The Bayesian Phylogenetic Bootstrap, Application to Short Trees and Branches

Frédéric Lemoine[1,2]*, Olivier Gascuel[3]*

[1]*Institut Pasteur, Université Paris Cité, National Reference Center for Respiratory Viruses, 75015, Paris, France.*

[2]*Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, 75015, Paris, France.*

[3]*Institut de Systématique, Evolution, Biodiversité (ISYEB UMR7205 – CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), 75005 Paris, France.*

*Corresponding authors: frederic.lemoine@pasteur.fr, olivier.gascuel@mnhn.fr

Abstract: Felsenstein's bootstrap is the most commonly used method to measure branch support in phylogenetics. Current sequencing technologies can result in massive sampling of taxa (e.g. SARS-CoV-2). In this case, the sequences are very close, the trees are short, and the branches correspond to a small number of mutations (possibly 0). Nevertheless, these trees contain a strong signal, with unresolved parts but a low rate of false branches. With such data, Felsenstein's bootstrap is not satisfactory. Due to the frequentist nature of bootstrap sampling, the expected support of a branch corresponding to a single mutation is ~63%, even though it is highly likely to be correct. Here we propose a Bayesian version of the phylogenetic bootstrap in which sites are assigned uninformative prior probabilities. The branch support can then be interpreted as a posterior probability. We do not view the alignment as a small subsample of a large sample of sites, but rather as containing all available information (e.g., as with complete viral genomes, which are becoming routine). We give formulas for expected supports under the assumption of perfect phylogeny, in both the frequentist and Bayesian frameworks, where a branch corresponding to a single mutation now has an expected support of ~90%. Simulation results show that these theoretical results are robust for realistic data. Results with viral and non-viral low-homoplasy datasets show that Bayesian bootstrap supports are easier to interpret, with high supports for branches very likely to be correct. As homoplasy increases, the two supports become closer and strongly correlated.

Keywords: phylogenetic trees, branch supports, Felsenstein's bootstrap, perfect phylogeny, low homoplasy datasets, viruses (Ebola, SARS-CoV-2, Rift Valley Fever Virus).

Introduction

The phylogenetic bootstrap was proposed by Joseph Felsenstein to estimate the branch supports of trees inferred from multiple sequence alignments (Felsenstein 1985). Given a multiple sequence alignment (MSA) and a reference tree inferred from it, the principle is to sample with replacement the sites of the original alignment to obtain pseudo-alignments of the same length, infer pseudo-trees from the pseudo-alignments, and measure the support of each branch in the reference tree as the proportion of pseudo-trees containing that branch. The usefulness and simplicity of this method made it extremely popular. Felsenstein's 1985 article is among the top 100 most cited scientific papers of all time (Van Noorden et al. 2014).

The phylogenetic bootstrap is an application to MSA and trees, of the statistical bootstrap, which is commonly used to study the distribution of numerical estimates (Efron 1979; Efron and Tibshirani 1993). In this context, the sampling with replacement procedure has a frequentist basis and mimics the sampling of observations (or statistical units) in a large set of homogeneous and independent observations. The relevance and limits of these homogeneity and independence assumptions have been questioned on a biological basis (Sanderson 1995) when applied to the phylogenetic bootstrap, where MSA sites are considered as statistical units. The complexity of phylogenetic tree spaces and sequence evolution models (compared to simpler numerical spaces and estimates) makes statistical interpretation of bootstrap supports difficult (Holmes 2003). This issue was the subject of much debate in the 90s (Hillis and Bull 1993; Felsenstein and Kishino 1993; Berry and Gascuel 1996). Several methods (Zharkikh and Li 1995; Efron et al. 1996; Susko 2009, 2010) have been proposed to correct bootstrap supports to better agree with standard ideas of confidence levels and hypothesis testing. However, bootstrap correction methods are limited to relatively small data sets, and the original method is still highly used (~2,000 citations in 2023). As stated by Soltis and Soltis (2003): “consensus has been reached among practitioners, if not among statisticians and theoreticians”, and “many systematists have adopted Hillis and Bull’s “70%” value as an indication of support”. This 70% threshold indicates the very conservative nature of the standard bootstrap. The bootstrap is also computationally heavy, but it is easily parallelized and fast algorithms have been designed (Stamatakis et al. 2008; Minh et al. 2013)

Alternatives to the Felsenstein bootstrap are local supports (aLRT and aBayes; Anisimova et al. 2006, 2011; Guindon et al. 2010), which are very fast to compute, but can be too local and support branches that should not be supported from a global perspective (Gascuel and Lemoine 2024). Bayesian posterior probabilities are also widely used (Huelsenbeck and Rannala 2004), but they tend to be liberal (as opposed to bootstrap supports), are difficult to apply to large datasets, and are sensitive to model misspecification.

The Felsenstein bootstrap and most of the approaches mentioned above face the problem of "rogue" taxa (Sanderson and Shaffer 2002; Soltis and Soltis 2003), which are inevitable in large datasets with many taxa. These taxa are unstable within bootstrap trees, and the presence of a single rogue taxon in a dataset makes the bootstrap support very low when a branch is found almost identically in bootstrap trees, except for one or a few misplaced rogue taxa. To address this problem, we proposed the Transfer Bootstrap Expectation (TBE, Lemoine et al. 2018), in which we measure the degree of presence of reference branches in bootstrap trees using a quasi-continuous index (as opposed to the binary index

used in the standard bootstrap). The support of a reference branch corresponds to the average number of taxa that must be transferred (or deleted) to find that branch in the bootstrap trees. TBE supports are higher than Felsenstein bootstrap supports, but do not falsely support incorrect branches.

The Felsenstein bootstrap is also problematic with short trees and branches supported by only a few mutations (Wertheim et al. 2022). This type of data is becoming increasingly common with the massive sequencing of many species, especially viruses such as SARS-CoV-2, Ebola, Zika, Dengue and many others. However, viruses are not the only organisms concerned. Studies of closely related species and intraspecific analyses also produce this type of data and phylogeny, as well as single-cell data, which is increasingly used in cancer and developmental biology (Jahn et al. 2016). Under these conditions, we are close to perfect (or Hennigian) phylogeny assumptions (Felsenstein 1973; Gusfield 1997). The level of homoplasy is low or even zero at some sites, and the sites contradict each other little or not at all. Contradictions between sites (e.g., one site supports cherry {A, B} with a mutation separating taxa A and B from all other taxa, while another contradictory site supports cherry {A, C}) can be the result of random neutral mutations or, for example, caused by recombination and horizontal transfer. Under these near-perfect conditions, (Wertheim et al. 2022) show that phylogenies are poorly resolved, but that inferred branches have a high probability of being correct. However, theoretical analyses (Felsenstein 1985) show that the expected bootstrap support for a branch supported by a single mutation is 0.632 (≈63%). This value is well known in bootstrap theory (Efron and Tibshirani 1997), and corresponds to the probability that a given site (e.g. carrying a mutation supporting a given clade) will be present in a bootstrap sample randomly drawn with replacement. This is a clear shortcoming of the standard bootstrap when the data is near-perfect, and (Wertheim et al. 2022) recommends against bootstrap resampling when substitutions are sparse.

In fact, in the standard frequentist bootstrap, certain sites are removed (with probability 1 - 0.632 ≈ 37%) and part of the original information is lost, resulting in bootstrap samples that are less informative than the original sample. This becomes problematic with small samples and when the information is poor, for example with virus data, where a large number of sites are invariant. From a biological point of view, this resampling method can also be questioned, as it implicitly assumes that the original sample is randomly taken from a larger sample, typically a genome. However, in the case of viruses (e.g., SARS-CoV-2, Ebola …) the data generally corresponds to the entire genome, and we cannot think of it as a sample from a larger dataset. Sampling with replacement creates "pseudo-genomes" that are poorer in information than the original data, with only ~63% of the sites present in the complete genome. The situation is similar if we are studying the phylogeny of a particular gene, whose history often differs from that of the genomes and species from which it is extracted.

