On Phylogenetic Uncertainty and Ancestral Sequence Reconstruction

A directed research project for the department of Computer and Information Science

Victor Hanson-Smith
with advising from Joe Thornton and John Conery

February 23, 2009

Contents

1 Introduction **4**
- 1.1 Early Approaches to Infer Phylogenies 8
- 1.2 Markov Models 11
- 1.3 Likelihoods on Trees 14
- 1.4 Maximum Likelihood 15
- 1.5 Bayesian Methods 18
- 1.6 Phylogenetic Uncertainty 20
- 1.7 Ancestral Reconstruction 22
- 1.8 Uncertainty and Ancestral Reconstruction 25

2 Methods **27**
- 2.1 An Algorithm to Integrate Phylogenetic Uncertainty 27
- 2.2 Simulations 31
- 2.3 Criteria for Grouping the Results 34

3 Results **36**
- 3.1 The ML and EB methods infer the same state at almost all ancestral sites 36
- 3.2 The ML and EB methods infer disagreeing states at poorly-supported sites on poorly-supported trees. 38
- 3.3 When ML and EB reconstructions differ, EB is usually less accurate 41
- 3.4 The ML and EB methods yield slightly different posterior probability values, but the overall accuracy of those values is nearly the same 42
- 3.5 Phylogenetic uncertainty is not correlated with ASR accuracy 46

4 Conclusions **48**

Author's Summary

In this project, I consider the problem of inferring the ancient evolutionary history of molecular gene sequences. Given the extreme paucity of molecular fossils, the history of genes can be difficult to study. However, computational methods of ancestral sequence reconstruction (ASR) can be used to statistically infer the sequences of extinct genes; some of these reconstructions have been chemically synthesized and experimentally tested. Although ASR allows us to answer previously unknowable questions about evolutionary molecular mechanisms, results from ASR-based experiments rely on the accuracy of their underlying computational reconstruction. In this project, I investigate one aspect of the ASR algorithm which may impact accuracy: phylogenetic uncertainty. Most reconstruction algorithms assume the phylogeny is known with certainty; in practice, this assumption is rarely valid. Does ignoring phylogenetic uncertainty affect ASR accuracy?

To answer this question, I proposed an empirical Bayesian algorithm for integrating phylogenetic uncertainty in ASR. I examined this method in simulated and real conditions. My results are surprising and nonintuitive: phylogenetic uncertainty is not correlated with the accuracy of reconstructed ancestral states. The conditions which produce phylogenetic uncertainty result in ancestral states on alternate trees which are similar, if not identical, to the ancestral states on the maximum likelihood tree. Ultimately, integrating phylogenetic uncertainty does not significantly affect the accuracy of reconstructed ancestral sequences.

1 Introduction

The evolution of life likely started over 3.8 billion years ago [Fenchel, 2002, Knoll, 2004] (see also [Schopf, 2006]). However, evolutionary processes can be challenging to directly observe given the relative brevity of a human lifetime. Although biologists have traditionally used fossils to infer evolutionary history, the fossil record is incomplete and not easily searchable. Simply put, our knowledge of evolutionary history should not rely on lucky shovels. The advent of gene sequencing technology provides alternatives to fossil-based inference: there exists a growing body of algorithms to statistically infer evolutionary history from observed extant gene sequences. The computational challenges of phylogenetic inference and ancestral reconstruction are frontiers in computer science, uniting aspects of biology, chemistry, statistics, and information science. In this paper, I address the specific problem of phylogenetic uncertainty in ancestral reconstruction. Before discussing this topic in detail, I want to familiarize the computer science reader with necessary concepts in evolutionary biology and statistical inference. Thus, this introduction. . .

Evolutionary forces act on phenotypes: the physical, behavioral, and biochemical attributes of an organism. The individual organisms in a population vary in their phenotypes. Some individuals have phenotypic variants which are better adapted to the conditions of life; these individuals are more likely to survive and reproduce. As an example, consider the peppered moth *Biston betularia* [Ridley, 2004]. Before the British industrial revolution, naturalists observed the majority of individuals within peppered moth populations had light-colored bodies. The light coloration provided natural camouflage on lichen and bark. When industrial activity covered the forests of central Britain with dark soot, the majority of peppered moths were no longer camouflaged and became easy prey for hungry birds. At the same time, a curious phenotypic minority of moths had dark coloration. The dark moths were well-adapted to the sooty forest and were more likely to survive predation from birds. Over time, evolution selected for the dark phenotype; the light phenotype declined into minority.

Although evolution acts on phenotypes, the phenotypes themselves are encoded in genetic material – i.e. DNA. Within each living cell, coding regions of DNA are transcribed into RNA and then translated into proteins. Individual proteins and networks of proteins are expressed in patterns which determine an individual's phenotype. The flow of information from DNA to RNA to protein is considered to be the central dogma of biology. (As a corol-

lary, the discovery of reverse transcriptase extends and challenges the dogma [Baltimore, 1970, Temin and Mizutani, 1970]). Individuals which survive and reproduce pass their genes to future generations; in this way, an individual's phenotype is informed by its ancestors' phenotype.

Over time, genetic lineages diverge and independent lineages are free to accumulate unique mutations. A history of evolutionary lineage-splitting can be expressed as a type of tree graph, called a cladogram. The terminal nodes of a cladogram correspond to extant taxa ; the internal branching pattern expresses the shared ancestry of the terminal nodes. Some cladograms are rooted, in which case the root node corresponds to the most-common-shared ancestor of all taxa on the tree. Figure 1 illustrates a simple rooted cladogram for the history of the family *Hominidae*.

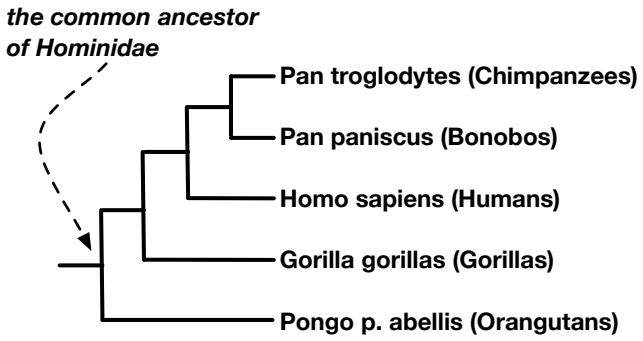


Figure 1: The shared evolutionary history of the family *Hominidae* can be expressed as a cladogram. Although this cladogram does not measure time, the direction of evolution progresses from left to the right.

A phylogram is a special type of cladogram in which the branch lengths (i.e. the edge lengths) correspond to a measure of evolutionary distance. "Distance" corresponds to the number of observed (or expected) character differences between nodes. The problem of phylogenetic inference is to determine the correct tree topology and branch lengths for a given set of extant taxa.

