Prior Probabilities of Phylogenetic Trees

draft as of 5/28/07

Joel Velasco

Abstract: Given that modern computing algorithms and computing power allow the calculation of posterior probabilities on phylogenetic trees with a high degree of accuracy, the main objection to the use of Bayesian methods in phylogenetics is that the posterior probability is not a good optimality criterion for trees. The theoretical problem mentioned most often in the biological literature is "the problem of the priors" - how to assign prior probabilities to various tree hypotheses. In the rare case that a solution is attempted, the proposals typically take one of two forms: 1) The particular priors that are used are unimportant because they have negligible effect on the posteriors or 2) Uninformative priors can be used in order to ensure that the posteriors are not biased. In this paper I show that the priors used can affect the results dramatically and that the so-called "uninformative" priors typically used do in fact bias results significantly. Thus neither proposed solution solves the problem of priors. After showing that a recent objection – that an appropriate assignment of priors is impossible – is based on a misunderstanding of what ignorance and bias are, I consider different methods of assigning prior probabilities to trees. I argue that in the general case, priors must be derived from a model of how outcomes are produced. In the specific case at hand, this means that priors need to be derived from an understanding of how distinct taxa have evolved. The appropriate evolutionary model is that of lineage splitting, which is abstractly 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.

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 mammals arrived in Madagascar (multiple separate rafting events rather than a land bridge – Poux et al 2005) all require a 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 a 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 authors, 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 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 alone is called the tree topology and is generally the primary object of inference for the systematist.

Bayesian phylogenetics

Maximum Parsimony and Maximum Likelihood are two families of methods that have dominated phylogenetics discussion for the past twenty 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 1997, Huelsenbeck et al 2001). This new trend 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 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:

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

Pr(Tree), called the prior probability of the tree in question, is determined from a probability distribution over all possible trees given before the data are examined. The probability of the data is a normalizing constant simply used to make sure that the posteriors 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

$$\Pr(\text{Data}) = \sum\nolimits_{\text{Ti}} \Pr(\text{Data} \mid \text{Tree}) \times \Pr(\text{T}_{\text{i}})$$

But we aren't done "simplifying" yet. Pr(Data | Tree) is called the likelihood of the tree, but it cannot be directly calculated as 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 along with some model of evolution which will contain its own parameters to be estimated such as the nucleotide substitution rates.

The Bayesian method for dealing with these nuisance parameters 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 & 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:

$$\mathrm{Pr}(\mathrm{Data} \,|\, \mathrm{T_i}) = \int_v \int_\theta \mathrm{Pr}(\mathrm{Data} \mid \mathrm{T_i}, v, \theta) \times \mathrm{Pr}(v, \theta)\, dv\, d\theta$$

Substitution in both the numerator and denominator yields an extremely unpleasant looking formula:

$$\mathrm{Pr}(\mathrm{T_i} \,|\, \mathrm{Data}) = \frac{\int_v \int_\theta \mathrm{Pr}(\mathrm{Data} \mid \mathrm{T_i}, v, \theta) \times \mathrm{Pr}(v, \theta)\, dv\, d\theta \times \mathrm{Pr}(\mathrm{T_i})}{\sum_{T_i} \int_v \int_\theta \mathrm{Pr}(\mathrm{Data} \mid \mathrm{T_i}, v, \theta) \times \mathrm{Pr}(v, \theta)\, dv\, d\theta \times \mathrm{Pr}(\mathrm{T_i})}$$

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, 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 some other parameter 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 entirely. However, there is reason to be hopeful here. Computational methods for numerically estimating multi-dimensional integrals are in fact quite advanced. The standard idea is to use Markov Chain Monte Carlo (MCMC) methods to estimate the posterior distribution. For and introduction and review of these methods see Larget and Simon (1999) and Larget (2005).

