Biol Philos (2008) 23:455–473
DOI 10.1007/s10539-007-9105-7

The prior probabilities of phylogenetic trees

Joel D. Velasco

Received: 25 June 2007 / Accepted: 16 December 2007 / Published online: 17 January 2008

Abstract Bayesian methods have become among the most popular methods in phylogenetics, but theoretical opposition to this methodology remains. After providing an introduction to Bayesian theory in this context, I attempt to tackle the problem mentioned most often in the literature: the "problem of the priors"—how to assign prior probabilities to tree hypotheses. I first argue that a recent objection—that an appropriate assignment of priors is impossible—is based on a misunderstanding of what ignorance and bias are. I then consider different methods of assigning prior probabilities to trees. I argue that priors need to be derived from an understanding of how distinct taxa have evolved and that the appropriate evolutionary model is captured by the Yule birth–death process. This process leads to a well-known statistical distribution over trees. Though further modifications may be necessary to model more complex aspects of the branching process, they must be modifications to parameters in an underlying Yule model. Ignoring these Yule priors commits a fallacy leading to mistaken inferences both about the trees themselves and about macroevolutionary processes more generally.

Keywords Base rate fallacy · Bayesianism · Phylogenetic trees · Phylogenetics · Prior probabilities · Systematics · Tree shape · Yule process

Introduction

Finding the solution to biological problems such as determining whether or not a Florida dentist passed HIV on to his patients (he did—Metzger et al. 2002), calculating whether or not brain size and testicle size are adaptively correlated in bats (they are anti-correlated—Pitnick et al. 2006), and determining how terrestrial

J. D. Velasco (✉)
Department of Philosophy, University of Wisconsin, 5185 White Hall, 600 North Park Street, Madison, WI 53706, USA
e-mail: jdvelasc@wisc.edu

mammals arrived in Madagascar (multiple separate rafting events rather than a land bridge—Poux et al. 2005) all require knowledge of the evolutionary history of certain groups. Recovering this history is the project of phylogenetic inference—the goal is to build a phylogeny, or genealogical history, of a group of genes, species, higher taxa, or whatever the objects of study happen to be.

This paper advocates the use of a particular methodology of phylogenetic inference—Bayesian inference. Before I offer any justification for this, I briefly describe the problem of phylogenetic inference and describe Bayesianism in this context. I then provide a minimal defense for the Bayesian approach. For many systematists, the reason to prefer other methods comes not from their belief in the correctness of their preferred methodology, but rather is a response to a supposed problem for Bayesianism—the "problem of the priors." By correcting serious misunderstandings about this problem and developing the beginning of a solution, I hope to bolster the overall defense of Bayesian phylogenetics.

In a typical problem of phylogenetic inference, we are concerned with recovering facts about the genealogies of particular biological groups at a variety of levels. The data used to construct these phylogenies can in theory be morphological, ecological, molecular, or any number of different types of information, but the majority of published phylogenies today come from DNA sequences of individual organisms. It is assumed that the true underlying history of these sequences is that of common ancestry and descent with modification. This history is then represented as a binary branching tree. The tips, or "leaves" of the tree are the DNA sequences and the internal nodes represent common ancestors—the points in the past of "coalescence" when the descendant sequences trace back to the same token sequence present in a single individual. Though a conclusion about the phylogeny of species is nearly always drawn, the philosophically minded reader is sure to recognize that moving from a tree of DNA sequences to a tree of another kind such as a species tree requires a conceptual leap; this second stage of inference needs a separate discussion of its own and can safely be ignored here.

The "phylogeny", the "evolutionary tree", or just simply "the tree" may or may not contain information such as branching dates, rates of change along branches, or ancestral character states, but it must give at least a branching diagram with the tips labeled. This information uniquely specifies any and all clades, or monophyletic groups, on the tree. This branching diagram is called the tree *topology* and is generally the primary object of inference for the systematist because knowledge of the topology is a prerequisite for most further inquiries about the history. Unless specified otherwise, "tree" here refers just to the tree topology.

Bayesian phylogenetics

Maximum Parsimony and Maximum Likelihood are two families of methods that have dominated phylogenetics discussion for the past 20 years and both have their advocates (Felsenstein 2004). Although there is a long and rich history of the study of Bayesian statistics generally, it is only in the past ten years that Bayesian methods of inference have been used in phylogenetic studies (Rannala and Yang 1996, Huelsenbeck et al. 2001). Bayesianism has taken some time to catch on in

popularity and the details and their consequences certainly have not been as widely discussed as those attaching to other methods (Randle et al. 2005). For example, Felsenstein in his attempt at a comprehensive textbook *Inferring Phylogenies* (2004) spends only one of 35 chapters on Bayesian methods. However, Bayesian methodology is gaining popularity with time and today it is widely used alongside other methods in published results.

The central idea in Bayesian phylogenetics is that all inferences should be made by utilizing the posterior probability distribution of the trees. Bayes' theorem has the following consequence:

$$\text{The probability that a tree is correct given the sequence data that we have} = \Pr(\text{Tree} \mid \text{Data}) = \frac{\Pr(\text{Data} \mid \text{Tree}) \times \Pr(\text{Tree})}{\Pr(\text{Data})}$$

Pr(Tree), called the prior probability of the tree, is determined from a probability distribution over all possible trees given before the data are examined. The probability of the data—Pr(Data)—is a normalizing constant simply used to make sure that the posterior probabilities sum to 1. It is equal to the sum of the probabilities of getting the data on every possible tree weighted by the particular tree's prior probability. Labeling each tree topology as T_1, T_2,...T_i, we have:

$$\Pr(\text{Data}) = \sum\nolimits_{\text{Ti}} \Pr(\text{Data} \mid \text{Ti}) \times \Pr(\text{Ti})$$

Pr(Data|Tree) is called the likelihood of the tree, but it cannot be directly calculated since the tree topology alone does not give us sufficient information to assign a probability to the data. Rather, we need additional information such as the branch lengths (the expected number of changes per site along a particular branch) along with some model of evolution that will contain its own parameters to be estimated such as the nucleotide substitution rates.