In this paper, we propose a Bayesian approach to phylogenetic bootstrapping in which all sites remain present in the bootstrap samples. This approach is derived from the Bayesian bootstrap proposed by Rubin (1981) and others (Shao and Tu 1995) in the context of numerical data and estimates. The principle is to weight the n observations of the original sample, rather than drawing them with replacement. Each observation i thus has a weight w_i drawn from a non-informative prior distribution that is identical and independent for all observations (or statistical units). Rubin suggests using a Dirichlet distribution, but other distributions have been studied (Lo 1991). Theoretical results show that, under certain conditions, the distribution obtained for the estimated parameters (e.g., the mean or

variance of a normal distribution) converges to the posterior of those parameters when their priors are non-informative (Lo 1991). The Bayesian bootstrap is a relevant alternative to the frequentist bootstrap, which is smoother and generally more suitable for small samples (Shao and Tu 1995). In the following, we introduce and study the application of this approach to phylogenetics. We specifically analyze zero-length branches, which may be significantly supported by most software programs, even if they are not supported by any mutation. We provide the expected values for standard frequentist and Bayesian bootstrap supports under the assumption of perfect phylogeny. We compare the results of the two approaches on simulated data that gradually deviate from perfect phylogeny. We analyze real datasets of viruses (Ebola, SARS-CoV-2, Rift Valley Fever Virus) as well as a plant dataset (Psiguria genus, Cucubirtacea) with low homoplasy. The results show that for this type of data with low homoplasy, Bayesian bootstrap supports are easier to interpret, with high supports for branches that are essentially correct. As homoplasy increases, the two supports become closer and strongly correlated.

Methods and theoretical analyses

The data is a multiple sequence alignment (MSA; DNA or protein) containing n sites and s sequences. The goal is to construct a phylogenetic tree from this MSA and measure the support of its branches. To simplify the presentation, we will mainly use the maximum likelihood framework, but the methods presented can easily be adapted to other approaches, in particular distance-based methods.

The length of branches in a phylogeny is expressed as the expected number of mutations per site. If a branch has a length of zero, this essentially means that it is not supported by any of the mutations present in the input MSA. In this case, the branch's support should be zero (or non-significant) whatever the support measure considered. However, when using the standard phylogenetic bootstrap, it is not uncommon for branches of zero length to receive substantial support, which can confound analyses (Guindon et al. 2010; Gascuel and Lemoine 2024). In what follows, we will explain this phenomenon, illustrate it with examples and show how it can be remedied. We will then examine the support of branches carrying one or more mutations (i.e. of non-zero length), first in the framework of the standard frequentist phylogenetic bootstrap, then in the framework of the Bayesian bootstrap.

Bootstrap support of zero length branches

Here we use a simple example as shown in Figure 1(a). In this example, the tree contains the cherry (2-taxon clade) {A, B}, which is contained in the clade {A, B, C}. If there is one or more mutations on the branch *ab* (located at the root of the {A, B} cherry, the same notation is used for the other branches), we expect this branch to be supported, and the {A, C} and {B, C} cherries to have no support. Conversely, if there is no mutation on *ab*, which is therefore of length zero, none of the three resolutions should have significant support.

However, this is not the case with standard bootstrap and usual phylogeny software, which produce binary topologies, mainly because they resolve the tree in a more or less deterministic way in the case of polytomies and null branches. For example, RaxML-NG (Kozlov et al. 2019) and PhyML (Guindon et al. 2010) produce caterpillars whose leaf order depends on the sequence order in the MSA.

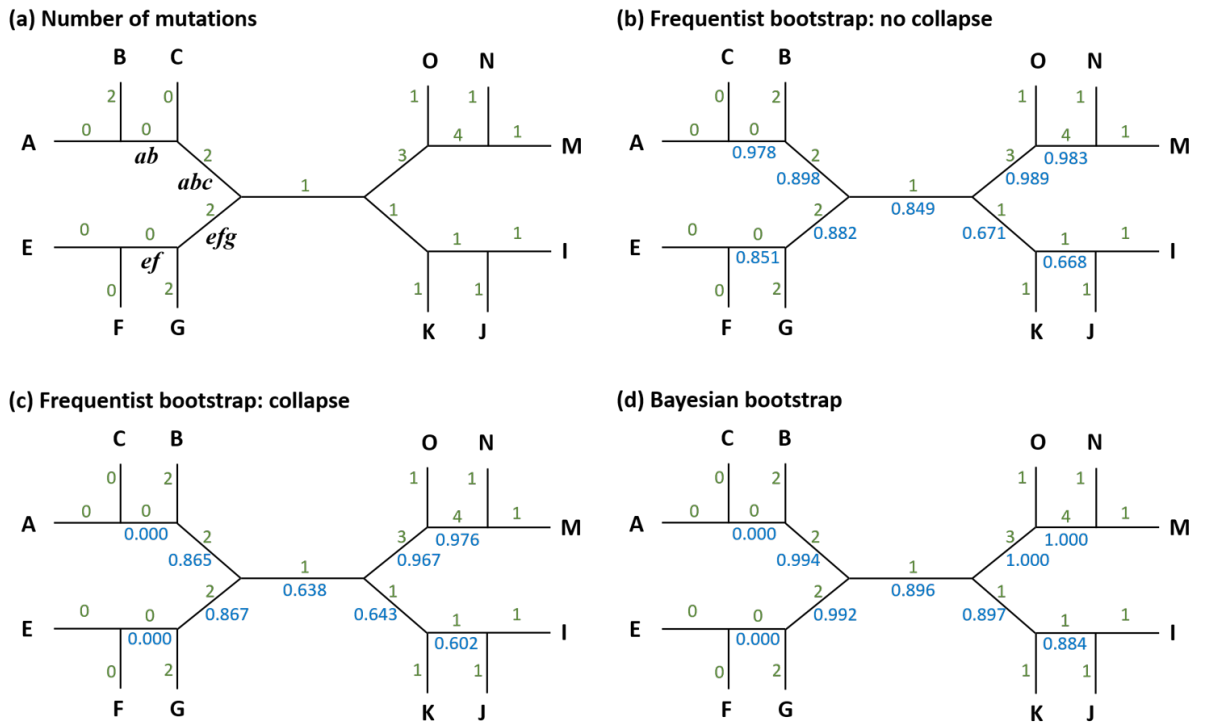

Figure 1 - Branch supports on a simple example. (a) Model tree with the number of mutations present in the MSA; this contains 1000 sites and each mutation belongs to a different site, the phylogeny is "perfect". **(b)** Application of the standard bootstrap with PhyML; the estimated branch lengths (in green, rounded to number of mutations as in (c) and (d)) are very close to the numbers of mutations present in the MSA, but taxa A and C are grouped together; the two zero-length branches are highly supported; results are very similar with RaxML-NG and IQTREE. **(c)** Supports obtained with the same bootstrap trees as in (a), but collapsing branches whose length is less than 1/10,000 (i.e., <0.1 expected mutations over all 1000 sites); the two zero-length branches now have zero support, and the branches supported by 1, 2, 3, 4 mutations have supports close to the theoretical expectations (e.g., 0.632 with 1 mutation). **(d)** Bayesian supports obtained with PhyML and our BBOOT pipeline; the values obtained are also close to the theoretical expectations (e.g. 0.904 with 1 mutation). See text for details.

If there are no mutations on branches *a*, *c*, and *ab*, sequences A and C are identical. If there are mutations on branches *b* and *abc* (Fig. 1(a)), A and C will generally form a cherry in the inferred reference tree (false, with a root branch of zero length) and B will be isolated. If the number of mutations on *b* and *abc* is relatively high, the bootstrap trees will be analogous to the reference tree and will strongly support the false cherry {A, C}, for which there is in fact no signal. The bootstrap associated with false resolutions originating from null branches may therefore strongly support false clades.

The standard bootstrap can also strongly support correct clades for which there is no signal. This case is illustrated by taxa E, F and G in Figure 1(a). Here, there are no mutations on the *e*, *f* and *ef* branches, but there are mutations on the *g* and *efg* branches. In this case, the cherry {E, F} will frequently be found in bootstrap trees, even though it should have no support.

Figure 1(b) shows the supports obtained with an alignment corresponding to the two cases above and the number of mutations indicated in Figure 1(a). The supports for the {A, C} and {E, F} clades are very high, while there is no signal for these clades and little or no support is expected. The results (Fig. 1(b)) are from PhyML, but the exact same effect is obtained with RaxML-NG and IQTREE (Minh et al.

2020). The reason for this is the hidden determinism of these programs, which is easily explained here, since sequences A and C are identical, as are sequences E and F. By default, these programs delete duplicated sequences, build the tree, and reinsert duplicates to form monophyletic clades. But even when this option is turned off in PhyML, we still find non-negligible support, on the order of 50% or more, for clades that should not be supported due to the absence of mutation. Other sources of determinism exist in all software, notably related to sequence order (Gascuel and Lemoine 2024), which can lead to supports close to 1 on zero-length branches (cf. results obtained on Ebola and SARS-CoV-2 data, see below).