In fact, it may be surprising that it is the practical aspects of Bayesian phylogenetics that have won many of its converts. MCMC methods are possible because nuisance parameters are integrated out rather than being point estimated such as in frequentist methods like Maximum Likelihood. Because of this, Bayesian inference can often be several orders of magnitude faster than a heuristic Maximum Likelihood search. These differences are then magnified again when bootstrap replicates are generated to test the reliability of the Likelihood inference. (Larget and Simon 1999) This practical aspect can make the difference between a problem being computationally feasible in a reasonable amount of time and not being computable in the scientist’s lifetime. In addition, free computer programs such as Mr. Bayes (Huelsenbeck and Ronquist 2001) have now been around long enough for many phylogeneticists to become familiar and comfortable with them, causing an explosion in the use of Bayesian methods.

While there is certainly much more that needs to be discussed regarding the practical issues surrounding Bayesian phylogenetics, it is to the theoretical issues that I now turn. The question here is whether posterior probabilities are in fact the quantities that we want to infer rather than the Parsimony score, the frequentist-interpreted Likelihood score, or some other quantity. It is precisely in this theoretical realm that I will argue that Bayesian methods are clearly superior.

The theoretical justification for Bayesian inferences is that all inferences about the values of the parameters should be based on the joint posterior distribution. For example, the tree topology which is best supported after examining the data is the tree that has the highest posterior probability. In fact, its support is measured exactly by its posterior probability and the posterior probability represents exactly what we want to know – the probability that the tree is true given our data and the model of evolution used. Other facts about the problem, such as which tree would require the least amount of evolutionary change, which is what the Parsimony score captures, is of significant interest only in so far as it is a reliable guide to which tree is probably true.

In addition, we can judge the strength of evidence for other parameters at the same time without needing to reanalyze the data using different methods. The support for a particular clade is its posterior probability – that is, the sum of the posterior probabilities of all trees which contain that clade. The probability that the transition:transversion rate is greater than 2:1, the probability that two particular sequences have coalesced in the last one million years, and the probability that sequence A is more closely related to B than to C are similarly interpreted using the posterior distribution. None of these questions are easily analyzed with other methods. While particular tests have been developed (Bootstrap values or decay indexes for clades, Kashino-Hasegawa tests for more general tree hypotheses, and many more – 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 theoretical justification.

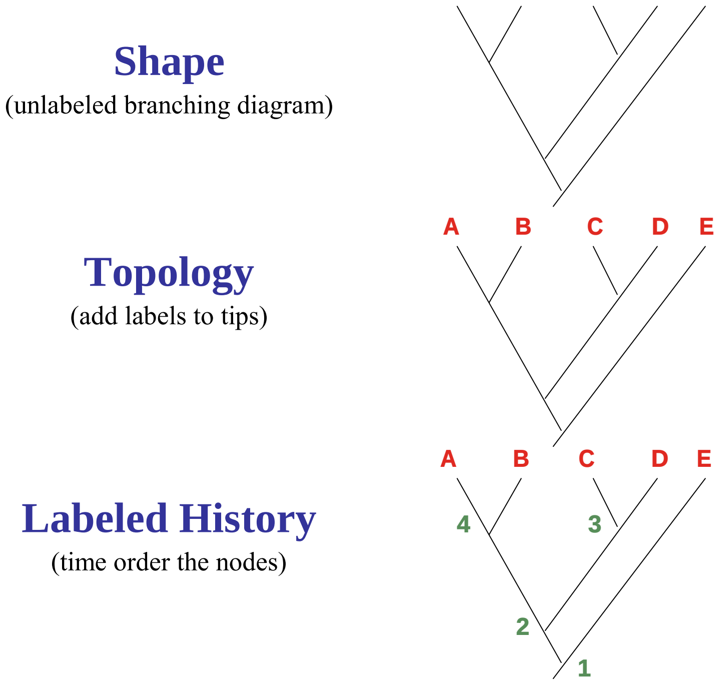
While a theoretical justification can be constructed for using posterior probabilities to guide our inferences, there are a few reasons why one might object. Some systematists believe that probabilities and perhaps even all statistical methods simply cannot be used to make inferences concerning a particular 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). 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 (for example, Huelsenbeck and Ronquist (2005)) and I will not discuss them further.