The Bayesian method for dealing with these nuisance parameters (parameters that aren't of primary interest) is to "average over" them by integrating them out. In the frequentist method called "Maximum Likelihood", for each tree, nuisance parameters such as branch lengths and substitution model parameters are set at the value that would maximize the probability of the data on that particular tree. The Maximum Likelihood tree is by definition the tree which is a conjunct in the tree-plus-nuisance-parameters conjunction which makes the data most probable. Thus, confusingly, the likelihood of the tree used in Bayes' Theorem is not the same as the tree's Likelihood score used for Maximum Likelihood inferences.

Treating nuisance parameters in the Bayesian way, if we denote a fixed set of branch lengths as v and a fixed set of parameter values of the model as θ we now have:

$$\Pr(\text{Data}|\text{Ti}) = \int_v \int_\theta \Pr(\text{Data}|\text{Ti}, v, \theta) \times \Pr(v, \theta|\text{Ti}) \text{d}v\text{d}\theta$$

Substitution in both the numerator and denominator yields this formula:

$$\Pr(\text{Ti}|\text{Data}) = \frac{\int_v \int_\theta \Pr(\text{Data}|\text{Ti}, v, \theta) \times \Pr(v, \theta|\text{Ti}) v\text{d}\theta \times \Pr(\text{Ti})}{\sum_{\text{Ti}} \int_v \int_\theta \Pr(\text{Data}|\text{Ti}, v, \theta) \times \Pr(v, \theta|\text{Ti}) \text{d}v\text{d}\theta \times \Pr(\text{Ti})}$$

The above formula tells us the posterior probability of any particular tree hypothesis. If we are interested in something else, say the probability that a particular group forms a clade, since this is, in effect, a large disjunction (the true tree could be any one of the trees that contains that clade), the posterior probability of that clade is simply the sum of the posterior probabilities of all trees which contain that clade. The probability distribution of any other parameters such as a branch length, the individual substitution rates, or the ratio of transitions to transversions are all similarly calculated. The Bayesian philosophy thus provides a framework for answering a host of relevant theoretical questions all at the same time. Of course actually calculating the full posterior distribution is another matter. However, there is reason to be hopeful here. Computational methods for numerically estimating multi-dimensional integrals are quite advanced. The standard idea is to use Markov Chain Monte Carlo (MCMC) methods to estimate the posterior distribution. For an introduction to these methods in phylogenetics see Larget and Simon (1999) and Larget (2005).

How we can actually infer the posterior probability and how we can do it in a computationally efficient manner are important practical questions, but it is to the theoretical issues that I now turn. These questions assume that we have access to the various posterior probability distributions and ask why we should use these to make our phylogenetic inferences rather than use some other quantity. If there are deep theoretical problems with Bayesian methodology, it hardly matters if we have an efficient way of calculating the relevant probabilities.

One important, though hardly decisive, consequence of Bayesian methodology is the ease of interpreting results. Since the posterior probability of a tree just is the probability that the tree is correct (given our data and our model of evolution), the tree with the highest posterior probability is the tree which is the best supported. In fact, the strength of its support is measured directly by the posterior. Other facts about the problem, such as which tree would require the fewest nucleotide substitutions, which is what the Parsimony score captures, are of interest only in so far as they are a reliable guide to which tree is true (which they often aren't).

In addition, unlike other methods, we can judge the strength of the evidence for all aspects of the tree at the same time without needing to reanalyze the data using different techniques. The probability that a particular group forms a clade, that two particular sequences have coalesced in the last one million years, that sequence A is more closely related to B than to C, or any other question about the tree is measured using the posterior distribution. None of these problems are easily analyzed with other methods which are usually designed just to find the best topology. While particular tests have been developed (see Felsenstein (2004) for a host of examples) none has a straightforward statistical interpretation that is useful and as such they generally appear to be disjoint, ad-hoc tests with no underlying, unified justification.

While a theoretical justification can be constructed for using posterior probabilities to guide our inferences, I will not attempt to do so here (for that and a host of similar references, see Howson and Urbach (2005)). Rather, I will focus on a few reasons why one might object to Bayesianism in this context. Some systematists believe that probabilities and perhaps even all statistical methods

simply cannot be used to make inferences concerning a particular group's evolutionary history since it is a "unique event" – meaning it has occurred only once (Siddall and Kluge 1997, but see Haber 2005) or believe that Parsimony has some special justification apart from its statistical behavior (see Farris 1983; Kluge 2005 or any of a host of papers in-between, but see Sober 1988). From those who are more statistically minded, there are worries that the posteriors might be overly sensitive to the choice of an evolutionary model or that Bayesian inference treats nuisance parameters as random variables and thus is not properly frequentist as Maximum Likelihood appears to be (though see Yang 2006). While these are important objections, they have been dealt with elsewhere (besides the above references, see for example Huelsenbeck and Ronquist 2005) and I will not discuss them further.

While there are certainly more issues to discuss, the central problem which has yet to be adequately dealt with and is perhaps the most common objection to Bayesian phylogenetics and to Bayesian inference more generally is the "problem of the priors"—how to assign prior probabilities to the hypotheses under test. As Felsenstein, an advocate of frequentist methods, puts it: "If the prior is agreed by all to be a valid one, then there can be no controversy about using Bayesian inference" (Felsenstein 2004: 300). While there would of course still be controversy, his point is that to the statistically minded theoretician, there shouldn't be.

While a subjective Bayesian may respond that prior probabilities ought simply to represent the prior beliefs of the particular investigator, it is certainly a worthwhile project to attempt to model a certain kind of ignorance for the use of priors. This is a direct attempt to avoid biasing the results in favor of our prior conceptions. After all, we want results that ought to be taken seriously by a wide range of scientists and we may want to know what "just this data" should lead us to believe. This is one of the goals of so-called "objective Bayesianism" (perhaps better called "Interpersonal Bayesianism"—Kadane 2006). As long as we have a proper understanding of ignorance, it would appear, at least in certain cases, that we should attempt to model ignorance in the priors. But there are many things that we appear to be ignorant about—the tree topology, its branch lengths, which groups form clades, etc. It might seem that modeling ignorance with respect to some of these factors is simple—for example, to model ignorance with respect to tree topologies we should assign equal prior probabilities for all topologies. This distribution is called a uniform prior on topologies. However, there are many different ways of conceiving of a tree. The shape of the tree refers to the branching diagram with no labels at the tips and so has less information than the topology. The *topology* is simply an unlabeled shape with labels added to the tips. In addition, we may be interested in more than just the topology. The *labeled history* (sometimes called "ranked topology"—e.g. Semple and Steel 2003) refers to the topology plus a temporal ordering of the nodes. These differences will become important later; they are depicted in Fig. 1.