Phylogenetic inference occurs in three steps. First, we collect observations about evolutionary characters for a set of descendant taxa. Second, we arrange those observations into a sequence alignment. Finally, we use the sequence alignment to infer a phylogeny. Here I describe these three steps in detail.

The first step is to select a set of extant taxa and observe evolutionary characters for those taxa. These characters can be discrete or continuous. For example, we might consider the character "presence of mammary glands" (a discrete binary character) to distinguish mammals from non-mammals. If we are analyzing birds, we might consider the character "beak length" (a continuous character) to help distinguish among bird species. Gene sequencing technology allows us to observe evolutionary characters at the molecular level. Nucleotide sequences (i.e. DNA sequences) include characters from an alphabet with four possible states: A, C, G, and T. These characters correspond to the four nucleotides which compromise the structure of DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). Nucleotide three-tuples encode for the alphabet of amino acids. For example, the nucleotide combinations GCT, GCC, GCA, and GCG all encode for the amino acid alanine; the combinations TAT and TAC both encode for tyrosine. There exists 2^3 possible nucleotide three-tuples, but redundant coding yields only twenty possible amino acids.

Figure 2 illustrates how a short string of amino acid characters might evolve over time. This example begins with an ancestral sequence $NEDP$, which stands for the amino acids asparagine, glutamic acid, aspartic acid, and proline. $NEDP$ speciates into two divergent lineages. In one lineage, the character N evolves to D and then E, giving rise to the extant descendant $EEDP$. In the other lineage, the character D evolves to V, giving rise to the intermediate ancestral sequence $NEVP$. In this example, the intermediate ancestor also speciates, leading to extant descendants $NQVP$ and $NQVA$. Although figure 2 explicitly shows the evolutionary history of the three descendant sequences, in practice the history is hidden from us.

The second step in phylogenetic inference is to arrange the character observations into a matrix called a sequence alignment. Each row in a sequence alignment corresponds to a string of character observations from one taxa. Each column corresponds to a character which may or may not appear in several taxa. Sequence alignments are critical for phylogenetic inference because alignments allow us to observe which characters are shared (or not shared) between extant taxa. For the descendant amino acid sequences in figure 2, the alignment is trivial:

```
EEDP
NQVP
NQVA
```

A more complex example would involve insertion and deletion characters, otherwise known as indels. Over time, sections of DNA are inserted and

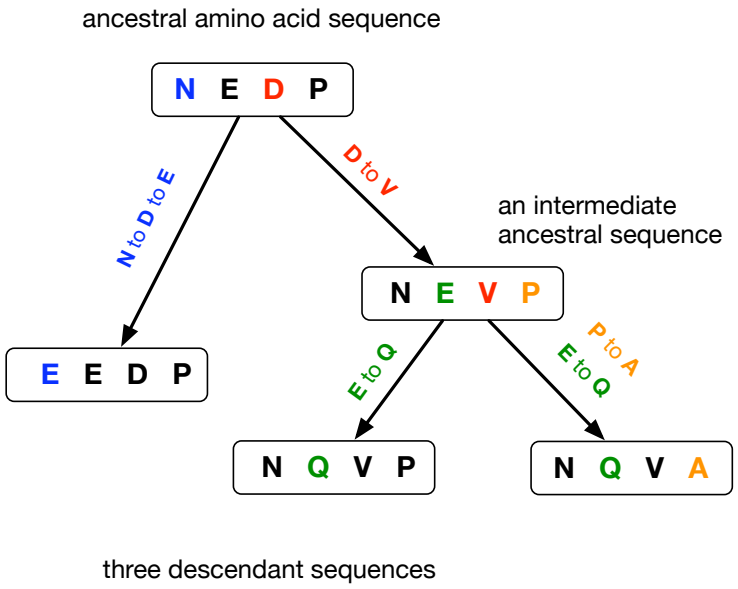
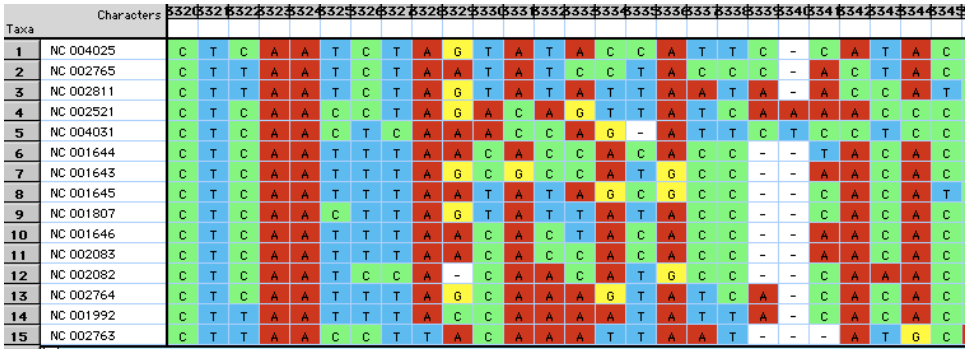


Figure 2: In this example, an ancestral amino acid sequence $NEDP$ evolves into three descendant sequences.

deleted. Consequently, a particular gene might be encoded with x number of nucleotides in one species, but $x + \Delta$ number of nucleotides in another species.

In a practical analysis, one of the challenges of aligning sequences is to determine where the insertion and deletion events occurred. There exist several algorithms and tools for aligning molecular sequences [Chenna et al., 2003, Thompson et al., 1994, Cedric Notredame, 2000, Edgar, 2004, O'Sullivan et al., 2004, Do et al., 2005, Armougom et al., 2006, Subramanian et al., 2008]. Figure 3 shows part of a sequence alignment for mitochondrial primate DNA. The full alignment is 18,201 characters long; for brevity I've shown only sites 3320 to 3345. Figure 3 shows that some characters are well conserved across all extant taxa, while other characters are wildly divergent. A comprehensive discussion of alignment algorithms is outside the scope of this paper; for more information, see Edgar and Batzoglou's review [Edgar and Batzoglou, 2006].