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 hypothesis under test. As Felsenstein, a strong advocate of Likelihood 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." While there would of course still be controversy, his point is really that to the statistically minded theoretician, there shouldn't be. In any case, specifying prior probabilities does seem to be a major point of contention in phylogenetics.

While a subjective Bayesian may respond that prior probabilities ought to simply 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. After all, we want results that ought to be taken seriously by a wide range of scientists. This is one of the goals of so-called "objective Bayesianism" (perhaps better called "Interpersonal Bayesianism" – Kadane 1996). Of course the systematist carrying out the inferential work is almost certain to have lots of evidence which precludes total ignorance about the underlying tree, but the interpersonal Bayesian ideal is to not build such evidence into the priors but rather, to operate "as if" we were ignorant. In the particular case at hand, we may want to know what the DNA sequences in our data alone ought to lead some impartial observer to believe even though the systematist herself knows quite a bit more. Of course "alone" is a tricky word here. It is quite proper to use logical and mathematical knowledge for inferences. And of course it would be absurd to deny the relevance of "independent" biological knowledge such as the fact that DNA is made of only four types of nucleic acids and that gene sequences are transmitted from one generation to the next via reproduction. These facts and many others that are implicitly relied on are simply part of the background knowledge agreed upon by all in the context of the problem. What seems inappropriate is to use facts that differ among the taxa in question – like the fact that certain taxa live spatially near to each other or share a large number of morphological characteristics. Of course taking these facts into consideration should be part of the overall evidence that we have for particular trees, but for our purposes, it is inappropriate to attempt to build these facts into a prior distribution. If we wanted to know what we should believe based on the molecular and the geographical evidence combined, both would be part of our data and for that we need some further procedure for combining different types of data (for which Bayesian analysis is again ideally suited.)

As long as we have a proper understanding of ignorance, it would appear 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, its shape, 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. However, there are many different ways of conceiving of a tree. The shape of the tree refers to the branching diagram with the labels at the tips removed. 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" – Semple and Steel (2003)) refers to the topology plus a temporal ordering of the nodes. These differences will become much more important later so I will display each of them here:

FIGURE 1



Shape

(unlabeled branching diagram)

Topology

(add labels to tips)

Labeled History

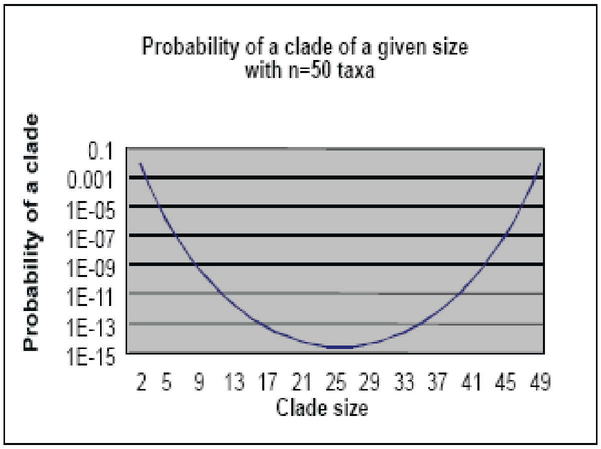
(time order the nodes)

Since some shapes are consistent with more topologies than others; if each topology has an equal prior, not all shapes will. Similarly, some topologies are consistent with more labeled histories than others. This is supposed to be a problem since it would appear that we are ignorant with respect to each but that 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 particular debate that has sprung up recently regarding the use of priors for phylogenetics.