To get a topology from a shape, labels are added to the tips. In this example, if "B" and "C" were switched, we would have the same shape but a different topology. Recall that topologies do not specify the time at which the nodes occur. In a labeled history, the nodes are labeled to represent their relative temporal ordering. In this example, C and D split from each other (node 3) before A and B split (node

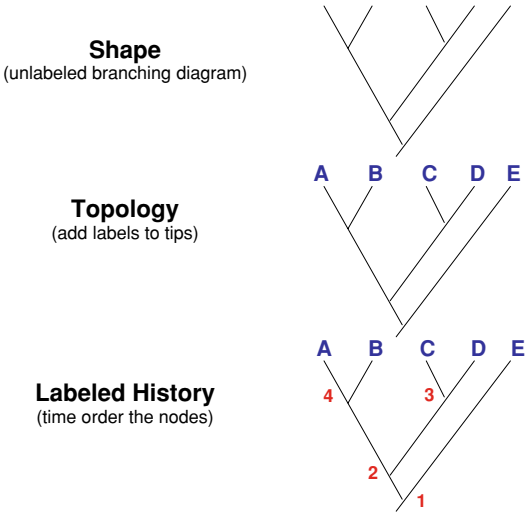

Fig. 1 Three different aspects of a tree

4). If "3" and "4" were reversed, we would have the same topology but a different labeled history.

Some shapes are consistent with more topologies than others; if each topology has an equal prior, not all shapes will be equally probable. Similarly, some topologies are consistent with more labeled histories than others so assigning equal priors to all topologies means that not all shapes nor all labeled histories will be equally probable. This is apparently a problem since it would appear that we are ignorant with respect to each but yet we cannot model ignorance with respect to all three. However, I suggest that this way of thinking about ignorance is a mistake. We are not ignorant of *everything* regarding topology and shape—after all, we know the logical facts that connect them. The kind of ignorance we ought to be modeling does not always lead to uniform priors. As an example of what I mean, I now turn to a recent example that purports to show that the use of priors in phylogenetics inevitably leads to bias results.

Priors on clades

Nearly every published paper using Bayesian methods uses a uniform prior distribution on tree topologies which assigns equal prior probability to each possible topology. Partly this is motivated by the simplicity of the proposal combined with its being the only distribution available (other than entering your own constraints for particular clades) in popular computer programs such as Mr. Bayes (Huelsenbeck and Ronquist 2001). And without careful examination, the proposal does seem sensible—after all, why should we have a prior preference for one topology over another when the topology itself is the primary object that we are trying to infer? In fact, by not using priors at all, if used as a guide to truth, Parsimony and Likelihood analysis are carried out in a way that effectively treat all topologies as equally

probable a priori. This fact has not been traditionally seen as biasing results in any way. But as Pickett and Randle (2005) (henceforth "P&R") point out, a uniform prior distribution on topologies implies a non-uniform distribution on the prior probabilities of clades—in particular, the probability that a particular group forms a clade depends on its size relative to the total number of taxa in the analysis. Smaller and larger groups have higher probabilities while middle-sized groups have the lowest probabilities. Figure 2 provides an example, when sets of different sizes are drawn from 50 taxa placed at the tips (or "leaves") of the tree.

While the particular values would change with a different number of taxa in the study, the shape of the curve will not. Any arbitrary group of taxa is a possible clade and P&R contend that all such groups regardless of their size should have the same prior probability of forming an actual clade in a given problem with a fixed group of taxa. Analyzing simulated data as well as data from seventeen published empirical studies, P&R argue that the use of the uniform distribution has biased the posterior probabilities in predictable ways, namely, that the very smallest and largest clades typically have the highest posteriors probabilities and the middle-sized clades have the lowest. This result corresponds to the prior distribution on clades imposed by setting a uniform prior over topologies. Several subsequent papers and books have cited this fact (e.g. Goloboff and Pol 2005 and Yang 2006) and different examples have been produced which lead to the same results. The authors agree that these facts lead to devastating conclusions for the Bayesian.

This line of thinking is based on misunderstanding what it is for the posterior to be biased and what the appropriate understanding of ignorance is. It is entirely proper for different sized clades to be more or less probable a priori since the appropriate understanding of a priori in this context builds in relevant background knowledge. P&R's claim that you can't have both uniform priors on topologies and on clades is correct; in fact, Velasco (2007) strengthens their proof by showing that on any probability distribution on trees (not just the uniform one) not all clades can

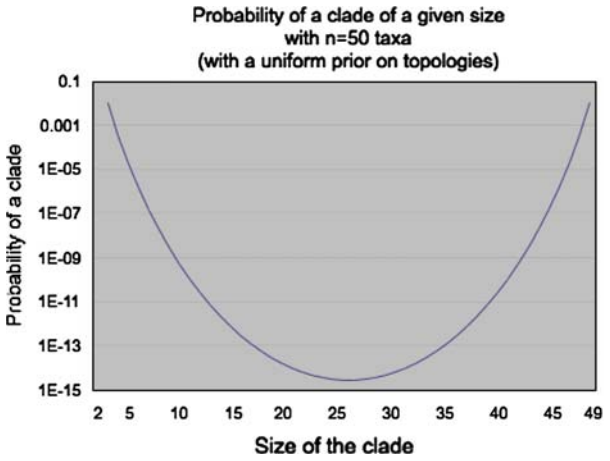

Fig. 2 A graph depicting the probability that a group of a given size forms a clade on a tree with 50 taxa when a uniform prior on topologies is used

be equally probable. There is nothing special about the uniform prior on topologies which conflicts with uniform priors on clades—uniform priors on clades is simply inconsistent. Having every possible clade be equally probable is not something that we could have even if it were desirable (which it isn't.) Once we see why this is so, it becomes easier to see what conclusions we should draw from it.

There is an easy explanation for why it is impossible for every possible clade to have an equal probability of forming an actual clade on the true tree. In a given problem with a fixed set of taxa, the probability that a group of a particular size forms a clade is just the expected number of clades of that size divided by the number of possible clades of that size. Lets use two specific sizes (clades of size 2 and 3) as examples to show that they can't be equally probable. The fact that not all of the probabilities can be equal can be deduced from the following two facts:

(1) Since smaller clades are nested inside larger ones, on any tree (and therefore on the true tree), there are at least as many actual clades of size two as there are of size three. Therefore, on any probability distribution over trees:

the expected number of clades of size $2 \geq$ the expected number of clades of size 3.