Given a sequence alignment, the third step of phylogenetic inference is to actually infer the tree and its branch lengths. Figure 4 shows an inferred phylogeny for the primate DNA from our previous example. Although numerous approaches have been proposed for inferring phylogenies,



Taxa		Characters	3320	3321	3322	3323	3324	3325	3326	3327	3328	3329	3330	3331	3332	3333	3334	3335	3336	3337	3338	3339	3340	3341	3342	3343	3344	3345
1	NC 004025		C	T	C	A	A	T	C	T	A	G	T	A	T	A	C	C	A	T	T	C	-	C	A	T	A	C
2	NC 002765		C	T	T	A	A	T	C	T	A	A	T	A	T	C	C	T	A	C	C	C	-	A	C	T	A	C
3	NC 002811		C	T	T	A	A	T	C	T	A	G	T	A	T	A	T	T	A	A	T	A	-	A	C	C	A	T
4	NC 002521		C	T	C	A	A	C	C	T	A	G	A	C	A	G	T	T	A	T	C	A	A	A	A	C	C	C
5	NC 004031		C	T	C	A	A	C	T	C	A	A	A	C	C	A	G	-	A	T	T	C	T	C	C	T	C	C
6	NC 001644		C	T	C	A	A	T	T	T	A	A	C	A	C	C	A	C	A	C	C	-	-	T	A	C	A	C
7	NC 001643		C	T	C	A	A	T	T	T	A	G	C	G	C	C	A	T	G	C	C	-	-	A	A	C	A	C
8	NC 001645		C	T	C	A	A	T	T	T	A	A	T	A	T	A	G	C	G	C	C	-	-	C	A	C	A	T
9	NC 001807		C	T	C	A	A	C	T	T	A	G	T	A	T	T	A	T	A	C	C	-	-	C	A	C	A	C
10	NC 001646		C	T	C	A	A	T	T	T	A	A	C	A	C	T	A	C	A	C	C	-	-	A	A	C	A	C
11	NC 002083		C	T	C	A	A	T	T	T	A	A	C	A	C	C	A	C	A	C	C	-	-	A	A	C	A	C
12	NC 002082		C	T	C	A	A	T	C	C	A	-	C	A	A	C	A	T	G	C	C	-	-	C	A	A	A	C
13	NC 002764		C	T	C	A	A	T	T	T	A	G	C	A	A	A	G	T	A	T	C	A	-	C	A	C	A	C
14	NC 001992		C	T	T	A	A	T	T	T	A	C	C	A	A	A	A	T	A	T	T	A	-	C	A	C	A	C
15	NC 002763		C	T	T	A	A	C	C	T	T	A	C	A	A	A	T	T	A	A	T	-	-	-	A	T	G	C

Figure 3: An alignment of fifteen primate mitochondrial genomes: The full alignment is 18,201 characters long. For brevity, only the characters in positions 3320 to 3345 are shown. The sequences are identified by their GenBank accession numbers [Burks et al., 1992]. These sequences were aligned using Clustal X [Chenna et al., 2003], and visualized using MacClade [Maddison and Maddison, 1992]. Each cell is colored according to its nucleotide state. The white cells indicate indels (i.e. insertions or deletions).

they generally fall into one of four categories: parsimony-based methods, distance-based methods, maximum likelihood methods, and Bayesian methods [Felsenstein, 2004]. Sections 1.1 through 1.5 of this manuscript discuss these approaches.

1.1 Early Approaches to Infer Phylogenies

Maximum parsimony (MP) is one of the earliest methods for inferring a phylogeny from a sequence alignment. According to the principle of parsimony, the explanation which requires the fewest number of ad-hoc hypothesis should be accepted in lieu of more complicated explanations. In the context of biological evolution, MP seeks the phylogeny which requires the minimum "evolutionary cost" between ancestral nodes and leaf nodes [Edwards and Cavalli-Sforza, 1963]. "Evolutionary cost" typically corresponds to the number and type of character changes. Several parsimony criteria have been proposed for measuring cost [Camin and Sokal, 1965, Kluge and Farris, 1969, Farris, 1970, Fitch, 1971, Farris, 1977].

Despite it's computational simplicity, MP is not statistically consistent [Felsenstein, 1978]. A method of phylogenetic inference is said to be consistent if, as the length of the observed sequences increases, the method converges on the true phylogeny. Given (impossible) infinite-length sequences, a consistent method will recover the true phylogeny every time. Incon-

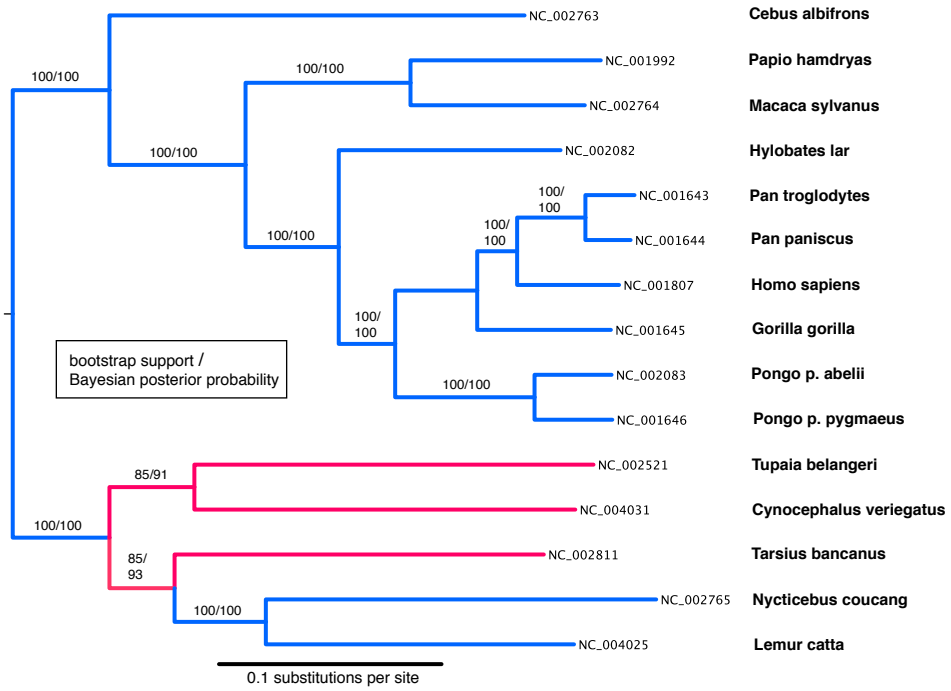


Figure 4: The primate phylogeny, reconstructed using the nucleotide sequences for cytochrome-B (CYTB) and cytochrome oxidase I (COI) genes. Here, I downloaded fifteen mitochondrial nucleotide genomes and two non-primate outgroup genomes from GenBank [Burks et al., 1992]. Using ClustalX [Chenna et al., 2003], I aligned the fifteen genomes using the *slow / accurate* method. I explored a range of gap opening costs (2, 10, 15, 20), and a range of gap extension penalties (0.05, 0.1, 0.2), for a total of twelve unique alignments. Unsure of which alignment is best, I engineered a small software engine to iteratively infer maximum likelihood phylogenies for each alignment (using PhyML [Guindon and Gascuel, 2003]) and to find the best-fit model (using AIC [Akaike, 1973]). Surprisingly, all twelve alignments produce the same ML topology and were all best-fit by the GTR model [Tavare, 1986]. Alternatively, I could have incorporated alignment uncertainty by eliding the twelve alignments [Wheeler et al., 1995]. The labels for terminal taxa correspond to Genbank accession numbers. The bootstrap and Bayesian support values are listed at each internal node. The pink branches form a topology which disagrees with Krishnan's original CYTB/COI phylogeny [Krishnan et al., 2004].

sistent methods pull us towards an incorrect hypothesis (i.e. an incorrect phylogeny). Although MP yields the correct tree in many cases, Felsenstein showed the method can be inconsistent when the true evolutionary history did not occur according to a fixed evolutionary rate [Felsenstein, 1978].