Priors on clades

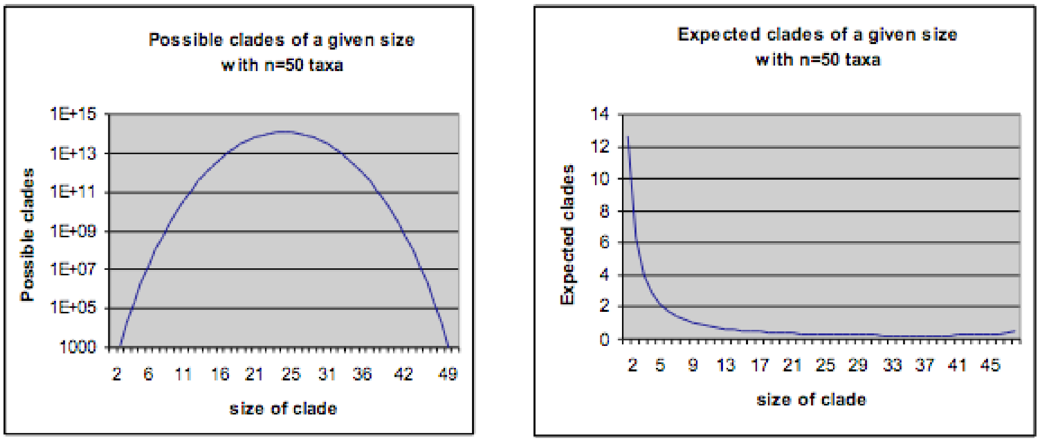
Nearly every published paper using Bayesian methods uses a uniform prior distribution on tree topologies. Partly this is motivated by the simplicity of the proposal combined with its being the only distribution available (other than entering your own constraints for particular clades) in popular programs such as Mr. Bayes. 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, 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 in 2005, Pickett and Randle (henceforth "P&R") produced a paper pointing out the simple fact that a uniform 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. As an example, here is the probability graph for *n*=50 taxa:

FIGURE 2



While the particular numbers would change with a different number of taxa in the study, the shape of the curve will not. There is an easy explanation for this. 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 trees yields a particular distribution on the expected number of clades of any particular size first calculated in Brown (1994). Here is a side-by-side comparison of the number of possible clades of a given size versus the expected number of clades of that given size using the uniform distribution on trees both using *n*=50:

FIGURE 3



Notice that in the first graph, the scale is logarithmic meaning that the number of possible clades of size 25 is vastly more than say, the number of possible clades of size 10. Assuming that all clades of the same size have the same probability, then that probability is just the expected number of clades of that size divided by the possible number of clades. Since there are vastly more possible middle-sized clades than larger or smaller ones, each one has to have a smaller probability. For all possible clades to have the same probability, the two above graphs would need to have exactly the same shape (the "expected" graph would be the "possible" graph multiplied by a constant factor – the probability).

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 uniform priors over topologies. Several papers and books since have already cited this fact (e.g. Goloboff and Pol 2005 and Yang 2006) and different examples have been produced which all apparently lead to the same conclusion. Each author agrees that these facts lead to devastating conclusions for the Bayesian.

I will argue that 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 apriori in this context builds in background knowledge which is relevant in this case. P&R's proof that you can't have both uniform priors on topologies and on clades is correct; in fact, I strengthen their proof by simply showing that on any probability distribution on trees (not just the uniform one) not all clades can be equally probable. Uniform priors across possible clades is not something that we could have even if it were desirable (which it isn't.) Once we see why this is the case, it becomes easier to see what conclusions we should draw from it.

The proof that there is no probability distribution that assigns equal probabilities to every possible clade is actually quite simple. This result is easily seen to follow from two elementary facts: 1) 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 and 2) There are many more possible clades of size three than of size two. So not all possible clades of size two and three could be equally probable and *a fortiori* not all clades can be equally probable.

Here is a formal version of the proof which holds for any number of taxa n>4. For n taxa, there is a total of n choose 2 = $\frac{n!}{2!(n-2)!}$ possible clades of size two while there are n choose 3 = $\frac{n!}{3!(n-3)!}$ possible clades of size three. Therefore, for n>5, there are more possible clades of size three than of size two. If each possible clade of size three or size two had the same probability of being an actual clade there would be more expected clades of size three than of size two which is impossible on any bifurcating tree and therefore on any distribution over trees (since each clade of size three entails the existence of a distinct clade of size two.)