(2) When there are at least five leaf taxa:

the possible number of clades of size $2 <$ the possible number of clades of size 3.

Therefore, (when we have at least five taxa),

$$\frac{\text{the expected number of clades of size 2}}{\text{the possible number of clades of size 2}} \neq \frac{\text{the expected number of clades of size 3}}{\text{the possible number of clades of size 3}}$$

So not all possible clades of size two or three could be equally probable and a fortiori not all possible clades can be equally probable.

To determine the actual numerical probabilities, we need to know two things: the numbers of possible and actual clades of each size. The number of possible clades of size x is just the number of possible ways of choosing a group of size x from the collection of n taxa which is just n choose $x = \frac{n!}{x!(n-x)!}$. The number of actual clades of a given size will depend on the tree. A uniform distribution on tree topologies yields a particular distribution on the expected number of clades of any particular size first calculated in Brown (1994). The above facts are perhaps more easily appreciated by attending to the following graphs in Fig. 3:

The first graph plots how the size of a clade determines the number of possible clades of that size. I have used $n = 50$ taxa as an example, but the shape of the curve is the same for any number of taxa. Notice that the scale is logarithmic meaning that there are vastly more possible clades of size 25 than, say, size 10. The second graph plots how the size of a clade determines the expected number of clades of that size on the uniform distribution on topologies. Since the probability of a clade is just the expected number of clades of that size divided by the number of possible clades of that size (assuming all clades of the same size have the same probability), if the probability of a clade is to be the same for every size, these two curves must have the exact same shape (one should be the other multiplied by a constant—the probability). Notice that the "expected clades" curve is calculated under a uniform prior on topologies (as in

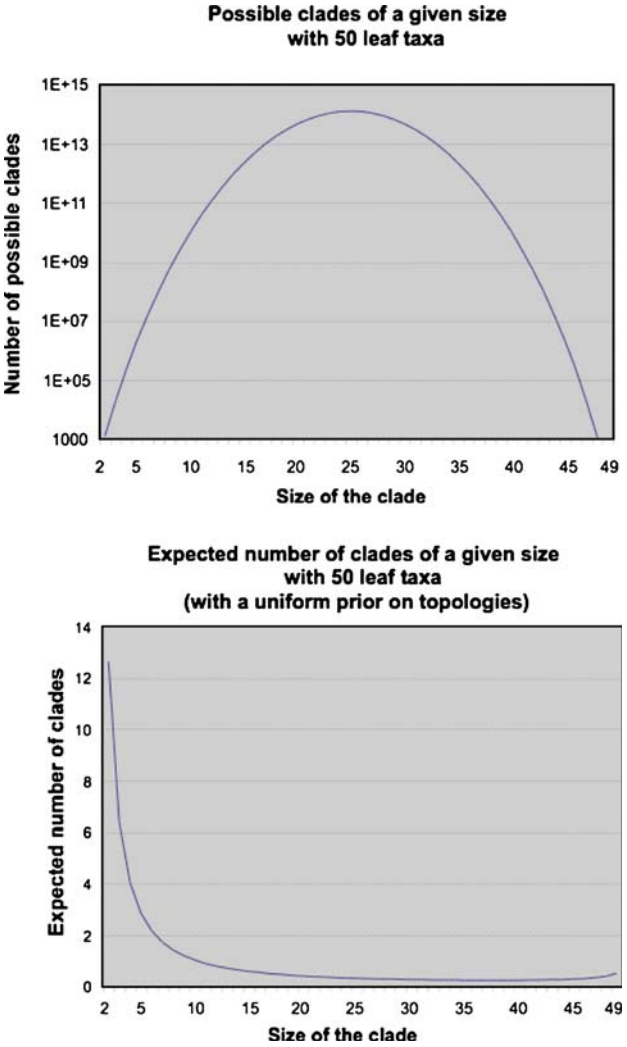

Fig. 3 Two graphs comparing the number of possible clades of a given size to the expected number of clades of that size. The expected number of clades (the value on the bottom) divided by the number of possible clades (the value on the top) is the probability that that group forms an actual clade. This figure is taken from Velasco (2007)

Picket and Randle 2005)—for other topology distributions the curve varies in shape slightly, but a few aspects remain constant, such as the fact that its peak must be at size 2. Since no distribution on trees gives it the same shape as the "possible clades" curve, the probabilities of all possible clades can never be identical. A formal proof of this fact is given in Velasco (2007).

So what should we make of this theorem? It might be thought that we have just shown that Bayesianism is a flawed methodology. After all, haven't we just shown that it is impossible to model ignorance with respect to clades since clades of

different sizes must have different probabilities? And isn't this obviously bad? As P&R put it,

> Few, if any, systematists believe a priori that the probability of monophyly has anything to do with the number of taxa hypothesized to be monophyletic. Certainly, the prior assertion that small clades and large clades are more probable than mid-sized clades lacks biological relevance. As such, a return to optimality per se is warranted. (Pickett and Randle 2005: p. 209)

P&R as well as Goloboff and Pol (2005) and Yang (2006) claim that uniform priors on topologies introduce a bias in favor of smaller and larger clades and against medium sized ones. We have just seen that this disparity in probabilities is guaranteed to occur regardless of our choice of priors on trees. Their choice of the word "bias" indicates that they think that this is a bad thing. P&R think this justifies abandoning Bayesianism as it is currently practiced and they suggest three alternative, incompatible methods for being Bayesian while attempting to artificially "correct" for this bias. However, their conclusion that using priors introduces an unacceptable bias into the problem rests on a mistake. We *want* the probabilities of clades to depend on their size. Artificially changing the posteriors or altering how we measure the strength of the evidence to correct for this would actually *introduce* bias. There *is* biological relevance to the fact that clades of size two should have higher priors than those of size three—we know from the way that clades are produced that clades with larger numbers of taxa have smaller clades within them. Basic mathematical facts combined with background biological facts indicate that we *should* believe that groups with more taxa are less likely to be clades. Claiming ignorance with regard to whether the true tree contains a particular clade of size two or whether that tree contains a particular clade of size three is like claiming ignorance with respect to whether some random integer is divisible by 2 or divisible by 4. Ignorance does not entail equally probable.