The solution we propose here is to collapse zero-length bootstrap branches that are not supported by any mutation. This avoids supporting branches that should not be supported. In practice, we collapse bootstrap branches whose estimated length is less than χ/n, where n is the length of the MSA and χ corresponds to a "number" of mutations ($\ll 1$; $\chi = 0.1$ in our experiments). This approach is implemented in Figure 1(c), using the data set of Figure 1(a) and the same bootstrap trees as in Figure 1(b), but with collapsed zero-length branches; the {A, C} and {E, F} clades now have zero support (as opposed to Fig. 1(b)), expressing the fact that the {A, B, C} and {E, F, G} clades are polytomies. Another beneficial effect of the collapse of the null bootstrap branches is that the bootstrap supports generated by different software programs become virtually identical, whereas without this constraint they show significant differences due to different hidden determinisms (see Supp. Fig. S1, with Ebola sequences analyzed by PhyML and RaxML-NG). Note that the goal here is not to use the bootstrap to detect those branches that should not be supported (for that we can use local tests such as aLRT, or simply look at the branch length expressed in number of mutations), but to correct the standard bootstrap for the sake of consistency.

Expected supports with the standard frequentist bootstrap

Here we analyze the expected supports with the standard frequentist bootstrap, for branches of the reference tree whose length corresponds to at least 1 mutation. In the theoretical analysis that follows, we assume that the sequence data are close to the perfect phylogeny hypothesis (Felsenstein 1973; Gusfield 1997). This implies that no or few convergent, parallel, and revertant mutations are observed at a single site. It also implies that the different sites in the alignment contain no or few conflicting signals. The hypothesis of a strictly perfect phylogeny is rarely verified, but many data sets come close, especially viral data and those with a low number of mutations and closely related sequences. This is the case, for example, with the Ebola and SARS-Cov-2 datasets, which we shall study in the next section. With such data, the estimated branch lengths clearly represent mutation numbers (Fig. 2).

The perfect phylogeny framework was used by Felsenstein (1985) to study the relationship between the length (in number of mutations) and the bootstrap support of tree branches. We follow the same path here, starting by analyzing the case where a branch of the reference tree is supported by 1 mutation, and thus by a single site within the MSA given as input. In a bootstrap sample, the probability of this site being absent is equal to $(1-1/n)^n$; in this case, the corresponding branch in the bootstrap tree will not be inferred or will have an estimated length very close to zero. Switching to the logarithm and assuming n is large, we can easily show that the limiting value of $(1-1/n)^n$ is e^{-1}. The probability that the site is present at least 1 time and that a bootstrap tree contains and supports this branch is therefore equal to $1-e^{-1} \approx 0.632$. This value is well known in bootstrap theory (Efron and Tibshirani 1997). In

phylogenetics, assuming (nearly) perfect data, 0.632 represents the expected bootstrap support for a branch whose length corresponds to (approximately) 1 mutation. In the case where a branch of the original tree has a length corresponding to k ($\ll n$) mutations, the above values become $\left(1-k/n\right)^n \approx e^{-k}$ and the expected bootstrap support becomes $1-e^{-k}$, which is approximately 0.865, 0.950, and 0.982 for k = 2, 3, and 4, respectively.

The results in Figure 1(c) (PhyML trees with collapsed zero-length bootstrap branches) show that these theoretical values are very close to the values observed with our "perfect" dataset (Fig. 1(a)): for 1 mutation the supports are equal to 0.638, 0.602 and 0.634 (expected value = 0.632); for 2 mutations: 0.865 and 0.867 (expected value = 0.865); for 3 mutations: 0.967 (expected value = 0.950); for 4 mutations: 0.976 (expected value = 0.982). In the following, we analyze simulated and real data with a relatively low number of mutations, but deviating from the assumption of a perfect phylogeny. On these realistic/real data, the observed supports are generally very close to the theoretical ones. The simulated data also show that branches with a length of about 1 mutation, and thus with a support close to 0.632, are generally true branches belonging to the model tree used to generate the sequences (Wertheim et al. 2022). For this type of data, 0.632 is therefore a high support, albeit a low value, below the traditional 70% or 80% threshold that is generally recommended to select the well-supported branches (Hillis and Bull 1993; Soltis and Soltis 2003). On these datasets, a ("high") support of 0.632 may seem very arbitrary to users and in no way represents a probability of being correct or the confidence level of a statistical test. This brings us back to the conservatism of the frequentist bootstrap and the difficulty of interpreting its supports, which have been discussed many times (see the Introduction).

The Bayesian phylogenetic bootstrap, expected supports

In the Bayesian bootstrap, sites in the bootstrap samples are drawn with weights from an a priori non-informative distribution, and the resulting estimates represent (under certain assumptions) the posterior distribution of the parameters of interest (Rubin 1981; Lo 1991). Several non-informative distributions have been proposed, the most widely used is the Dirichlet distribution of parameters (n; 1, 1, ...1). Rubin (1981) proposes a definition of this distribution and a computation based on a set of intervals in [0, 1]. An equivalent definition is to independently draw n weights w_i (attached to each of the sites i) in an exponential distribution with expectation 1.0 and normalize these weights by their sum. To make it closer to the usual phylogenetic framework, we multiply these normalized weights by the number of sites, i.e., $w_i \leftarrow w_i \times n / \left(w_1 + w_2 + ... w_n\right)$. The sum of the normalized weights is n, and the expectation of each weight is 1.0. This distribution of weights thus represents a continuous version of the frequentist sampling with replacement procedure. The main difference is that each site is present with a non-zero weight in each of the Bayesian bootstrap samples. Thus, a signal carried by a single site remains present in all samples, but with a different weight from one sample to another. Within the maximum likelihood framework, the likelihood of a weighted MSA is calculated simply by multiplying the log-likelihood of each site by its weight. Standard software programs (e.g. RaxML-NG) already use this weighting system with integer weights to account for identical sites. In PhyML, which we use in our experiments, it is possible to have continuous weights, as required here.

To derive the theoretical support as a function of the branch length of the original tree, we again place ourselves in the perfect phylogeny framework. In principle, in this framework and with a perfect inference method, we should find the original phylogeny regardless of the bootstrap sample and

Bayesian weights, since each site is present with a positive weight and there is no conflicting signal. If a branch *b* in the reference tree is supported by 1 mutation (i.e., has length $1/n$), and if the weight of the corresponding site in a bootstrap sample is *w*, then the length of *b* in the corresponding bootstrap tree will be w/n. If *w* is small, the length of the bootstrap branch will be very small.

In practice, short branches must be collapsed to avoid the effects of the hidden determinisms described above, numerical approximations, and also because real data are never perfect, with a certain amount of homoplasy and contradiction, which can induce branches of non-zero (but almost) length that are not supported by any mutation. In our experiments, we collapsed branches of length less than $0.1/n$ (i.e., with an expected number of mutations less than $\chi = 1/10$). Consider a reference branch *b* of length $1/n$, if the weight *w* of the single site supporting this branch is less than $\chi = 1/10$, this branch will be inferred but collapsed and therefore not counted as recovered in the corresponding bootstrap tree. The theoretical support of the branch will therefore be equal to $(1-p)$, where *p* is the probability of having a weight *w* less than χ. For relatively large *n* (say >100), the normalized Dirichlet (see formula above) is practically equal to an exponential law with parameter 1.0, and the distribution in the neighborhood of 0 is asymptotically close to the uniform distribution in [0, 1]; in other words, $p \approx \chi$. For example, *p* is equal to ~0.0952 when $\chi = 0.1$ and $n = 1000$. This means that by collapsing branches shorter than $0.1/1000$ in the Bayesian bootstrap trees, the theoretical support of branches corresponding to 1 mutation in the reference tree will be 0.9048. In short, collapsing branches shorter than $\chi(=10\%)/n$ gives support very close to $1-\chi(=90\%)$.

Now, let us examine how this analysis works in the more general case where a branch of the reference tree corresponds to *k* mutations. In this case, we consider the sum S_k of *k* weights, each distributed according to the same law (see above), and we search for the probability that S_k is less than χ and that the corresponding branch is collapsed. Using the exponential approximation (fully justified in practice, since *n* is rarely less than 100), we are interested in the distribution of a sum of exponential laws with parameter 1.0, which is an Erlang law whose distribution function is easy to compute. We find theoretical supports of 0.9048, 0.9953, 0.9998, and 0.9999 for *k* = 1, 2, 3, and 4 mutations.