At roughly the same time that parsimony methods were being developed, other methods based on distance matrices were proposed [Cavalli-Sforza and Edwards, 1967, Fitch and Margoliash, 1967]. The key idea of these methods is to construct an $n \times n$ distance matrix, where n is the number of extant taxa in the analysis. Each entry $[i, j]$ expresses the fraction of characters in which sequences i and j differ. For example, consider the descendant sequences in figure 2. The descendants $NQVP$ and $NQVA$ differ by one character out of four; their distance is 0.25. This distance is actually an underestimate of the true evolutionary distance. Although $NQVP$ and $NQVA$ appear to be separated by 0.25 substitutions per site, they are actually separated by 0.75 substitutions per site: the change $E \rightarrow Q$ occurred twice and $P \rightarrow A$ occurred once. The key idea is that some mutational events are not observable. If all character states in all lineages changed to all other possible states at the same rate, then converting observed distances into true evolutionary distances would be relatively straightforward [Jukes and Cantor, 1969]. In practice, evolutionary rates are not always constant or unbiased, so calculating the correct distance between two sequences remains problematic.

If we momentarily assume distances are correct, a phylogeny can be constructed from a distance matrix by considering all possible topologies, treating branch lengths as a variable parameter, and selecting the tree which best fits the pairwise distances. In practice, an exhaustive examination of all possible trees is often computationally intractable. Therefore, heuristics are used to find the "best" tree in a practical amount of time. The unweighted pair group method with arithmetic mean (UPGMA) [Sokal and Sneath, 1963] and the neighbor-joining method [Saitou and Nei, 1987] are two such heuristics. Once a tree is constructed, its optimality can be measured using the least-squares [Cavalli-Sforza and Edwards, 1967, Fitch and Margoliash, 1967] or minimum evolution [Kidd and Sgaramella-Zonta, 1971].

Distance methods have limitations [cite Farris 1982?]. As already mentioned, converting observed distances into evolutionary distances remains problematic. Furthermore, we lose information when we convert sequences into distance matrices. This shortcoming has been explained in terms of a person arriving in one city from another place [Osborn and Smith, 2005]: The distance method takes into account only how far the person has travelled, whereas other methods (such as parsimony) attempt to reconstruct the actual route taken. Finally, Felsenstein observes that distance methods

are incapable of propagating information across the tree [Felsenstein, 2004]. He explains this limitation in terms of rate variation: if we observe a fast evolutionary rate in one lineage, we should use this information to affect our interpretation of evolutionary rates in other lineages. Distance methods cannot do this.

In the remainder of this introduction, I turn our attention away from parsimony and distance methods and focus on probabilistic methods. These methods rely on explicit models of character substitution. As we shall see, one of the challenges of probabilistic methods is to design evolutionary models which combine realism with computational tractability.

1.2 Markov Models

By treating molecules as character sequences, molecular evolution can be understood within an information theoretic framework. To motivate this view, let's begin with the the simplest case: a sequence with one character x. The evolution of a single ancestral amino acid x into some descendant amino acid x' can be modeled as the transmission of data over a discrete memoryless channel. Evolutionary mutations introduce noise into this channel: in other words, x' might not equal x.

Because our channel is memoryless, we can model the evolution of $x \to x'$ as a continuous-time Markov process. A Markov process describes a system evolving according to the Markov property: the conditional probability of the system's future state depends only on the present state and not on past states. As a demonstration of the Markov property, suppose the ith site in some ancestral protein sequence is asparagine (represented as 'N'). After some time, suppose N mutates into aspartic acid ('D'). After more time, suppose D mutates into glutamic acid ('E'). The mutation D $\to$ E occurs independently of the previous mutation N $\to$ D. The inverse is also true: the mutation N $\to$ D occurs independently of the future mutation D $\to$ E. In this way, the channel is memoryless: the path N $\to$ D $\to$ E is forgotten and we observe only the present state E.

Although we could apply any number of alternate models, the Markov model provides an acceptable trade-off between biological realism and computational tractability. That said, the basic Markov process for molecular evolution makes a key assumption: each site within a sequence evolves independently of other sites. As an example, suppose an ancestral sequence with two characters x_i and x_j evolves into a sequence with characters x'_i and x'_j. The basic Markov model assumes the evolution of $x_i \to x'_i$ occurs independently of $x_j \to x'_j$.

Here I discuss how to calculate the probability of ancestral state x evolving into descendant state x' over time t. Typically, the timing and quantity of mutation events over some time t is unknown. However, we can use a Poisson distribution as an approximation. Given a mutation rate μ, the probability of k mutations over time t is:

$$p(k|t,\mu) = \frac{(\mu t)^k e^{-\mu t}}{k!} \tag{1}$$

Let ε be the set of all possible states in our character alphabet. For DNA sequences, $\varepsilon = \{A, C, G, T\}$; for protein sequences, ε equals the set of twenty amino acids. Let Q be the instantaneous substitution rate matrix. Q_{ij} expresses the relative rate at which character i mutates to character j at a differential moment in time, where $i, j \in \varepsilon$. We combine the matrix Q with the Poisson distribution (from equation 1) to calculate the probability of specific character changes over time t. Before describing that process, I will discuss some properties of the matrix Q.

Q matrices are specified by one or more parameters. For example, the Q matrix for the JC69 model of evolution is specified in terms of a single parameter μ, representing the overall mutation rate [Jukes and Cantor, 1969]. Thus:

$$Q = \begin{Bmatrix} -\frac{3}{4}\mu & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & -\frac{3}{4}\mu & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3}{4}\mu & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3}{4}\mu \end{Bmatrix}$$

A more complex example, the general time reversible (GTR) model, is specified in terms of six rate parameters $\{a, b, c, d, e, f\}$ and four base frequencies $\{\pi_a, \pi_c, \pi_g, \pi_t\}$ [Tavare, 1986]. Thus:

$$Q = \begin{Bmatrix} -\mu(a\pi_c + b\pi_g + c\pi_t) & \mu a\pi_c & \mu b\pi_g & \mu c\pi_t \\ \mu a\pi_a & -\mu(a\pi_a + d\pi_g + e\pi_t) & \mu d\pi_g & \mu e\pi_t \\ \mu b\pi_a & \mu d\pi_c & -\mu(b\pi_a + d\pi_c + f\pi_t) & \mu f\pi_t \\ \mu c\pi_a & \mu e\pi_c & \mu f\pi_g & -\mu(c\pi_a + e\pi_c + f\pi_g) \end{Bmatrix}$$

In addition to JC69 and GTR, there exist several other nucleotide models of varying complexity. To name a few: K80 [Kimura, 1980], HKY85 [Hasegawa et al., 1985], and F81 [Felsenstein, 1981].