To head off a possible response, notice that the idea of clades nested within clades explains why smaller clades should be more probable, but this doesn't explain why larger clades also have higher priors. But this is not a problem. The high probability of very large clades is simply an artifact of the design of the problem. If our problem uses 10 taxa, for nine of them to form a clade, all it takes is for the tenth to be outside of the rest. However, that same group of nine taxa is much less likely to form a clade if the problem considered 50 taxa. Unlike a problem with 10 taxa, with 50 taxa, clades of size 8 are more probable than clades of size 9. The bias toward very large clades essentially comes from assuming that all taxa under consideration form a clade. Just as the conditional probability that A and B form a clade is relatively high given that A, B, and C do, the conditional probability of nine taxa forming a clade is high given that we are acting as if they are inside a clade of 10 taxa. This fact is more easily appreciated by recalling that "forming a clade" is only meaningful in the context of a particular problem. For a group to form a clade in a particular problem, the members of the group must be more closely related to each other than to any other taxa under consideration. Humans and Gorillas form a clade as long as Chimpanzees are not one of the taxa under study. There is nothing objectionable about this either.

This argument shows that the probability of a clade must depend on its size, but if we do not carefully formulate the question, there might appear to be obvious counterexamples. If we think of particular groups, it is tempting to conclude that P&R might be correct after all—for example, what should the prior probabilities of monophyly be for the following groups: apes, mammals, and vertebrates? According to the above reasoning, the prior on apes should be low, mammals extremely low, and vertebrates unbelievably tiny. But our actual confidence in the three groups doesn't appear to depend on size. So are P&R correct after all? No. There are several problems with the supposed analogy, but the major statistical error is that this is an instance of sampling bias. Ignore the fact that many systematists would simply define these groups in such a way as to guarantee that they are monophyletic and imagine that we are working with a more traditional definition based on characters—or think of "vertebrates" as rigidly designating some set of taxa which we currently believe are vertebrates. The sample is biased because we have selected clades that have a high *posterior* probability of being monophyletic and then we are asked to imagine what their *priors* should be. For example, they each have what appear to be uniquely derived characters. Of course clades of different sizes can have the same posterior probabilities. But this is not the claim. P&R are claiming that *before* we examine *arbitrary* groups of taxa that we known nothing about, we should be equally confident that they are monophyletic regardless of their size. But this is absurd. Imagine I assign each of 100 primate species a different number and then randomly select some of those numbers. What are the chances that the numbers I have selected pick out a monophyletic group? The chances will obviously vary with the number of taxa that I select. If I select two primate species at random, the odds that I have selected a monophyletic group are low, but they are vastly higher than the odds that I have selected a monophyletic group if I had selected fifty random species. Yet this is exactly analogous to the question under consideration. Size does matter.

Possible priors and the principle of indifference

The above argument shows that we have to be careful when we wish to model ignorance, but it does not tell us how we actually ought to do so. We need to further constraints to guide our priors. The above arguments only show that clades of different sizes should have different probabilities but it is clearly correct that all possible clades of size 2 should be equally probable, all possible clades of size 3 should be equally probable, etc. In other words, if we ask some question about a group of n taxa that are otherwise unknown to us, it shouldn't matter which n taxa we select. If we want to know the probability that A is closer to B than to C or that A and B coalesce in the past million years, it shouldn't matter which taxa A, B, and C represent. Distributions that satisfy this condition are called label-invariant. If we want to model ignorance with respect to the particular taxa we choose, we must use a label-invariant prior. While this is certainly helpful, it still leaves us with an infinite number of choices. For example, a uniform prior on topologies satisfies this condition, but so do many distributions that entail that some shape has probability

one. While these second types of distributions are certainly implausible, we can't rule them out simply on the basis of the condition that we must treat each taxon equally.

Although uniform priors on topologies are typically used, we have already seen that several authors believe that it leads to biased results that can be uncovered by examining other factors such as particular clades. While unequal priors on clades are not a good reason to give up uniform priors on topologies, perhaps looking elsewhere will provide just such a reason. For example, with four taxa there are 15 different topologies—12 have the pectinate A(B(C,D)) shape (this notation means that C and D form a clade which is nested inside B, C, and D which form a clade) while only three have the balanced (A,B), (C,D) shape where there are two clades of size 2. So uniform priors on trees introduces a skewed distribution on shapes. Is this acceptable? A traditional defense for uniform priors on topologies might appeal to the principle of indifference—when there is no epistemic reason to prefer one topology over another, they should all have equal priors. Of course most versions of the principle of indifference have well-known problems and typically lead to inconsistency (Joyce 2005), but there may be some less general principle which applies in this case that isn't problematic. But even a principle tailored specifically for phylogenetics is going to be question-begging in this context as the obvious response is that there is a reason to weight topologies differently—namely, some shapes are consistent with more topologies than others. If we believed that shapes should be equally probable, this (together with label invariance) would determine a particular distribution on topologies that favors topologies that are more balanced. In addition, we might also wish to assign equal probabilities to each labeled history. Each distribution is different so which distribution is to be preferred?

In other cases in science where we think that there is a good answer to this type of question, the correct prior is always determined by looking at the physical process that generates the values for the probabilities. In many cases, the process can vary. Selecting a day "at random" might yield a prior probability of 1/365 for any particular day being selected, but if the process of selection involves first selecting a month at random and then selecting a day within that month, the probabilities would be different. Regardless of the process, the point is that if we know the method of selection, then we can determine how to model ignorance. Assigning priors is problematic only in cases where we do not have an understanding of the underlying process.

In the phylogenetic case, the tree is a result of the biological process of common ancestry and descent with modification. We want to know the probability distribution that results when a tree is produced by this process. Trees are the result of the sequences passing down from organism to organism via reproduction on the branches and splitting at the nodes when the organism gives rise to multiple offspring which lead to different, extant taxa. A perfectly random branching process is captured by the Yule birth process in which particles reproduce with a constant probability of giving birth per particle per unit time so the Yule birth process seems the ideal place to start our investigation.