The results in Figure 1(d) show that these theoretical values are close to the values observed with our "perfect" data set: for 1 mutation, the supports are 0.897, 0.884, and 0.896 (expected value = 0.904); for 2 mutations: 0.994 and 0.992 (expected value = 0.995); for 3 and 4 mutations: 1.0 (expected value >0.9998). We thus have a method to approximate Bayesian supports as a function of branch length. We will see on large simulated and real data sets that this approximation is close to the values of Bayesian support observed for correct branches, when the data do not deviate drastically from the perfect phylogeny assumption. It may seem arbitrary that the support depends on the collapse threshold (χ/n). However, in the standard frequentist bootstrap, the support is just as arbitrary and not related to any branch probability (e.g., 0.632 for 1 mutation, see above). An advantage of the Bayesian approach is that we can control this support to make it meaningful and interpretable in practice by adjusting the collapse threshold. The simulation results below confirm the relevance of the approach, which produces interpretable, probability-like supports with a collapse threshold based on $\chi = 1/10$.

Simulated data

To evaluate the properties of the various approaches and the robustness of our theoretical results with large realistic datasets, we simulated sequences inspired by the Ebola dataset (Dudas et al. 2017; Zhukova et al. 2023) that is analyzed below. The complete Ebola MSA consists of 1,610 sequences and 18,996 sites. For computational purposes, we randomly subsampled this dataset to select 800 sequences. We inferred a tree from these 800 sequences using PhyML+GTR+G4 (Guindon et al. 2010). This "Ebola" tree has a total length of 0.0926, corresponding to ~1760 mutations (i.e. ~1.10 mutation per branch). We used the Ebola tree to simulate the evolution of sequences. However, to avoid having too many zero-length branches with simulated data, we replaced all near-zero branch lengths in the Ebola model tree with values drawn from an exponential distribution with an expectation corresponding to 1/2 mutation. Such correction is required because in simulations short branches can generate zero mutation, even if their length is larger than zero. The resulting tree (same topology, slightly longer length) is called "CorrectedEbola" tree in the following.

We rooted the CorrectedEbola tree using mid-point rooting and took as ancestral sequence the sequence made up of the most common nucleotide at each site in the original alignment. This root sequence was evolved along the tree branches using the GTR+G4 model estimated by PhyML with Ebola sequences. We then inferred a tree from the simulated sequences using PhyML+GTR+G4. This "Simulated" tree is very close to the Ebola tree. When collapsing the branches with nearly zero length, the Simulated tree has 421 internal branches in common with the CorrectedEbola tree, and only 1 specific internal branch. This confirms that short trees are not fully resolved (with many zero-length branches), but almost all non-zero-length branches are correct (Fig. 3). We also measured the level of homoplasy in this dataset. Based on the simulated MSA, the minimum number of mutations is equal to 1,942. To obtain this result, we assume a perfect phylogeny and count for each site the number of characters present in the site and subtract 1 (e.g., 2 characters = 1 mutation); then we sum the values thus obtained for all sites. The actual size of the tree, in number of mutations (sum of branch lengths x alignment length), inferred from the simulated MSA is equal to 2,217, which (as expected) is larger than the theoretical minimum, due to homoplasy, convergence, reversion, etc. However, both values are close with 14% additional mutations $(=(2217-1942)/1942))$. The same measure applied to the Ebola MSA and tree is equal to 15%, indicating the closeness of the two datasets in terms of homoplasy. Finally, the branch lengths in the Simulated tree show a similar distribution to that of the Ebola tree, with very clear spikes corresponding to 0, 1, 2, 3 ... mutations (Fig. 2). However, as expected, the spikes are slightly less pronounced in the real data than in the simulated data, especially when the number of mutations is large, with a few intermediate lengths corresponding to non-integer numbers of mutations.

To study the properties and limitations of the approaches, we simulated MSAs and trees that gradually deviate from the assumptions of perfect phylogeny up to conditions where these assumptions are clearly violated. To achieve this we multiplied the branch lengths in the CorrectedEbola tree by a constant factor, and divided the number of sites in the MSA by the same factor. In this way, the number of expected mutations remains constant, i.e. the phylogenetic signal remains similar, but the degree of homoplasy increases and one gets further and further away from a perfect phylogeny. The factors used in our simulations were equal to 4, 16 and 64. For example, with a factor of 64, the tree length is close to 6.0 mutations per site (instead of 0.0926), while the length of the MSA is 296 (instead of 18,996), a radically different setting corresponding to, for example, animal or plant genes rather than virus

genomes. Sequence simulation and tree inference were then performed using the same programs and options as for the original CorrectedEbola tree. The estimated level of homoplasy (relative to perfect phylogeny, see above) was equal to 60%, 222%, and 518% for factors of 4, 16, and 64, respectively, a clear violation of the assumptions of perfect phylogeny (= 0%). In the following, the settings corresponding to the different multiplicative factors are called "X1" (i.e. the original setting corresponding to Ebola data), "X4", "X16" and "X64".

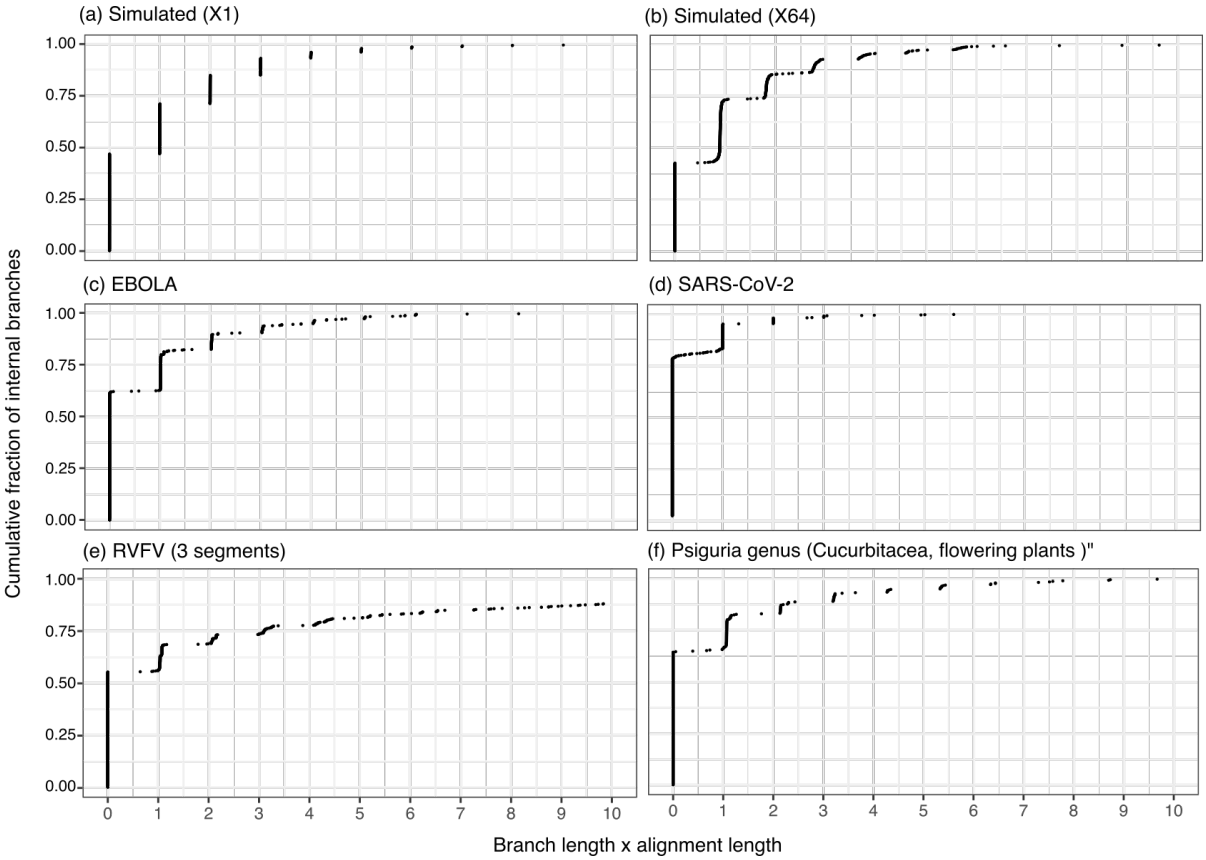

Figure 2: Cumulative distribution of internal branch length estimates. Branch lengths (X-axis) are expressed in number of substitutions (i.e., standard branch lengths multiplied by MSA length). Y-axis: cumulative fraction of internal branches. Panel **(a)**: Simulated data in X1 (Ebola-like) condition. Panel **(b)**: Simulated data in X64 (high-homoplasy) condition. Panel **(c)**: Ebola data (Dudas et al. 2017). Panel **(d)**: SARS-CoV-2 data (Zhukova et al. 2021). Panel **(e)**: Three datasets of RVFV virus, segments M, L, V (combined distribution of the 3). Panel **(f)**: Psiguria genus (Cucubirtacea, flowering plants) low-homoplasy dataset. See text and Supplementary Table 1 for details.