Unlike nucleotide models, substitution matrices for amino acid models are often derived empirically from large databases of sequences [Dayhoff et al., 1978, Jones et al., 1991, Adachi and Hasegawa, 1996]. This approach occurs in three steps. First, all the amino acid sequences in a database are phylogenetically arranged. For instance, we could build a distance-based tree for all the

sequences in the SWISS-PROT database [Bairoch and Boeckmann, 1992]. Second, pairs of sequences with similarity greater than some cutoff are identified. For example, the process for building the JTT model finds pairs of sequences with $\geq$85% similarity. Finally, for all amino acids i and j, the proportion of each substitution $i \rightarrow j$ is counted in all sequence pairs. These counts are converted into overall proportions and expressed as a substitution matrix. The approach I just described is essentially a distance method, and has been shown to underestimate the true number of character changes [cite Goldman? 1990]. Several likelihood-based approaches have been proposed to overcome distance underestimation [Adachi and Hasegawa, 1996, Adachi et al., 2000]. Whelan and Goldman proposed the WAG model, which combines attributes of the counting methods and the likehood-based methods [Whelan and Goldman, 2001].

Regardless of which model we choose, the algorithm for calculating the probability of state change over a phylogenetic branch of length t is the same for nucleotide and amino acid data. Let R be the matrix satisfying the expression $Q = R - I$, where I is the identity matrix. (Nucleotide models are often specified in terms of Q, whereas protein models are often specified in terms of R.) Let $P(t)$ be a matrix where $P(t)_{ij}$ expresses the probability of mutating from state i to state j over time t. We construct $P(t)$ by combining the the Poisson distribution (from equation 1) with an exponentiated form of the matrix R. Thus:

$$P(t) = \sum_{k=0}^{\infty} (R^k) \frac{(\mu t)^k e^{-\mu t}}{k!} \tag{2}$$

And with a rearrangement, we get:

$$P(t) = e^{-\mu t} \sum_{k=0}^{\infty} (R^k) \frac{(\mu t)^k}{k!} \tag{3}$$

The summation in equation 3 has the same form as a matrix exponent series:

$$e^{R\mu t} = \sum_{k=0}^{\infty} (R^k) \frac{(\mu t)^k}{k!} \tag{4}$$

Therefore, equation 3 can be simplified to:

$$P(t) = e^{-\mu t} e^{R\mu t} \tag{5}$$

If we introduce the identity matrix and rearrange, we get:

$$P(t) = e^{-\mu t I} e^{R\mu t} \tag{6}$$

$$P(t) = e^{(R-I)\mu t} \tag{7}$$

$$P(t) = e^{Q\mu t} \tag{8}$$

Although equation 8 gives a compact and direct expression of $P(t)$, the formula is not practical for software implementation. Equation 8 is typically computed using Eigen decomposition as follows:

$$e^{Q\mu t} = Ae^{D\mu t}A^{-1} \tag{9}$$

where D is the diagonal matrix of Eigenvalues and A are the Eigenvectors. A full treatment of Eigen decomposition is outside the scope of our discussion; a comprehensive description is found at [Golub and Loan, 1996].

1.3 Likelihoods on Trees

In section 1.2, I introduced the matrix $P(t)$, which expresses the probability of any given state mutating to another state over a single phylogenetic branch of length t. In this section, I extend our discussion to consider the probability of character mutations on multiple branches in a phylogenetic tree.

Consider a sequence alignment with N sites (i.e. columns) and M taxa (i.e. rows). For now, assume the tree and its branch lengths are given. Also assume the model of evolution and the parameters for it's Q matrix are given. Let θ be the vector containing the model parameters and the branch lengths. If we assume each site evolves independently, then the probability of the alignment is the product of probabilities for each site. In other words:

$$P(alignment|tree, \theta) = \prod_{i=1}^{N} P(site_i|tree, \theta) \tag{10}$$

The conditional probability $P(alignment|tree, \theta)$ is also known as the likelihood of the tree and θ. There exists a dynamic downpass algorithm to calculate $P(alignment|tree, \theta)$. For the sake of brevity, I describe this algorithm for DNA data (with four possible states). This algorithm can be trivially extended to accommodate protein data (with twenty possible states). The algorithm follows:

For each tree node p, we initialize the vector $G_p = \{g_p^A, g_p^C, g_p^G, g_p^T\}$. Each element of G_p expresses the conditional probability of the subtree rooted at p, assuming the particular state was assigned to p. For the terminal nodes, we set $g_p^i = 1$ if we observe state i; we set $g_p^j = 0$ for all other states. We proceed from the tips to the interior nodes, according to a postorder traversal. At each internal node p, we update G_p. For an internal node p, with two descendant branches of length r_1 and r_2 leading to nodes u_1 and u_2, we calculate g_p^ias follows:

$$g_p^i = \sum_{j}^{j \in \{ACGT\}} P(r_1)_{ij} g_{u_1}^j \times \sum_{j}^{j \in \{ACGT\}} P(r_2)_{ij} g_{u_2}^j \tag{11}$$

In other words, the conditional probability of state i at node p is the sum of probabilities that i changes to any state j (where we allow for $j = i$) along branch r_1 and r_2.

Eventually the downpass algorithm will calculate G_p for the root node, after which the total probability of the tree for site i is calculated by summing over all possible states, each weighted by its stationary state frequency:

$$P(site_i | tree, \theta) = \sum_{j}^{j \in \{ACGT\}} \pi_j g_{root}^j \tag{12}$$

Putting everything together, we calculate the likelihood of a tree in two steps:

1. Use Eigen decomposition to calculate $e^{Q\mu t}$, as shown in equation 9.
2. Perform a post-order traversal of the tree, applying equation 11 for every state and every site.

The first step is much more computationally demanding than the second. If we have N sites, M taxa, and C possible states, step one takes $O(MC^3)$ time, and step two takes $O(MNC^2)$ time [Bryant et al., 2005]

1.4 Maximum Likelihood

Until now, our discussion assumed the phylogeny and its branch lengths were known. In practice, the phylogeny (and its branch lengths) are typically unknown. The challenge of finding the best phylogeny can be computationally nontrivial, especially when dealing with large taxa sets. In this section,

I discuss how to use maximum likelihood (ML) to infer phylogenies from sequence alignments.

Maximum likelihood starts with a model of evolution, which explains how the observed data arose. Given a set of input parameter values, the model gives the probability of observing the data. The basic idea of ML is to select parameter values that maximize the probability of observing the data. In the context of phylogenetic inference, the model parameters include the values of the substitution matrix, the topology of the tree, and the tree's branch lengths.