The Yule process

In 1924, G.U. Yule developed a statistical model to help explain why some genera have many more species than others (Yule 1925). The model was based on thinking of speciation as lineage splitting—one lineage gives birth to another without dying. In the simplest case, the idea is that we start with a common ancestor and then the probability of any particular lineage splitting in some small unit of time is the constant λdt. Two splitting events happen in the same time period with probability $o(dt)$. As time passes, there are more and more lineages present, each with the same probability of splitting until we reach the final result of n taxa. If at each slice of time, each existing lineage has an equal chance of splitting, we call the process a Yule pure birth process.

Another way to think about this process is by looking at the present and working backwards. The coalescent process imagines n gene sequences existing at the present. Then as we move back in time they will begin to coalesce. Each sequence has an equal probability of coalescing with any other particular sequence and then we go from n to $n-1$ sequences and repeat the process again. This process is obviously just the inverse of the birth process and so the same mathematical rules apply yielding the same probabilities for certain parameters such as shape and topology (Kingman 1982).

For our purposes, we want to know the probability of getting a particular tree as the result of a Yule process. The answer is that a Yule process produces each labeled history with equal probability (Edwards 1970). Thus the distribution that each labeled history should be equally probable a priori can be given a justification. The justification is not the one provided by the principle of indifference, which says "I can't think of a reason why one labeled history should be more probable than another." Rather, the justification is that if the evolution of different taxa is the result of random lineage splitting, then for n random taxa, the probability that they form a particular tree topology is proportional to the number of labeled histories that are consistent with that topology.

One might be worried that we are ignoring extinction. We could easily add another parameter μ where the probability of any particular lineage going extinct is μdt. This is known as a birth–death process. Importantly, it leads to exactly the same distribution on tree topologies. As long as the extinction happens randomly across lineages, the prior probabilities will be the same (Thompson 1975). The pure birth process, the birth–death process, and the coalescent process all lead to exactly the same distribution—all labeled histories are equally probable.

The idea that the Yule process represents a "randomly branching tree" is not new in the mathematical literature (Harding 1971; Aldous 2001). This idea is also fairly standard in the biological literature. The Yule birth process (or more typically a birth–death process) is widely used to study macroevolutionary trends. For example, the discovery of broad-scale biogeographical patterns and the detection of differences in speciation or extinction rates across lineages are standardly thought to depend on comparing the accepted phylogeny to a null model of random branching. The null model typically used for such comparisons is the Yule model (e.g. Mooers and Heard 1997 and many of the very large number of references therein). The Yule process is

also widely used to study microevolutionary processes. The standard method of studying intraspecies diversity will use a coalescent process to build gene genealogies which are essential to testing hypotheses such as those concerning the strength of selection at a particular site or testing the amount of gene flow between distinct populations (Halliburton 2004, Hein et al. 2005). Despite the near-universal acceptance of the Yule process being the underlying physical process for common descent and therefore the production of phylogenetic trees, biologists virtually never take this process into account when actually constructing trees! (For exceptions, see Rannala and Yang 1996; Yang and Rannala 1997). The use of prior probabilities in Bayesian phylogenetics makes thinking about the probabilities of trees unavoidable, but the idea of a null model for a tree is required even in methods that do not specifically attempt to use a prior probability distribution. As we shall see later, ignoring these facts can lead to mistaken conclusions not only in constructing trees which are best supported by the evidence, but also when we attempt to use those trees to make further inferences about the evolutionary process. Theoretically, it is well motivated to start insisting on such a change in methodology, but I now turn to the question of what, if any, consequences making such a change will actually have.

We have already noted that the "Yule distribution"—the probability distribution of trees induced by a Yule process—is a different distribution than the uniform distribution. With four taxa, there are 15 topologies and 18 labeled histories. Since some topologies (those with the pectinate, asymmetric shape) are consistent with only one labeled history and some are consistent with two (the balanced shape), the priors shift from 1/15 on the uniform topology to either 1/18 or 2/18 depending on whether we are looking at the asymmetric or the balanced tree. In general, more asymmetric topologies will have their prior probabilities lowered and more balanced trees will have theirs raised. There are many ways that the overall balance of a tree could be measured (Mooers and Heard 1997), but certainly in the clear cases, the result of a Yule process is that a tree that is more balanced will be consistent with more labeled histories (there are more pathways to reach it) and thus is more probable than any particular unbalanced tree.

The idea that balanced trees are consistent with more labeled histories and therefore are more probable than unbalanced trees is exactly analogous to the claim that if we flip a fair coin 100 times, we are more likely to get 50 heads than some other number of heads. If the coin is fair, each particular sequence of heads and tails is equally probable. Since 0 heads is only consistent with one sequence, it is far less probable than 50 heads which is consistent with $\approx 10^{29}$ sequences. An important side note is that we should not conclude that the Yule process will probably result in a balanced tree. The appropriate conclusion to draw is that $\Pr(T_1|T_1$ is balanced$) > \Pr(T_2|T_2$ is unbalanced) not that Pr(Tree will be balanced) > Pr(Tree will be unbalanced). Far fewer tree topologies are balanced than unbalanced, so even though each has a higher probability than those that are unbalanced, the unconditional probability that a tree is balanced is still relatively low.

So we know that if we replace uniform priors with Yule priors, the prior probabilities of unbalanced trees will go down while those of balanced trees will go up. But does this difference really matter to their posterior probabilities? This will depend on the particular problem. Problems can be constructed where the priors

matter. Problems can be constructed where they don't. With enough data, the likelihoods of the various trees will completely swamp differences in the priors between trees. But how much data is required and just how much this difference in priors matters in realistic cases is something that will require careful quantitative investigation.

It is widely known that the number of possible trees with n taxa $= 2n-3!! = \prod_{2}^{n} 2n-3$ (Felsenstein 2004). Steel and McKenzie (2001) provide a recursive algorithm for calculating the number of labeled histories consistent with a particular topology. For each vertex v (node) let $\delta(v)$ be the number vertices that are its descendants (including itself). Note that this is the same as the number of taxa in the subtree formed by that node minus 1. Now, the number of labeled histories consistent with any particular tree topology $= \frac{(n-1)!}{\prod_v \delta(v)}$. For example, the number of labeled histories consistent with the perfectly balanced four taxa tree $= \frac{(4-1)!}{3\times 1\times 1} = 2$. Combined with the formula for the total number of possible labeled histories for n taxa: $\frac{n!(n-1)!}{2^{n-1}}$ (Edwards 1970) we now can calculate the prior probability of any particular tree under the Yule model. To see directly whether this will affect the posterior probability of any individual tree, we would need to calculate the normalizing constant—Pr(Data)—which we can't do. Another option is to run an MCMC on some particular data set with uniform priors as is typically done and then run the MCMC on the same data set with Yule priors instead of uniform priors and simply check for differences in the results. This method requires us to recalculate the entire posterior distribution just to see if there will be any significant difference in the posteriors of particular trees. But there is another method that can tell us at least some of what we want to know.