We applied three bootstrap methods to each of these simulated MSA and trees: (1) the original frequentist approach of Felsenstein (1985), where all branches in the bootstrap trees are kept, even if they have nearly zero length (called "NoCollapse" in the following); (2) the frequentist approach where bootstrap branches of length $< 0.1/n$ are collapsed (called "Frequentist" in the following; n is the length of the MSA, see explanations above); (3) our Bayesian approach where the sites are weighted and the bootstrap branches with length $< 0.1/n$ are collapsed (called "Bayesian" in the following). For each of the four simulated trees (corresponding to X1, X4, X16, and X64) and each branch of these trees, we thus obtained three supports corresponding to the three bootstrap methods. The branches were classified

depending on their length in the Simulated trees. The “0 mutation” branches have length estimates less than $0.5/n$, the “1 mutation” branches have length estimates in the range $[0.5, 1.5[/n$, the “2 mutations” branches in the range $[1.5, 2.5[/n$, etc. Indeed, we do not expect the estimated branch lengths to correspond exactly to 0, 1, 2 ... mutations, although the distribution of branch length estimates (Fig. 2) shows that we are relatively close to this pattern, especially with the Ebola-like (X1) data and tree. In addition, we distinguished the "true" and "false" branches in simulated trees, depending on whether the corresponding bipartitions are found in the model (Ebola) tree or not. We counted as false those branches that (1) were not present in the Ebola tree, or (2) were present in the Ebola tree but along which no mutation had occurred in the simulation of sequence evolution, since such branches can only be recovered by chance (see Supp. Fig. S2 for additional results on zero-length branches, where false branches are simply those that are not in the Ebola tree).

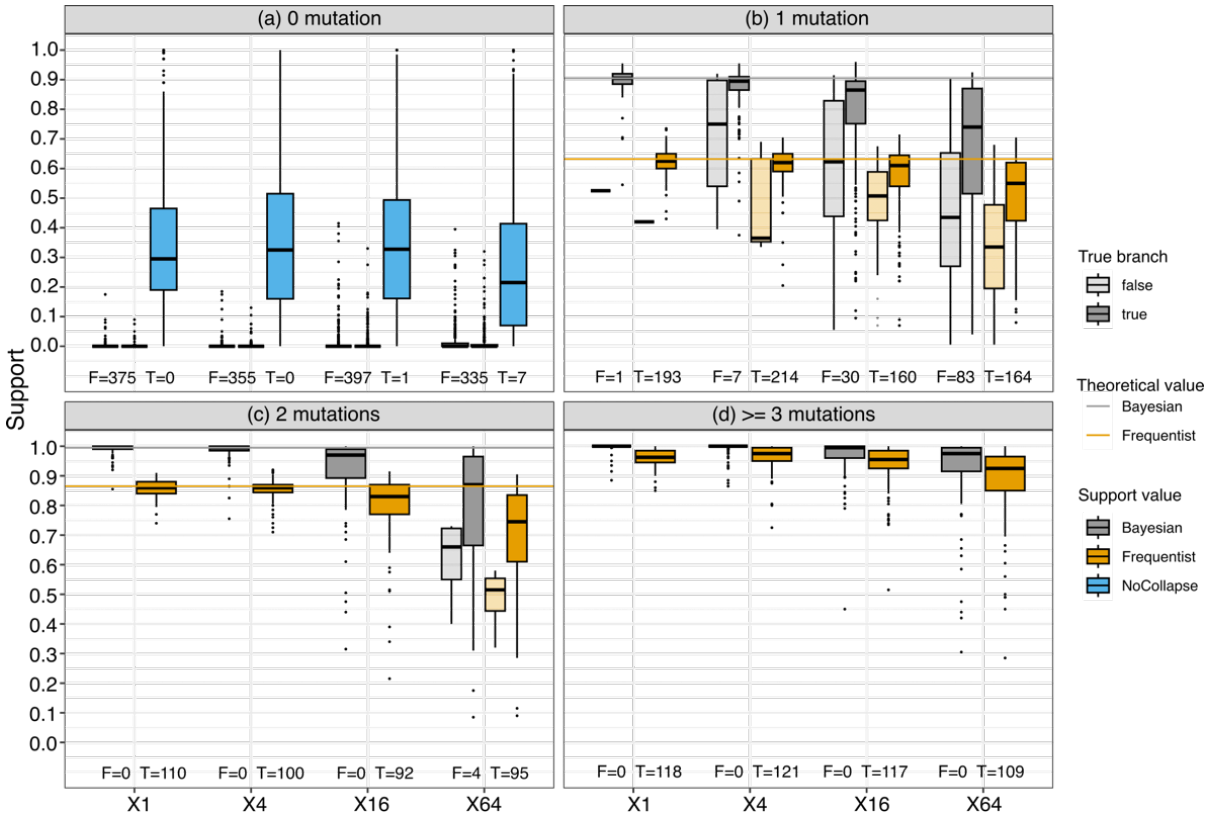

Figure 3: Simulation results. X-axis: the four experimental conditions, X1 (Ebola-like, low homoplasy: 14%), X4 (moderate homoplasy: 60%), X16 (substantial homoplasy: 222%), and X64 (high homoplasy: 518%). Y-axis: branch supports, between 0 and 1. Panels **(a)** to **(d)**: supports of branches with estimated lengths in the range [0, 0.5[, [0.5, 1.5[, [1.5, 2.5[and [2.5, ∞[, respectively. NoCollapse: standard frequentist bootstrap, where short bootstrap branches are not collapsed. Frequentist: frequentist bootstrap, where bootstrap branches with length corresponding to <0.1 mutation are collapsed. Bayesian: our approach with weighted sites, where branches with length corresponding to <0.1 mutation are collapsed. Within each panel, we provide the number of false (F) and true (T) branches, for each experimental condition (e.g., F = 335 and T = 7 for 0 mutation and X64). We also provide the theoretical support values for the Frequentist and Bayesian approaches, under perfect phylogeny assumptions. See text for details.

The results are shown in Figure 2 and 3:

As we approach perfect phylogeny, the estimated branch lengths (multiplied by the MSA length) are remarkably close to an integer number of mutations (Fig. 2(a), X1). This trend can also be seen in the analysis of real virus data (Fig. 2(c) and 2(d); see below). As expected, this trend diminishes as we move away from perfect phylogeny and the phylogenetic signal is perturbed (Fig. 2(b), X64). Under these conditions, the peaks of the distribution over integer values remain visible, but they are less pronounced due to the high level of homoplasy.

When the estimated length corresponds to ~0 mutations (Fig. 3(a)), the inferred branches are generally false (e.g., 98% and 100% in X64 and X1 conditions); they are either not present in the Ebola tree, or they are present in the Ebola tree but no mutations occurred along these branches in the simulations and they can only be found by chance. However, some of these branches have high support with the standard frequentist NoCollapse approach, as predicted by the theoretical analyses above. Conversely, when the null branches of bootstrap trees are collapsed, the support of these branches (essentially false or unrecoverable) becomes low for both the frequentist and Bayesian approaches. This demonstrates the importance of collapsing the null branches of bootstrap trees, regardless of the type of support. In this case (Fig. 3(a), in contrast to Fig. 3(b, c, d)), the distributions of the three supports remain fairly stable, regardless of the level of homoplasy.

With an estimated length of ~1 mutation (Fig. 3(b)), the signal becomes worse (as expected) when the multiplicative factors are high: with X1 we find 1 false branch out of 194 (~0.5%), while with X64 ~35% of branches are false. The Bayesian support is significantly higher than the Frequentist support. For X1, X4, and to some extent X16, the median (Frequentist and Bayesian) supports of the true branches are close to theoretical values (63% and 90%, respectively), although we are relatively far from a perfect phylogeny, especially for X16 (homoplasy = 220%). For X64 (very far from perfect phylogeny, homoplasy = 518%), we deviate from the theoretical values, and the observed supports are significantly lower, even for the true branches. In this case (as in others), the bootstrap supports (Frequentist and Bayesian) of the false branches are significantly lower than those of the true branches, justifying (if necessary) the use of the bootstrap. The length of branches gives a good indication of their support, but cannot be used alone, for example in the presence of contradictions between sites.

With an estimated length of ~2 mutations (Fig. 3(c)), very few branches are false (4 in total for X64). The difference between Frequentist and Bayesian bootstrap is still significant, and the values observed for X1, X4, and to some extent X16 are still close to the theoretical values (86% and 99%, respectively) derived from perfect phylogeny assumptions. Again, the support values for X64 deviate significantly from the theoretical values, and the 4 false branches have much lower bootstrap support (Frequentist and Bayesian) than the true branches.