ML was originally developed by Fisher [Fischer, 1922] (see also [Aldrich, 2007]). Felsenstein introduced ML to the problem of phylogenetic inference; he showed that ML is consistent for selecting an evolutionary model and for estimating branch lengths [Felsenstein, 1981]. However, Felsenstein's proof falls short of showing that ML is consistent for selecting tree topologies. In response, Yang articulated a proof that ML is consistent for selecting topologies [Yang, 1994], and Rogers found a stronger proof [Rogers, 1997].

The problem of finding the maximum likelihood phylogeny is an optimization problem, where we seek the optimal combination of tree topology, branch lengths, and model parameters. The space of possible trees can be immense. For n extant taxa, there exist $\frac{(2n-3)!}{2^{(n-2)}(n-2)!}$ possible unrooted topologies, each with $2n-2$ branches. For a small dataset, with only twenty taxa, there exist 2.22×10^{39} possible unrooted topologies, each with 40 branches. Each topology has an infinite number of branch lengths and model parameters. For most practical analyses, an exact exhaustive search through the space of possible topologies, branch lengths, and model parameters is computationally intractable. Consequently, a large number of heuristic solutions have been proposed and implemented. In general, these solutions construct an initial tree and then use hill-climbing algorithms to optimize the topology, branch lengths, and model parameters.

Swofford reviews five methods for constructing initial trees [Swofford et al., 1996]: the distance method, random trees, sequential insertion, star decomposition, or approximate likelihood. Of these options, the software package PhyML uses the distance method [Guindon and Gascuel, 2003]; the software package PAML uses sequential insertion [Yang, 1997]; and the package PAUP implements all five methods [Swofford, 2003].

After constructing an initial tree, hill-climbing algorithms can be used to optimize the topology, branch lengths, and model parameters. Broadly speaking, optimization algorithms seek to optimize either one dimension or multiple dimensions at each iterative step. PAML uses the one-dimension

Newton-Raphson method [van der Vaart, 2000] to independently optimize tree topology and branch lengths, whereas PhyML uses a multidimensional variant of this method [Guindon and Gascuel, 2003]. PAUP implements four optimization methods: the single-dimension Newton-Raphson method, Brent's one-dimension algorithm [Brent, 1972], a multi-dimensional variant of the Newton-Raphson method [Brent, 1972], and a multi-dimensional variant of the Simplex method [Dantzig, 1963]. The question of single versus multi-dimensional optimization appears unaddressed in phylogenetic literature; this might be a fruitful area of future research.

In addition to the aforementioned ML algorithms, several computational strategies are being investigated. Matsuda introduced a genetic algorithm for inferring maximum likelihood phylogenies from protein sequences [Matsuda, 1996]; Lewis introduced a similar method for nucleotide sequences [Lewis, 1998] and implemented the method in the software package named GAML. Later, Zwickl reimplemented GAML into a software package named GARLI; the reimplementation uses the general-time reversible model (and all its subsets), in addition to using a more efficient genetic selection process. Genetic algorithms aside, Kolaczkowsi and Thornton introduced a simulated annealing algorithm to estimate ML tree topologies, model parameters, and the number of branch length categories [Kolaczkowski and Thornton, 2008]; Kolaczkowski's algorithm is prototypically implemented in the software package SAML.

ML phylogenetic inference can be computationally demanding, especially when using genetic algorithms or simulated annealing approaches. These demands are being addressed by software parallelization and advances in multiprocessor architectures. Stamatakis ported his software package RaxML [Stamatakis et al., 2005] to the IBM BlueGene/L [Ott et al., 2007] and the IBM Cell processor [Stamatakis et al., 2007]. Depending on the parallelization strategy (fine-grained versus coarse-grained) the parallelized version of RaxML achieved between logarithmic and linear speedup. Genetic algorithms have also been parallelized; in fact, genetic algorithms lend themselves to so-called embarrassing parallelization. Lewis' GAML software was parallelized with nearly linear speedup [Brauer et al., 2002], and later GARLI was parallelized to obtain significantly more optimal phylogenies than the serial implementation. Unlike genetic algorithms, simulated annealing algorithms do not lend themselves to easy parallelization. That said, approaches have been proposed for the general non-phylogenetic case [Azencott, 1992]. Given that simulated annealing was only recently applied to the problem of inferring phylogenies, much work remains to implement computational optimality.

1.5 Bayesian Methods

In recent years, Bayesian methods have been proposed as an alternative to maximum likelihood methods. As with ML, Bayesian methods use a model of evolution to explain how the observed data arose. In other regards, ML and Bayesian methods are philosophically different. Unlike ML, Bayesian methods require us to incorporate prior beliefs about the model's parameter values. Whereas ML calculates point estimates of the model parameters, Bayesian methods calculate integrated estimates. Finally, ML yields likelihoods, while Bayesian methods yield posterior probabilities.

The Bayesian posterior probability of a tree, branch lengths, and model parameters is calculated according to Bayes' theorem. Let D be a set of observed data (i.e. a sequence alignment), let T be a given tree topology, let M be a model of evolution, and let θ be the vector containing the model parameters. The posterior probability of T, M, and θ, given D, is

$$P(T, M, \theta|D) = \frac{P(T, M, \theta)L(D|T, M, \theta)}{P(D)} \tag{13}$$

In the numerator, $P(T, M, \theta)$ specifies the prior probability of the tree topology, model, and model parameters. $L(D|T, M, \theta)$ is the likelihood function: the likelihood of observing the data, given the topology, model, and model parameters. In the denominator, $P(D)$ is the total probability of the data summed and integrated over all tree topologies, models, and model parameters. Thus:

$$P(D) = \int_{T,M,\theta} P(T, M, \theta)P(D|T, M, \theta)d\{T, M, \theta\} \tag{14}$$

For the purposes of unraveling evolutionary history, we are concerned with finding the best topology given all possible models and sets of model parameters. In other words, the Bayesian approach to finding the best topology is to integrate over uncertainty about the model and the model parameters. Consequently, equation 15 is recapitulated in these terms:

$$P(T|M, \theta|D) = \frac{P(T)L(DM, \theta|T)}{P(D)} \tag{15}$$

Although Bayes' theorem was introduced in the 18th century, Bayesian phylogenetic inference was adopted only recently due to the difficulty of analytically integrating posterior probabilities. A solution to this problem came by way of Stanislaw Ulam, John von Neumann, and Nicholas Metropolis: in the 1940's they developed the so-called "Monte Carlo" algorithm as a means

to numerically approximate solutions to differential equations on the ENIAC machine [Metropolis, 1949, Metropolis, 1987]. Metropolis et al. later articulated the Monte Carlo algorithm in the context of a Markov chain, thus inventing the Monte Carlo Markov Chain (MCMC) [Metropolis et al., 1953]. Several groups worked to developed MCMC methods for Bayesian phylogenetic inference [Rannala and Yang, 1996, Mau, 1996, Mau and Newton, 1997, Mau et al., 1999, Li, 1996, Larget and Simon, 1999, Yang and Rannala, 1997, Newton et al., 1999]. The summation of most this work is implemented in the popular software package MrBayes [Ronquist and Huelsenbeck, 2003].