Imagine that we perform the calculations with uniform priors and get the result that T_1 has a higher posterior probability than T_2. How probable is it that the results would be different if we used Yule priors instead? For the order to switch, the ratio of the posteriors would have to switch from being greater than 1 to being less than 1. By Bayes Theorem, the ratio of the posterior probabilities is equal to the ratio of the priors times the ratio of the likelihoods:

$$\text{Bayes Theorem(Odds} - \text{Ratio form)}\ \frac{\Pr(T_1|D)}{\Pr(T_2|D)} = \frac{\Pr(D|T_1)}{\Pr(D|T_2)} \times \frac{\Pr(T_1)}{\Pr(T_2)}$$

Since the likelihoods themselves will not change, we can directly calculate the effect of changing the priors. Since the old prior ratio was 1:1, if we want to switch the ordering on trees, we need the new prior ratio to be greater than the reciprocal of the likelihood ratio. So how large is the ratio of the priors? In the four taxa case, the most balanced to least balanced ratio is only 2:1. But like all other effects that depend on the number of possible trees, this is going to increase combinatorially. To give an extreme example, the perfectly balanced tree with 64 taxa (it splits into two subtrees of 32, each of those splits into two subtrees of 16, etc.) is consistent with $\frac{(63-1)!}{63\times 31^2\times 15^4\times 7^8\times 3^{16}} \approx 2.61 \times 10^{63}$ labeled histories. Since the maximally unbalanced tree, which has splits of 1:63 then 1:62, then 1:61, etc., is consistent with only one labeled history, 10^{63} is also the ratio of the prior probabilities of the trees. For $n = 128$, this ratio rises to $\approx 4.1 \times 10^{163}$. While the likelihood ratio can easily be

greater than this for several thousand independent sites, these massive numbers should certainly give pause to anyone who claims that using different priors would not make any difference. Certainly they will make some difference to the overall posterior distribution. Exactly how much difference they will make will depend on the particular case and general conclusions will require further study. Regardless of how often changing the priors dramatically affects the posteriors, we have at least the beginnings of an explanation of why we should use one prior probability distribution rather than another.

An important feature of this discussion is that it is not essential to this argument that the Yule process perfectly captures the causal process by which evolutionary trees are produced. In particular, it is clearly unrealistic that at a given time, each extant lineage has the same probability of splitting. This paper takes the important first step of using priors that at least attempt to be biologically relevant. There is no known biological process that would lead to a uniform distribution on topologies. As such, there can be no justification for using these priors. In addition, we can think of the Yule model with no other effects as the simplest among a whole class of branching models which might be used to generate priors. Further biological investigation can help us improve our branching models and thus improve the accuracy of our prior probabilities. The Yule model, not the uniform model, will form the essential backbone of any such future investigations.

In addition, it needs to be pointed out that Yule process models how all of the tips resulting from a common ancestor are expected to be related. This means that taxa sampling will severely affect the model. If we are examining all of the tips that have descended from some ancestor, then the Yule process will be adequate. Similarly, if we randomly sample tips, this will not affect the distribution. However, if we use some non-random method—such as sampling two organisms from each species under investigation—it is easy to see that we should expect the tree to have a different shape. A clear instance of this is the use of outgroups which guarantees that the tree will be very imbalanced at the root—something that is improbable on a Yule model. Depending on how we sample, we might be able to correct this bias (in the two above cases, this is easy). But often times, we sample taxa non-randomly, but not by answer process which we can build into our model. But this entire discussion simply reinforces the point that it is essential to realize that not all trees are equally probable a priori and that this fact can affect our results when it is ignored.

The base-rate fallacy

Since using different priors on topologies could lead to different results in particular phylogenetic studies, it could also lead to different results in studies that use phylogenies to make further biological inferences. This is particularly relevant since I have argued that phylogenies produced without attending to the Yule model are not just in error, but that they are in error in a particular way. I will now examine an example of this error. The base-rate fallacy is a common mistake made in everyday reasoning. That mistake is to ignore the base-rate, or prior probability, of events

when making inferences. Here is a standard example of medical testing often found in the literature.

We take a random person in the United States and administer an HIV test which is accurate in 95% of all cases. The test shows up positive. The proper conclusion to draw is that this person probably does not have HIV. We can reach this conclusion by noting that the prior probability that they have HIV can be approximated by the base-rate of HIV in the population. In 2005, the CDC estimated that there are about 438,000 people living with HIV out of over 300 million in the US and its dependencies giving us a prior probability of .00146 (Centers for Disease Control and Prevention 2005). By Bayes' Theorem,

$$\begin{aligned}\Pr(\text{HIV}|+\text{test}) &= \frac{\Pr(+\text{test}|\text{HIV}) \times \Pr(\text{HIV})}{\Pr(+\text{test})} \\ &\approx \frac{0.95 \times 0.00146}{(0.95 \times 0.00146) + (0.05 \times 0.99854)} \approx 0.027\end{aligned}$$

In other words, there is only 2.7% chance that this person actually has HIV. The explanation is simple—5% of the people who don't have HIV will get a positive test result and this group is much larger than the group of individuals who actually have HIV. Certainly, the positive test result raises the probability that this person has HIV. In fact, it raises it by a factor of almost 20—but this only raises the probability from 0.15% to about 2.7%. In general, if the false-positive rate is higher than the base-rate, then there will be a less than 50% chance that they actually have the disease in question. If we look at only the likelihood of having HIV (0.95) and ignore the base-rate, we are committing the base-rate fallacy. While ignoring very skewed base-rates is particularly bad, it is important to realize that it is always an error to ignore base-rates regardless of what they are.