With an estimated length of ~3 mutations and more, there are no longer any false branches, and the Bayesian and Frequentist supports come very close to 1.0, even for X64.

In Figure 3, we see again what we observed in Figure 1: (1) it is essential to collapse null branches; (2) with low homoplasy the Bayesian support is higher and more interpretable than the frequentist bootstrap, with values close to 100% for correct branches; (3) the two supports are similar

and close to 100% for long branches, which are generally correct. Moreover, as expected, the two supports result in substantially different values but are highly correlated (Pearson correlation of 98% in the 4 simulated conditions).

We compared the performance of the two approaches, Frequentist and Bayesian, using the AUC (Area Under the "receiver operating characteristic" (ROC) Curve), which provides a measure of the trade-off between Type 1 error (number of false branches that are supported, divided by the number of false branches) and power (number of true branches that are supported, divided by the number of true branches). The AUC integrates results over all possible selection thresholds between 0% and 100% and is not dependent on any particular threshold. We have grouped the results for all four conditions to give an overall view, and because the number of false positives for `X1` and `X4` is very low. The two ROC curves are similar (Supp. Fig. S3), with a slight advantage for the Bayesian approach over the Frequentist approach (AUC = 0.919 and 0.903, respectively).

In practice, bootstrap users want to know the threshold at which branches should be considered well supported. In the case of Frequentist bootstrap, an empirical threshold of 70% is commonly used, as suggested by Hillis and Bull (1993), Soltis and Soltis (2003), and others. For the Bayesian bootstrap, our analyses above suggest a threshold of ~90%, which is easier to interpret in the context of a test. We measured the performance of these two thresholds based on all four experiments grouped together, for non-zero branches (estimated length ≥0.5 mutations, i.e. with some phylogenetic signal). Under these conditions, the Frequentist bootstrap with a threshold of 70% selects no false branches (Type 1 = 0%), but the number of true branches that are supported is not very high (power = 50%). For the same conditions, the Bayesian approach with a threshold of 90% selects 5 false branches (Type 1 = 4%) and increases the power with 62% of true branches supported. These two thresholds are therefore reasonable (very low Type 1 error), but appear conservative in a testing framework where the alpha risk is generally set at 5% (and not 0% or 1%) to achieve greater power. By specifying a Type 1 error of 5%, the selection thresholds for the Frequentist and Bayesian approaches are 66% and 89%, with power of 56% and 68%, respectively. With a Type 1 error of 10%, these thresholds become 63% and 85%, with powers of 69% and 80%, respectively. Once again, we see a slight superiority of the Bayesian approach, with higher power for the same level of risk.

To summarize the results of these experiments: (1) Across all four conditions (with very different levels of homoplasy), the Bayesian bootstrap performs slightly better in terms of AUC. (2) A threshold of 70% for the Frequentist bootstrap seems very conservative; the optimal threshold is closer to 65% (or slightly less, given the theoretical value of 63.2%). (3) A 90% threshold for the Bayesian bootstrap is less conservative, giving higher power and a natural interpretation of the supports; the optimal threshold is slightly lower, between 85% and 89%, depending on the level of risk that is accepted.

Bayesian and frequentist supports with virus and low homoplasy phylogenies

We used three virus datasets, from Ebola in West Africa (Dudas et al. 2017), from the global SARS-CoV-2 pandemic (Zhukova et al. 2021), and from Rift Valley Fever Virus (RVFV) downloaded from NCBI Genbank. We also used one large low-homoplasy plant MSA collected from TreeBASE (Piel et al. 2009; Togkousidis et al. 2023). Details on all these datasets (and the simulated ones) are provided in Supplementary Table 1.

The West Africa Ebola epidemic began in 2013 in Guinea and circulated for two and a half years in neighboring Liberia and Sierra Leone. We downloaded the Ebola full-genome alignment of (Dudas et al. 2017) from (https://github.com/evolbioinfo/bdei). This multiple sequence alignment (MSA) consists of 1,610 sequences of 18,996 nucleotides. From this MSA, we randomly subsampled 800 sequences for computational reasons. We inferred the reference tree using PhyML with GTR+G4 and default options. To infer frequentist bootstrap trees, we generated 200 bootstrap alignments by sampling sites with replacement (as usual) and used PhyML+GTR+G4. To infer Bayesian bootstrap trees, we generated 200 weight vectors using a Dirichlet distribution with parameters (18996; 1,…,1) and used PhyML, which allows for continuous site weights, with GTR+G4 and default options. All these steps were performed using our BBOOT pipeline (see Supp. Mat. for details). The NoCollapse supports were computed as usual, while for the Frequentist and Bayesian supports, we first collapsed the bootstrap branches with length less than $0.1\,(\text{mutation})/18996\,(\text{MSA length})$. In all this, we used the very same approach and program options as for the simulated data, thus making the results comparable to the X1 (Ebola-like) condition. In fact, the amount of homoplasy is similar for the two datasets (~15%).

Our SARS-CoV-2 dataset covers the first five months of the pandemics (up to April 25, 2020). It was subsampled to avoid over-representation of certain countries and periods (Zhukova et al. 2021). We extracted the identifiers from the publication and downloaded the sequences from GISAID. We again randomly subsampled the sequences for computational reasons and aligned them to obtain an MSA of 796 sequences with a length of 29,903, representing the emergence of the global pandemic. Alignment sites known to contain noisy phylogenetic signal were masked. The reference and bootstrap trees were inferred using the same protocol as for the Ebola MSA, with a collapse threshold of $0.1/29903$. The size of the tree (in number of mutations) is 1062, while assuming a perfect phylogeny, the lower bound for the number of mutations is 787, resulting in a homoplasy level of 35%. This is appreciably higher than for the Ebola dataset (15%), although the number of mutations is significantly lower (~0.67 mutations per branch versus ~1.10 with Ebola).

Our RVFV dataset was downloaded from Genbank (on April 17, 2024) and consists of 300, 240, and 237 sequences of its three respective segments (S, M, and L). We aligned the three segments separately to obtain three MSAs of length 1690, 3885, and 6404 nucleotides for the S, M, and L segments, respectively. The reference and bootstrap trees were inferred using the same protocol as before, with a collapse threshold of $0.1/\text{MSA length}$. The lengths of the three trees in number of mutations are 955, 2177, and 2976 for S, M, and L segments (i.e. ~1.6, ~4.6 and ~6.3 mutations per branch), while assuming a perfect phylogeny, the lower bounds for the number of mutations are 505, 1302, and 1778, resulting in estimated homoplasy levels of 89%, 67%, and 67%, respectively. This is significantly higher than for the two previous data sets (15% and 35%) and similar to the simulated X4 condition. The branch lengths and supports of the three trees were combined in the results.

To investigate the behavior of bootstrap supports with non-viral data, we used a large low-homoplasy dataset collected from TreeBASE (Piel et al. 2009; Togkousidis et al. 2023; Supp. Mat.). We used the largest nucleotidic MSA with a homoplasy level of less than 50%, containing 385 (324 unique) sequences from the genus Psiguria (Cucubirtacea, flowering plants). This genus has 6 species, the dataset is both inter and intra specific. The number of sites is 1140 (much less than Ebola and SARS-CoV-2, but similar to RVFV segment S). The number of mutations per branch is 1.1 (similar to viruses) and the level of homoplasy is 46%, which is relatively high compared to viruses. The reference and bootstrap trees were inferred using the same protocol as for viruses.

The results for these four datasets are displayed in Figure 2 (distribution of branch lengths) and Figure 4 (branch supports). We observe:

Tree branch lengths are similar between the simulated data in condition X1 and the real Ebola data (Figures 2(a) and 2(c)), with marked peaks on the first integer values corresponding to 0, 1, 2, and 3 mutations. However, with the real data (Fig. 2(c)), there are intermediate points between these integer numbers of mutations, and the graph becomes more blurred beyond 3 mutations. This is a sign of a less perfect phylogenetic signal than with the simulated data, in particular due to some violation of the substitution model (GTR+G4) used to infer the tree and estimate branch lengths. With SARS-CoV-2, there are fewer mutations than with Ebola, and many branches of zero length (~80%). There are peaks at 0, 1 and 2 mutations, but a substantial number of intermediate points, particularly between 0 and 1 mutations. This indicates an increased deviation from the perfect phylogeny, as indicated by the homoplasy level (35%), due to misspecification of the substitution model and probably also because convergences in these sequences were observed very early (Van Dorp et al. 2020), thus perturbing the phylogenetic signal. Despite a rather high homoplasy level (>65%) in the RVFV trees and MSAs, we still see relatively well-marked peaks for 0, 1, 2, and 3 mutations, but the values are less well centered on integer values than for Ebola. The same can be seen for the plant dataset (homoplasy = 46%), the peaks are visible with 0, 1, 2, and 3 mutations, but they tend to be shifted to the right due to model violation and the presence of duplicated sequences, gaps and unknowns that blur the signal and make it difficult to assign a mutation to a unique branch in the tree.