The key idea of MCMC is to use a very large statistical sample to approximate a multidimensional integral. We can estimate the posterior distribution using a sequence $\{s_1, s_2, ..., s_n\}$ of independently and identically distributed samples. The MCMC algorithm begins with a random sample s_1. The algorithm proposes the next sample s_2 based on a transition kernel $q(s_{i+1}|s_i)$. In some literature, this is called the jumping kernel or the proposal density. Given s_i, the transition kernel selects the next sample s_{i+1} based on a sliding window w around the current sample s_i. For example, in one implementation, s_{i+1} is a random value drawn from the interval $[s - \frac{w}{2}, s + \frac{w}{2}]$. The key idea of the value w is to control the size of steps taken during the MCMC run.

If the probability of the new proposal is greater than the probability of the old proposal – i.e. if $P(s_{i+1}|D) \geq P(s_i|D)$ – the MCMC algorithm accepts the new proposal. Otherwise, the new proposal s_{i+1} is accepted with probability α, where

$$\alpha = min\left(1, \frac{P(s_{i+1}|D)}{P(s_i|D)}\right) \tag{16}$$

We can visualize the MCMC algorithm traversing a topographic landscape. The peaks correspond to values for s which yield large posterior probabilities. The transition kernel will propose samples which sometimes lead uphill, and other times lead downhill. If the number of samples n is sufficiently large, then the MCMC algorithm will spend time at each sample s in proportion to its posterior probability.

Three requirements must be satisfied for the MCMC algorithm to work:

1. Irreducibility: the Markov chain must be able to reach all parts of the posterior distribution.

2. Recurrence: if the chain is run for infinite samples, then all parts of the posterior distribution must be reached infinitely often.

3. Convergence: the sample mean of the estimated posterior distribution should converge to the expected mean.

If a chain meets these requirements, then it will converge on a stationary distribution (i.e. the MCMC estimate of the true distribution). A chain is said to have good mixing time if it converges on its stationary distribution after a relatively short number of samples.

The basic MCMC assumes the transition kernel is symmetric. In other words, MCMC assumes $q(s_{i+1}|s_i) = q(s_i|s_{i+1})$. Hastings extended MCMC to use non-symmetric kernels [Hastings, 1970]. Geyer proposed another improvement: running several parallel MCMC chains with different sizes of sampling windows [Geyer, 1991]. Geyer's so-called Metropolis-coupled MCMC (or MC^3) gives better mixing time than standard MCMC. In addition to Geyer's approach, simulated annealing algorithms have also been used to improve MCMC mixing time [Marinari and Parisi, 1992, Geyer and Thompson, 1995]. Like MC^3, the simulated annealing approach runs several parallel chains. Unlike MC^3, simulated annealing MCMC randomly interchanges the chains.

1.6 Phylogenetic Uncertainty

If our phylogenetic inference method is statistically consistent and we have infinite data, then our maximum likelihood phylogeny is guaranteed to be the true phylogeny. In practice, our datasets are finite and we cannot be sure our best inferred phylogeny is indeed correct. It is therefore desirable to have some metric of statistical confidence in the accuracy of an inferred tree. Here we discuss two such measures: the maximum likelihood bootstrap proportion (BP), and the Bayesian posterior probability (PP).

We can formally think of the BP and the PP as testing the limit on accuracy of the claim that a given clade is nonmonophyletic [Alfaro and Holder, 2006]. In theory, a high BP or PP value is intended to be interpreted as a strongly-supported rejection of nonmonophyly. Although it is tempting to think of these values as probabilities that a clade is a real historic taxonomic grouping, Hillis and Bull observe that the BP (and by extension, the PP) do not account for biases in the method or the prior probabilities [Hillis and Bull, 1993].

The bootstrap method works by resampling with replacement a set of sites in the original sequence alignment. Each sample is used to construct a phylogeny. For x samples, we construct x phylogenies. The bootstrap proportion for a given clade is the fraction of the x phylogenies in which the clade appears. Efron and Gong introduced the boostrap [Efron, 1979, Efron and Gong, 1983] (see also [Diaconis and Efron, 1983]), and Felsenstein later popularized the BP for use with phylogenetic inference [Felsenstein, 1985].

Hillis and Bull used computational simulations and laboratory-generated phylogenies to asses the accuracy of the BP [Hillis and Bull, 1993]. Their results show BPs consistently underestimate the accuracy of clades.

In response to the bootstrap's conservative estimates, several corrected boostrap methods have been introduced: Rodrigo's calibrated bootstrap [Rodrigo, 1993], the complete-and-partial bootstrap [Zharkikh and Li, 1995], and a two-level bootstrap [Efron et al., 1996]. Efron observes that BPs provides "first-order" confidence limits on the accuracy of the clade, whereas corrected BPs provide "second-order" limits [Efron et al., 1996]. Corrected BPs are rarely used in published literature, which might be attributed their difficult software implementation and their slow runtimes. If we have n sites, m taxa, and c possible states, the uncorrected bootstrap is $O(mn^2c^2)$, while the corrected bootstrap is $O(mn^3c^2)$. If we ignore computational limitations, we could theoretically calculate a more accurate third, fourth, or n-th order approximation [Efron et al., 1996].

Whereas the boostrap method samples the original alignment to generate new data sets, Bayesian MCMC (as described in section 1.5 of this manuscript) keeps the alignment static and samples variants of parameter-space [Alfaro and Holder, 2006]. Although previous analyses claim the Bayesian PP overestimates the probability that a particular clade is correct [Simmons et al., 2004], recent analysis shows that Bayesian PPs can overestimate or underestimate the probability, depending on the branching patterns of the true tree [Kolaczkowski and Thornton, 2007].

Should evolutionary biologists use the BP or the PP? Simmons et al. describe a three-way split among phylogeneticists [Simmons et al., 2004]:

> Thus far, conclusions are split among the view that Bayesian support values are more reliable than the bootstrap as indicators that clades are correctly resolved (Wilcox et al. 2002; Alfaro, Zoller, and Lutzoni 2003), the opposite view (Suzuki, Glazko, and Nei 2002; Cummings et al. 2003), and the view that Bayesian values may form a reliable upper bound, whereas bootstrap values may form a more valid lower bound (Douady et al. 2003).

In fact, all three of these viewpoints are incorrect. Neither the BP or the PP is accurate, and the BP is not necessarily the upper bound on accuracy. Although the problem of measuring statistical confidence is unsolved, Kolaczkowski and Thornton assert that the best approach is to carefully apply several methods and then evaluate results in light of each method's statistical properties [Kolaczkowski and Thornton, 2007]. According to this sensible advice, the results of the BP and PP should be compared to each

other and to other methods, such as the approximate-likelihood ratio test [Anisimova and Gascuel, 2006].