For the special case of n=5, there are 10 possible clades of size two and 10 of size three. If they were each assigned equal probability, the expected number of clades of size two and size three would have to be equal. This is only possible if the only tree topologies with positive probability have the pectinate shape, which has nested clades of size 2, 3, and 4. On this distribution, there are equally many clades of size two as size three, but there are also equally many clades of size four. Since there are only 5 possible clades of size four, on this distribution they cannot have the same probability as the clades of size two and three.

This simple proof shows that as long as we are dealing with more than 4 taxa, we have to accept that not all possible clades can be equally probable. P&R would undoubtedly see this as a crushing blow to Bayesians. After all, haven't we just shown that it is impossible to model ignorance with respect to clades? As they 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, 2005:209)

It is unclear exactly what they mean by a "return to optimality" but presumably this means returning to views that don't involve a posterior distribution and rather, use a particular optimality criterion such as Maximum Parsimony or Maximum Likelihood. However, "returning to optimality" in this sense has nothing to do with solving the problem and in fact introduces problems of its own. P&R conclude that if we must be Bayesians, we should use the tree which has the highest posterior probability and base our inferences on that tree rather than using the posterior probabilities of particular clades. But there are deep and unsolvable problems with this suggestion. The full tree is essentially a very large conjunction (Clade 1 & Clade 2 & … Clade n-2) and as such can, and typically does, have a very low posterior probability so we would apparently be concluding that no clades are well-supported. Second, even if it is supported to some degree, this commits the fallacy of thinking that since a large conjunction is the best supported, each conjunct must be. But this is wrong. For example, consider the set of all clubs and spades in a standard 52 card deck and replace each Ace with an Ace of Hearts before drawing randomly from the deck. The single card with the highest posterior probability is an Ace of Hearts with P = 2/26. But we should not conclude from this that we will probably draw an Ace of Hearts, nor should we should conclude that aces are better supported than kings or that hearts are the better supported than or spades. Bayesian posterior probabilities for such things as individual clades use information contained in the full posterior distribution including, essentially, trees that are sub-optimal. This is a strength of the method and ignoring it leads to mistaken inferences.

While focusing exclusively on the optimal tree would be a mistake, we still have the apparent problem of bias. In fact, P&R suggest two other incompatible solutions besides looking just at the optimal tree. A second suggestion is to use Bayes factors instead of posterior probabilities. A "Bayes factor" is calculated in the following way:

$$\text{Bayes factor} = \frac{\text{posterior}/(1-\text{posterior})}{\text{prior}/(1-\text{prior})}$$

Bayes factors are designed to measure the *change* from the prior to posterior. Perhaps they measure something like the strength of the evidence that the data alone provide, but they certainly don't measure the confidence we ought to have in a hypothesis after seeing all of the data, which is what we are after. That is precisely what the posterior probability measures. A third suggestion of theirs is to introduce a "correction factor" to artificially correct the bias. As we can now see, it would be impossible to completely correct this bias, but we could certainly go some way by artificially inflating the posteriors of medium sized clades and deflating those of others.

P&R as well as Goloboff and Pol (2005) and Yang (2006) all claim that uniform priors on topologies introduce a bias in favor of smaller and larger clades and against medium sized ones. Their choice of the word "bias" indicates that they think that this is a bad thing However, this conclusion rests on a mistake. We *want* the probabilities of clades to depend on their size. 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 large numbers of taxa have many smaller clades within them. Basic mathematical facts combined with these background biological facts indicate that we *should* believe that groups with more taxa are less likely to be clades.

While it is true that middle-sized clades tend to have lower posterior probabilities, this is as it should be. It requires much stronger evidence to achieve an equal level of confidence that a particular group forms a clade if this group is large. This is not because of any bias – at least not if "bias" has its typical negative connotation. Rather, it is the predictable result of the background knowledge that we have. 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 number is divisible by 2 or divisible by 4. Ignorance does not entail equally probable. Attempting to “correct” the results is what would in fact introduce a bias.