Despite these differences in branch length distribution and a noticeable deviation from perfect phylogeny, the branch supports (Fig. 4) are close to the theoretical values and to the results obtained with simulated data (Fig. 3). We see again (Fig. 4A-B-C-D(a), 0 mutation) the importance of collapsing the short branches of bootstrap trees, without which many inferred branches not supported by any mutation have substantial support. For some datasets (e.g. SARS-CoV-2, Fig. 4B(a)), there are still short branches that receive Frequentist and Bayesian support quite far from 0.0, presumably because of their intermediate length, but these supports remain substantially below the selection thresholds of 65% (Frequentist) and 85% (Bayesian) that emerge from the analyses of simulated data. Branches corresponding to 0 mutations can therefore reasonably be considered false. For branches corresponding to 1 mutation or more, the median supports for both Ebola and SARS-CoV-2 datasets are very close to the theoretical values. By using thresholds of 85% for the Bayesian bootstrap and 65% for the Frequentist bootstrap, we are very likely selecting branches that correspond to the true evolutionary history of these viruses. For RVFV and plant trees, branches corresponding to approximately 1 mutation

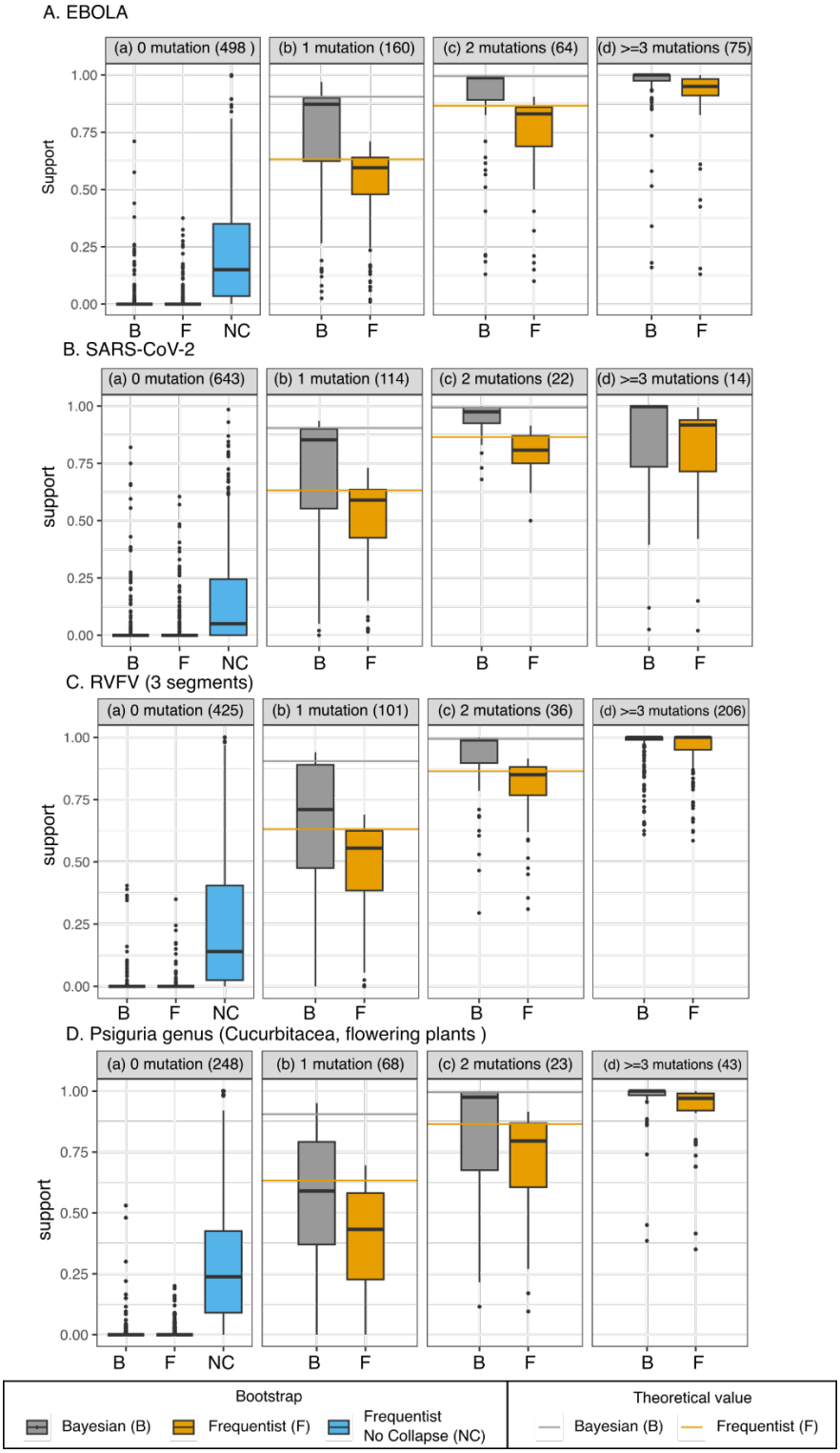

Figure 4: Analysis of viral and non-viral datasets. **A:** Ebola (Dudas et al. 2017); **B:** SARS-CoV-2 (Zhukova et al. 2021); **C:** Rift Valley Fever Virus (RVFV; extracted from Genbank); **D:** Psiguria genus (plants) low-homoplasy dataset collected from TreeBASE. Panels **(a)** to **(d)**: supports of branches with estimated lengths in the range [0, 0.5[, [0.5, 1.5[, [1.5, 2.5[, and [2.5, ∞[, respectively. NoCollapse: standard frequentist bootstrap, where short bootstrap branches are not collapsed. Frequentist: frequentist bootstrap, where bootstrap branches with length corresponding to <0.1 mutation are collapsed. Bayesian: our approach with weighted sites, where branches with length corresponding to <0.1 mutation are collapsed. Within each panel, we provide the theoretical support values for the Frequentist and Bayesian approaches, under perfect phylogeny assumptions. See text and Supplementary Table 1 for details.

in length have Frequentist and Bayesian supports below the theoretical values (Fig. 4C-D(b)), as expected given the relatively high level of homoplasy in these data. With thresholds of 85% (Bayesian) and 65% (Frequentist), very few branches of length ~1 will be selected, which is likely conservative despite the rather low level of these thresholds. For branch lengths of about 2 mutations and more, the results are consistent with the theoretical results for all data sets (Fig. 4A-B-C-D(c-d)), and the use of thresholds of 85% (Bayesian) and 65% (Frequentist) should ensure the selection of mostly correct branches. However, it is not possible to know (as with the simulated data) whether one support gives better accuracy than the other. It is likely that their performances are close (their Pearson correlation is ≥98% on the 4 datasets), but the Bayesian bootstrap shows again the advantage over the frequentist bootstrap of corresponding to thresholds that are more interpretable in the usual statistical sense.

Discussion

The Bayesian bootstrap is a smoothed version of the frequentist bootstrap in which the sampling with replacement procedure is replaced by weights on the statistical units (or observations) drawn from a non-informative distribution (Rubin 1981; Lo 1991 ; Shao and Tu 1995). Here, we have investigated the application of this principle (combined with the collapse of zero-length bootstrap branches) to the estimation of the branch supports of a phylogeny. Our results demonstrate the interest of this smoothed approach for short trees and branches inferred from highly conserved data sets with low numbers of mutations, as we observe in viruses (but not only). Such datasets contain a partial but reliable evolutionary signal, with parts of the inferred phylogenies that are poorly resolved, but few false branches (Wertheim et al. 2022). Our theoretical and experimental results on simulated and real data show that Bayesian supports are higher than standard frequentist supports, and more consistent with the usual statistical interpretation of a test. A selection threshold of 85-90% gives good results on sequences with low homoplasy, with very few false positives and fairly good power. To achieve similar performance with the standard bootstrap, a selection threshold of 60-65% must be used. In our simulations, the Bayesian bootstrap shows slightly better performance (ROC curves and AUC, Supp. Fig. S3), probably due to its smoothness. Our results also show that the two supports are strongly correlated, and that as the trees and branches lengthen, with more mutations and a high level of homoplasy, the two supports become very close (as expected from bootstrap theory; Efron and Tibshirani 1993).