While the debate over how to assign prior probabilities might be seen as a debate internal to Bayesianism, understanding the underlying process that generates phylogenies is essential to making correct inferences regardless of methodology. If the Yule process truly underlies the production of phylogenetic trees, then to ignore it as Parsimony and Maximum Likelihood methods do is akin to committing the base-rate fallacy. Similarly, using a prior distribution, but using the wrong one such as when the uniform distribution is used, leads to the wrong conclusions. If we are lucky enough to have data which show a very strong signal for particular clades, the data will overcome the bias that these mistakes introduce, but this will certainly not be the case in every instance.

As a practical example of this error, there is a large literature on how to make inferences based on the shapes of trees and the consensus in the field is that trees (based on published phylogenies) seem to be more asymmetric than we would expect by chance (Huelsenbeck and Kirkpatrick 1996; Mooers and Heard 1997). This has lead systematists to conclude, among other things, that effects such as clade selection are prevalent and that phylogenies are not just the result of random branching. What we would expect "by chance" is (appropriately) determined by examining a Yule distribution, but the published phylogenies typically do not use prior probabilities and if they do, they use a uniform distribution which is skewed

toward asymmetry relative to the Yule distribution. Since the Yule process represents random branching, the use of uniform priors on topologies (or the use of no priors at all) have biased the results in favor of more asymmetric trees. In other words, we should expect the result that published phylogenies are more asymmetric than expected by the Yule process. By thinking about the base-rate fallacy, we can see that if our data leads us to conclude that a tree is unbalanced, this might be a case where it is more probable that the tree is more balanced, but that the data is misleading. Of course not every case will be a false positive, effects such as clade selection and taxa sampling bias certainly do affect inferred tree shapes, but the above analysis points to an important project that still needs to be done – reexamining the data on tree shapes to see just how much of the apparent difference between actual history and randomly produced trees is simply an artifact of getting the history wrong in the first place due to ignoring the process by which trees are generated.

Acknowledgements I am grateful to David Baum, Matt Haber, James Justus, Bret Larget, Greg Novack, and especially Elliott Sober for their support and many helpful comments on earlier drafts of this paper. Thanks also to an anonymous reviewer who made several helpful comments which improved the presentation of this paper.

References

Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Stat Sci 16(1):23–34

Brown JKM (1994) Probabilities of evolutionary trees. Syst Biol 43(1):78–91

Centers for Disease Control and Prevention (2006) HIV/AIDS Surveillance Report, 2005, vol 17. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, pp 1–54. Also available at http://www.cdc.gov/hiv/topics/surveillance/resources/reports/2005report/

Edwards AWF (1970) Estimation of the branch points of a branching diffusion process. J R Stat Soc B (Methodological) 32(2):155–174

Farris JS (1983) The logical basis of phylogenetic analysis. Adv Cladistics 2:7–36

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, Mass

Goloboff PA, Pol D (2005) Parsimony and Bayesian phylogenetics. In: Victor AA (ed) Parsimony, phylogeny and genomics. Oxford University Press, Oxford, pp 148–159

Haber MH (2005) On probability and systematics: possibility, probability, and phylogenetic inference. Syst Biol 54(5):831–841

Halliburton R (2004) Introduction to population genetics. Pearson/Prentice Hall, Upper Saddle River

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab 3(1):44–77

Hein J, Schierup M, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, Oxford

Howson C, Urbach P (2005) Scientific reasoning: the Bayesian approach, 3rd edn. Open Court, La Salle

Huelsenbeck JP, Kirkpatrick M (1996) Do phylogenetic methods produce trees with biased shapes? Evolution 50(4):1418–1424

Huelsenbeck JP, Ronquist F (2001) MR BAYES: Bayesian inference of phylogenetic trees. Bioinformatics (Oxford) 17(8):754–755

Huelsenbeck JP, Ronquist F (2005) Bayesian analysis of molecular evolution using Mr. Bayes. In: Nielson R (ed) Statistical methods in molecular evolution. Springer, New York, pp 183–232

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science (Washington DC) 294(5550):2310–2314

Joyce J (2005) How probabilities reflect evidence. Philos Perspect 19:153

Kadane JB (2006) Is "objective Bayesian analysis" objective, Bayesian, or wise? (comment on articles by Berger and by Goldstein). Bayesian Anal 1(3):433–436

Kingman JFC (1982) The coalescent. Stochastic Process Appl 13(3):235–248
Kluge AG (2005) What is the rationale for 'Ockham's razor' (a.k.a. parsimony) in phylogenetic inference? In: Albert VA (ed) Parsimony, phylogeny and genomics. Oxford University Press, Oxford, pp 15–42
Larget B (2005) Introduction to Markov Chain Monte Carlo methods in molecular evolution. In: Nielson R (ed) Statistical methods in molecular evolution. Springer, New York, pp 44–61
Larget B, Simon DL (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16(6):750–759
Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM (2002) Molecular evidence of HIV-1 transmission in a criminal case. Proc Nat Acad Sci 99(22):14292–14297
Mooers AO, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 72(1):31–54
Pickett KM, Randle CP (2005) Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. Mol Phylogenet Evol 34(1):203–211
Pitnick S, Jones KE, Wilkinson GS (2006) Mating system and brain size in bats. Proc R Soc Biol Sci B 273(1587):719–724
Poux C, Madsen O, Marquard E, Vieites DR, de Jong WW, Vences M (2005) Asynchronous colonization of Madagascar by the four endemic clades of primates, tenrecs, carnivores, and rodents as inferred from nuclear genes. Syst Biol 54(5):719–730
Randle CP, Mort ME, Crawford DJ (2005) Bayesian inference of phylogenetics revisited: developments and concerns. Taxon 54(1):9–15
Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43(3):304–311
Semple C, Steel M (2003). Phylogenetics. Oxford University Press, Oxford
Siddall ME, Kluge AG (1997). Probabilism and phylogenetic inference. Cladistics 13(4):313–336
Sober E (1988). Reconstructing the past. Parsimony, evolution, and inference. MIT Press, Cambridge
Steel M, McKenzie A (2001). Properties of phylogenetic trees generated by yule-type speciation models. Math Biosci 170(1):91–112
Thompson EA (1975) Human evolutionary trees. Cambridge University Press, Cambridge
Velasco JD (2007) Why non-uniform priors on clades are both unavoidable and unobjectionable. Mol Phylogenet Evol 45:748–749
Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford
Yang Z, Rannala B (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol Biol Evol 14(7):717–724
Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philos Trans R Soc Lond B 213:21–87