The idea of clades nested within clades explains why smaller clades should be more probable, but this doesn’t explain larger clades. The high probability of very large clades is simply an artifact of the design of the problem. If our problem uses *n*=10 taxa, for 9 of them to form a clade all of takes is for the tenth to be outside of the rest. However, that same group of 9 taxa is much less likely to form a clade if the problem considered 50 taxa. Suddenly, a group of 8 is more probable than a group of 9 for this problem. 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 9 taxa forming a clade is high given that we are acting as if we are inside a clade of 10 taxa. This is an effect of the size of the problem. But assuming we are looking at less than half of the total taxa in consideration, bigger clades should always be less probable. Actually, this isn’t quite correct as there are distributions that allow clades of size 4 to be more probable then 3 – for example, if every time there was a clade of size four it always split into two groups of two and never into groups of 3 and 1. This can generalize to other sizes, but clades of size 2 will always be the most frequent and “most” distributions will simply scale up in size to leave clades of size 3 more probable than 4, 4 more probable than 5, etc. As for the larger clades (containing more than half of the total taxa in question), probabilities start to increase again. But it is important to see that large clades per se are not more probable – only clades that are large *relative to the total number of taxa being considered* are more probable.

This argument conclusively shows that the probability of a clade must depend on its size, but if we do not carefully distinguish the question, there might appear to be obvious counterexamples. If we think of particular examples, 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, they are not correct. There are several problems with the supposed analogy but the major blunder is that this is an instance of sampling bias. Ignore the fact that some 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 n 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 all the species under consideration a different number and then I 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 is low, but it is 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. Clearly 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. Knowing that we shouldn't try to ignore background knowledge of clades generally does tell us something, but of course it still seems correct that all possible clades of size two should be equally probable, all possible clades of size three 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 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 represent A, B, and C. This condition will be satisfied if, when specifying a tree topology with the taxa labeled 1, 2, 3, … n the probability of the tree can be determined without regard to which number represents which taxa. This is equivalent to the claim that the probability of a tree depends only on its unlabeled shape. 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 unfortunately still leaves us with an infinite number of choices. For example, the uniform prior on topologies satisfies this condition, but so does the distribution that says that the pectinate shape has probability one. While this second distribution is certainly implausible, we can't rule it out simply on the basis of the condition that we must treat each taxa equally.

Although uniform priors on topologies are typically used, we have already seen that several authors believe that it leads to biased results which can be uncovered by examining other "parameters" such as particular clades. While unequal priors on clades is 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 while 3 have the balanced ((A,B),(C,D)) shape. So uniform priors on trees introduces a skewed distribution on shapes. Is this acceptable? The traditional defense for uniform priors on topologies appeals 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 which would favor topologies that were 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 the case of playing cards we know that individual cards are randomly selected; you don't first choose the suit of the card and then choose a card from among that suit. If you did, the suits should be equally probable but they often aren't (if some of the cards are no longer in the deck, etc.). In many cases, the process can vary. Consider the case where Bob has said that he will "randomly" choose a time for his vacation next year. If we want to assign probabilities to which month the first day of that vacation will be in, complete ignorance about Bob might suggest that each month should receive a 1/12 probability, but it might be more reasonable to use a second method which weights the months by the number of days they contain. This second method assumes he selects randomly among the days. The first is only reasonable if we image something like the following: he turned his calendar to a random page then chose a random week or day from it. He may even select his vacation in other ways. For example: He is taking a week long vacation, but this week will be Sunday through Saturday. So if he selects randomly from the "weeks" then we should weight the months by the number of Sundays that they contain. Regardless of the process, the point is that if we know his method of random 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 as a result of 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. This process is captured by the Yule birth process in which particles reproduce with a constant probability of giving birth per particle per unit time.