For this type of data with low homoplasy, bootstrap supports are strongly correlated with branch lengths (Felsenstein 1985). This confirms recent results on the predictability of branch supports by machine learning (Wiegert et al. 2024; Ecker et al. 2024). However, the usefulness of the bootstrap remains, especially for short branches corresponding to 1 or 2 mutations, where the signal can be conflicting and blurred by the homoplasy of the data, even if it is low (Fig. 3(b), 3(d)). Our results show that to properly interpret the results of a bootstrap analysis, it is important to measure the level of homoplasy in the data and to plot the distribution of branch lengths to verify that the lengths cluster (or do not cluster) around integer values of the mutation number (Fig. 2). These statistics (homoplasy level, branch length distribution) and the branch supports of the Bayesian bootstrap can be obtained using our BBOOT pipeline (https://github.com/evolbioinfo/bayesian_bootstrap) based on PhyML for tree inference (Guindon et al. 2010). The computing time for the Bayesian phylogenetic bootstrap is the same as for the standard approach.

To compare the behavior of frequentist and Bayesian bootstrap supports on a wide variety of sequences, we applied these approaches to large datasets collected from TreeBASE (Piel et al. 2009; Togkousidis et al. 2023; Supp. Mat.). The results for DNA (300 datasets, Supp. Fig. S4) show that: (1) homoplasy levels in these datasets are highly variable, from less than 1% to over 30,000%, with an average of 1,210%; we are therefore often far from the homoplasy levels studied here (between 15% and 89%; Supp. Table 1), but these datasets are peculiar, virus-free (except for one Dengue MSA) and often from deep phylogenies. (2) Branch support differences between the two bootstrap approaches depend on the level of homoplasy (as expected). The differences are small (<10%) when homoplasy exceeds 1000%, but Bayesian supports remain higher than frequentist supports. (3) For each of the DNA data sets, the correlation between the two branch supports is very high (≥93%, 98% on average). The results are similar for proteins (100 datasets, Supp. Fig. S5), but we did not find any large dataset with low homoplasy (minimum value = 182%, average = 1,351%), probably because protein datasets in TreeBASE were generally used to elucidate the deep branches of the tree of life. Again, we observe that the gap between the two supports narrows with high homoplasy (<6% with >1000% homoplasy), and that the correlation between the two supports is strong for all datasets (>95%, 99% on average).

Taken together, our results show that the standard frequentist version of the bootstrap would benefit from being replaced by the Bayesian version, at least in the case of DNA data. The Bayesian bootstrap is smoother, its supports are higher and therefore less conservative, especially when there are few mutations and homoplasy is low. Several avenues are open for future research. In particular: (1) Theoretical and experimental comparison with the usual Bayesian approaches based on the posterior probability of branches (Huelsenbeck and Rannala 2004), and with the local aBayes support (Anisimova et al. 2011). (2) Development of fast methods, in continuation of the Rapid and Ultrafast bootstrap algorithms developed for the frequentist version (Stamatakis et al. 2008; Minh et al. 2013). (3) Application of the Bayesian approach to the Transfer Bootstrap Expectation (TBE; Lemoine et al. 2018) and to the RELL bootstrap used in the comparison and test of phylogenetic trees (Kishino et al. 1990; Goldman et al. 2000; Shimodaira and Hasegawa 2001).

Data availability: All our datasets and workflows are available from (see Supp. Mat. for details): https://github.com/evolbioinfo/bayesian_bootstrap_analyses/.

Acknowledgements: The authors would like to thank Stéphane Guindon for his help with PhyML and Marie Morel for early discussions and preliminary analyses.

References

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

Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60,** 685–699.

Berry, V., Gascuel, O. (1996). On the Interpretation of Bootstrap Trees: Appropriate Threshold of Clade Selection and Induced Gain, *Mol. Biol. Evol.* **13**, 999–1011.

Ecker, N., Huchon, D., Mansour, Y., Mayrose, I., Pupko, T. (2024). A machine-learning based alternative to phylogenetic bootstrap. *Bioinformatics* (in press). https://doi.org/10.1093/bioinformatics/btae255

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **7,** 1–26.

Efron, B. & Tibshirani R. J. (1993). An introduction to the bootstrap, Chapman & Hall, NY.

Efron, B., Halloran, E. & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci.* **93,** 7085-7096.

Efron, B. & Tibshirani R. J. (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Am. Stat. Assoc.*, **7**, 548–560.

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters, *Syst. Biol.* **22**, 240–249.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39,** 783–791.

Felsenstein, J., Kishino, H. (1993). Is there something wrong with the bootstrap on phylogenies a reply to Hillis and Bull. *Syst. Biol.* **42,** 193–200.

Gascuel, O., Lemoine, F. (2024). Measures of Branch Support in Phylogenetics, in *Models and Methods for Biological Evolution: Mathematical Models and Algorithms to Study Evolution* (Gilles, D., and Guindon, S., Eds), John Wiley & Sons, 213-234.

Goldman, N., Anderson, J. P., Rodrigo, A. G. (2000). Likelihood-Based Tests of Topologies in Phylogenetics, *Syst. Bio.* **49**, 652-670.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010). New Algorithms and Mehtods to Estimate Maximum-Likelihood Phylogenies: Asessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.

Gusfield, D. (1997). Algorithms on strings, trees and sequences, New York: Cambridge University Press.

Hillis, D. M. & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phyogenetic analysis. *Syst. Biol.* **42,** 182–192.

Holmes, S. (2003). Bootstrapping Phylogenetic Trees: Theory and Methods. *Stat. Sci.* **18,** 241–255.

Huelsenbeck, J. & Rannala, B. (2004). Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. *Syst. Biol.* **53,** 904–913.

Jahn, K., Kuipers, J. & Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biol.* **17**, 86.

Kishino, H., Miyamata, T., Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31**, 151–160.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455.

Lemoine, F., Domelevo Entfellner, JB., Wilkinson, E. et al. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456.

Lo A. Y. (1991). Bayesian Bootstrap Clones and a Biometry Function. *Indian J. Stat.* **53**, 320-333.

Minh, B. Q., Nguyen, M. A. T. & Von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30,** 1188–1195.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol.* **37**, 1530–1534.

Piel W. H., Chan L., Dominus M. J., Ruan J., Vos R. A., Tannen V. (2009) TreeBASE v. 2: A Database of Phylogenetic Knowledge, *e-BioSphere* 2009.

Sanderson, M. J. (1995). Objections to bootstrapping phylogenies: A critique. *Syst. Biol.* **44,** 299–320.

Sanderson, M. J. & Shaffer, H. B. (2002). Troubleshooting Molecular Phylogenetic Analyses. *Annu. Rev. Ecol. Syst.* **33,** 49–72.

Shimodaira, H., Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-1247.

Soltis, D. E. & Soltis, P. S. (2003). Applying the Bootstrap in Phylogeny Reconstruction. *Stat. Sci.* **18,** 256–267

Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst. Biol.* **57,** 758–771.

Susko, E. (2009). Bootstrap support is not first-order correct. *Syst. Biol.* **58,** 211–223.

Susko, E. (2010). First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. *Mol. Biol. Evol.* **27,** 1621–1629.

Rubin D. B. (1981). The Bayesian Bootstrap. *Ann. Stat.* **9**, 130-134.

Shao, J., Tu, D. (1995). The Jackknife and Bootstrap. Springer-Verlag, NY.

Togkousidis, A., Kozlov, O. M., Haag, J., Höhler, D., Stamatakis, A. (2023). Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty, *Mol. Biol. Evol.* **40**, msad227.

Van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. (2020) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol.* **83**, 104351.

Van Noorden, R., Maher, B. & Nuzzo, (2014). R. The top 100 papers. *Nature* **514,** 550–553.

Wertheim, J. O., Steel, M., Sanderson, M. J. (2022). Accuracy in Near-Perfect Virus Phylogenies. *Syst Biol.* **71**, 426-438.

Wiegert, J., Höhler, D., Haag, J., Stamatakis, A. (2024). Predicting Phylogenetic Bootstrap Values via Machine Learning, bioRxiv, doi.org/10.1101/2024.03.04.583288.

Zharkikh, A. & Li, W. H. (1995). Estimation of Confidence in Phylogeny: The Complete-and-Partial Bootstrap Technique. *Mol. Phyl. Evol.* **4,** 44–63.

Zhukova, A., Blassel, L., Lemoine, F., Morel, M., Voznica, J., Gascuel, O. (2021). Origin, evolution and global spread of SARS-CoV-2, *Comptes Rendus. Biologies* **344**, 57-75.

Zhukova, A., Hecht, F., Maday, Y., Gascuel, O. (2023). Fast and Accurate Maximum-Likelihood Estimation of Multi-Type Birth–Death Epidemiological Models from Phylogenetic Trees, *Syst. Biol.* **72**, 1387–1402.