1147, Nov 1992. URL `http://mbe.oxfordjournals.org/cgi/reprint/9/6/1119`.

Olga Zhaxybayeva, Pascal Lapierre, and J Peter Gogarten. Genome mosaicism and organismal lineages. *TRENDS in Genetics*, 20(5):254–260, 2004.