The Yule Process

In 1925, G.U. Yule developed a statistical model to help explain why some genuses have many more species than others (Yule 1924). 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 from the perspective of 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 pure 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 apriori 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 of 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). And this idea has also made its way into 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 is standardly thought to depend on comparing the accepted phylogeny to some sort of null model of phylogeny. The null model typically used for such comparisons is one of random branching (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 hypothesis such as those about the strength of selection at a particular site or testing the amount of gene flow between distinct populations. (Halliburton 2003, 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, taking this process into account when actually constructing trees is virtually never done. 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 unavoidable even in methods which 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 may be 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 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 symmetric 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? Of course it is going to depend on the particular problem. Problems can be constructed where the priors will matter. Problems can be constructed where they won'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 vertexes that are its descendents (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 4 taxa tree = $\frac{(4-1)!}{3\times1\times1} = 2$. 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 can now calculate the prior probability of any particular tree under the Yule model. To see directly whether this will affect the posterior probability we would need to calculate the normalizing constant – Pr(Data). To do this, we would need 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 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 which 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:

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

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 inverse of the likelihood ratio. So how large is the ratio of the priors? In the 4 taxa case, the most balanced: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, for the perfectly balanced tree with 64 taxa (it splits into two subtrees of 32, each of those splits into two subtrees of 16, etc.) it is consistent with $\frac{(63-1)!}{63\times 31^2 \times 15^4 \times 7^8 \times 3^{16}} \approx 2.61\times 10^{63}$. Since the maximally unbalanced tree which has splits of 1:63 then 1:62, then 1:61, etc. is only consistent with one labeled history, this is also the ratio of the prior probabilities of the trees. For n=128, this ratio rises to ≈ 4.1×10^{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.

The Base-Rate Fallacy

The base-rate fallacy is a refers to a common mistake made in everyday statistical reasoning. That mistake is to ignore the base-rate, or prior probability, of an event thereby leading to a mistaken inference. Here are two standard types of examples in the literature.

Example 1: Medical testing.

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

$$\Pr(\text{AIDS} \mid +\text{test}) = \frac{\Pr(+\text{test} \mid \text{AIDS}) \times \Pr(\text{AIDS})}{\Pr(+\text{test})} \approx \frac{0.95 \times 0.00146}{(0.95 \times 0.00146) + (0.05 \times 0.99854)} \approx 0.027$$

In other words, there is only a 2.7% chance that this person actually has AIDS. The explanation is simple – 5% of all those who don't have AIDS will get a positive test result and this group is much larger than the number of individuals who actually have AIDS. Certainly, the positive test result raises the probability that this person does have AIDS. 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 AIDS (0.95) and ignore the base-rate, we are committing the base-rate fallacy.

Example 2: Eyewitness testimony

Assume that a witness is 80% reliable when it comes to telling the difference between two colors of a car of a particular type. 95% of the cars of this type are green while only 5% are blue . If a witness testifies that he believes that the car was blue, he is probably wrong. By Bayes' Theorem,

$$\Pr(\text{Blue} \mid \text{witness says it is blue}) = \frac{0.8 \times 0.05}{(0.8 \times 0.05) + (0.2 \times 0.95)} \approx 0.174$$

As before, since the probability of a false positive (0.2) is greater than the base-rate (0.05), it is more probable that this is an instance of a false positive than a true one.

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 a base-rate fallacy. Similarly, using a prior distribution, but using the wrong one such as when the uniform distribution is used, commits the same fallacy. 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). What we would expect "by chance" is 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. Of course 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.

References

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Statist.Sci, 16*(1), 23-34.

Brown, J. K. M. (1994). Probabilities of evolutionary trees. *Systematic Biology, 43*(1), 78-91.

Centers for Disease Control and Prevention. *HIV/AIDS Surveillance Report, 2005.* Vol 17. Atlanta: U.S. Department of Health and Human Services, Centers for Disease

Control and Prevention; 2006:1-54. Also available at http://www.cdc.gov/hiv/topics/surveillance/resources/reports/2005report/.

Edwards, A. W. F. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society.Series B (Methodological), 32*(2), 155-174.

Farris, J. S. (1983). The logical basis of phylogenetic analysis. *Advances in Cladistics, 2*, 7–36.

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

Goloboff P.A. and Pol, D. (2005). Parsimony and Bayesian phylogenetics. pp. 148-159 in *Parsimony, Phylogeny and Genomics*, ed. Victor A. Albert. Oxford, OX ; New York: Oxford University Press

Haber, M. H. (2005). On probability and systematics: Possibility, probability, and phylogenetic inference. *Systematic Biology, 54*(5)

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

Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability, 3*(1), 44-77.

Hein, J. (2005). *Gene genealogies, variation and evolution : A primer in coalescent theory*. Oxford ; New York: Oxford University Press.

Huelsenbeck, J. P., & Kirkpatrick, M. (1996). Do phylogenetic methods produce trees with biased shapes? *Evolution, 50*(4), 1418-1424.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science (Washington D C), 294*(5550), 2310-2314.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford), 17*(8), 754-755.

Huelsenbeck, J. P., & Ronquist, F. (2005). Bayesian analysis of molecular evolution using MrBayes. pp. 183-232 in *Statistical Methods in Molecular Evolution*, ed. Rasmus Nielson. New York: Springer.

Joyce, J. (2005). How probabilities reflect evidence. *Philosophical Perspectives, 19*, 153.

Kadane, J. B. (2006) Is “Objective bayesian analysis” objective, bayesian, or wise? (comment on articles by Berger and by Goldstein). *Bayesian Analysis 1(3),* 433-436.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Process.Appl, 13*(3), 235-248.

Kluge, Arnold G. (2005). What is the rationale for 'Ockham's razor' (a.k.a. parsimony) in phlogenetic inference? pp. 15-42 in *Parsimony, Phylogeny and Genomics*, ed. Victor A. Albert. Oxford, OX ; New York: Oxford University Press.

Larget, B., & Simon, D. L. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution, 16*(6), 750-759.

Larget, B. (2005). Introduction to Markov Chain Monte Carlo methods in molecular evolution. pp. 44-61 in *Statistical Methods in Molecular Evolution*, ed. Rasmus Nielson. New York: Springer.

Metzker, M. L., Mindell, D. P., Liu, X. M., Ptak, R. G., Gibbs, R. A., & Hillis, D. M. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences, 99*(22), 14292-14297.

Mooers, A. O., & Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology, 72*(1), 31-54.

Pickett, K. M., & Randle, C. P. (2005). Strange bayes indeed: Uniform topological priors imply non-uniform clade priors. *Molecular Phylogenetics and Evolution, 34*(1), 203-211.

Pitnick, S., Jones, K. E., & Wilkinson, G. S. (2006). Mating system and brain size in bats. *Proceedings of the Royal Society Biological Sciences Series B, 273*(1587), 719-724.

Poux, C., Madsen, O., Marquard, E., Vieites, D. R., de Jong, W. W., & Vences, M. (2005). Asynchronous colonization of madagascar by the four endemic clades of primates, tenrecs, carnivores, and rodents as inferred from nuclear genes. *Systematic Biology, 54*(5), 719-730.

Randle, C. P., Mort, M. E., & Crawford, D. J. (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. *Journal of Molecular Evolution, 43*(3), 304-311.

Semple, C., & Steel, M. (2003). *Phylogenetics*. Oxford ; New York: Oxford University Press.

Siddall, M. E., & Kluge, A. G. (1997). Probabilism and phylogenetic inference. *Cladistics, 13*(4), 313-336.

Steel, M., & McKenzie, A. (2001). Properties of phylogenetic trees generated by yule-type speciation models. *Mathematical Biosciences, 170*(1), 91-112.

Thompson, E. A. (1975). *Human evolutionary trees*. Cambridge Eng. ; New York: Cambridge University Press.

Yang, Ziheng (2006). *Computational molecular evolution.* Oxford ; New York: Oxford University Press

Yang, Z., & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A markov chain monte carlo method. *Molecular Biology and Evolution, 14*(7), 717-724.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of dr. JC willis, FRS. *Philosophical Transactions of the Royal Society of London.Series B, Containing Papers of a Biological Character, 213*, 21-87.