Even if our measures of clade support were accurate, we face the problem of accommodating phylogenetic uncertainty in downstream analyses. Consider the following scenario. Suppose we have an aligned set of amino acid sequences. After running a Bayesian MCMC, our posterior distribution contains a tree t_1 with PP 0.6, another tree t_2 with PP 0.3, and trees t_3 through t_n whose PPs sum to 0.1. We know that evolution unfolded according to a single history, but here we observe a distribution of possible histories. Which phylogeny should we select as the "best?" According to a maximum likelihood framework, t_1 is the highest point estimate of the phylogeny and should therefore be selected. According to a Bayesian framework, the choice of t_1 ignores the 40% uncertainty embedded in the alternate trees. A true Bayesian would recommend an approach which somehow integrates this uncertainty.

1.7 Ancestral Reconstruction

Thus far, our discussion focused on phylogenetic inference. Here, I turn to the downstream topic of ancestral sequence reconstruction (ASR). Given an alignment of sequences, a phylogeny of those sequences, and a model of evolution, we would like to computationally infer ancestral sequences for internal nodes in the tree. Using the reconstructed sequences, we can chemically synthesize the ancestral molecules and then experimentally investigate their function. Pauling and Zuckerkandl originally proposed the idea of "resurrecting" ancestral molecules to test hypotheses about evolutionary history [Pauling and Zuckerkandl, 1963]. In the last decade, ASR has been used to investigate the evolution of elongation-factor proteins [Gaucher et al., 2003], steroid hormone receptors [Bridgham et al., 2006], and vertebrate rhodopsins [Chang et al., 2002] to name a few examples. For an introduction to ASR history and methods, see [Thornton, 2004].

In the early days of ASR, maximum-parsimony (MP) was the method du jour for reconstructing ancestral states: paleobiologists assigned states to ancestral nodes so as to minimize the number of state changes along the branches of the tree. Fitch developed the original MP ASR algorithm for rooted bifurcating trees [Fitch, 1971]; Hartigan later extended this algorithm for general trees [Hartigan, 1973]. (See also [Swofford and Maddison, 1987, Swofford and Maddison, 1992, Maddison and Maddison, 1992].) Among many examples, the MP method was used to reconstruct ancestral lysozymes [Malcolm et al., 1990], the mouse L1 protein [Adey et al., 1994], the bovid

ribonuclease [Stackhouse et al., 1990], and the artiodactyl ribonuclease [Jermann et al., 1995].

In the context of ASR, MP poses several problems. First, MP can yield several equally-best ancestral states at a given site, but provides no method for choosing the single-best state. This is troublesome if we are interested in chemically synthesizing ancestral molecules: the cost of manufacturing and investigating all the equally-best ancestral combinations can be prohibitively expensive. Second, when there exists asymmetry in transformation probabilities, MP can be systematically biased against changes from ancestral rare states to common extant states [Collins et al., 1994]. Third, MP can produce biased reconstructions when the rate of evolution is not constant across the phylogeny [Cunningham et al., 1998]. Finally, MP methods for ASR fail to incorporate information about branch lengths, mutation rates, or substitution rates. This means that a mutation from some state x to another state x' is equally likely over branch lengths of 0.01, 1.0, and 100.0 substitutions per site.

As an alternative to parsimony, Yang et al. proposed a likelihood-based ASR method which uses the same Markov models discussed in section 1.2 of this manuscript [Yang et al., 1995]. The likelihood-based algorithm makes two important assumptions. First, the evolutionary model is assumed to be symmetric, i.e. the probability of observing some state i mutate j is the same as observing state j mutate to i. Symmetric models makes the location of the tree's root unimportant and we can therefore perform ASR on unrooted trees. Second, we assume that each site evolves independently. With this assumption, the probability of observing an entire sequence alignment can be calculated as the product of probabilities for each site.

For simplicity, I begin by describing the ML ASR algorithm in terms of a single site. Let d be the observed data: the states for a single sequence site from several taxa. Let y be the vector of state assignments for all interior nodes of the phylogeny. Let $\hat{T}$ be the maximum likelihood phylogeny, and let M be our evolutionary model. Let $\hat{\theta}$ be the vector of maximum likelihood model parameters and branch lengths. Let $P(y|d, M, \hat{T}, \hat{\theta})$ be the conditional probability of assigning the state vector y to all interior nodes, given the data, the ML tree, the model, and the ML model parameters; $P(...)$ can be calculated as follows:

$$P(y|d, M, \hat{T}, \hat{\theta}) = \frac{L(d|y, M, \hat{T}, \hat{\theta})}{\sum_i L(d|y_i, M, \hat{T}, \hat{\theta})} \tag{17}$$

where $L(d|y, M, \hat{T}, \hat{\theta})$ is the likelihood of observing the data, given the set

of ancestral states specified by y, the model, the tree topology, and the model parameters. The denominator is the likelihood of the data summed over all possible ancestral state assignments. Whereas equation 17 considers only a single site, we now consider an alignment with n sites. The posterior probability of ancestral assignments for all internal nodes at all sites is calculated by multiplying the posterior probability of the ML ancestral assignment from each site, as follows:

$$\prod_{s=1}^{n} P(\hat{y_s}|d_s, M, \hat{T}, \hat{\theta}) = \frac{L(d_s|\hat{y_s}, M, \hat{T}, \hat{\theta})}{\sum_i L_k(d_s|y_{s_i}, M, \hat{T}, \hat{\theta})} \tag{18}$$

The likelihood-based ASR method articulated by Yang et al. is an emprical Bayesian (EB) approach; informally, EB methods are hybrids between maximum likelihood and Bayesian methods [Maritz, 1970]. EB methods are Bayesian because they include priors and calculate posterior probabilities. However, unlike a true Bayesian method, EB approaches fix some parameters at their maximum likelihood estimate. The EB method proposed by Yang et al. fixes the tree topology, branch lengths, and model parameters.

There exist two primary variants of ASR: marginal reconstructions and joint reconstructions. A marginal reconstruction estimates the state at a single ancestral node, integrating over all possible ancestral states at other nodes. A joint reconstruction (as shown in equation 18) estimates the states for all ancestral nodes in the tree. Koshi and Goldstein introduced a dynamic algorithm for marginal ancestral reconstruction [Koshi and Goldstein, 1996], and Pupko et al. introduced a dynamic algorithm for joint reconstruction [Pupko et al., 2000]. The dynamic versions of the joint and marginal reconstructions perform with equivalent computational complexity, scaling linearly with the number of taxa. Yang implemented both variants in the software package PAML [Yang, 1997, Yang, 2007].

Marginal and joint reconstructions can yield disagreeing ancestral reconstructions. Which method to use depends on the specific phylogenetic question being asked. For example, if we want to know about the variation of character state frequencies across the tree, then the joint reconstruction is appropriate. On the other hand, if we want to resurrect a specific ancestor, then a marginal reconstruction should be employed. Pupko et al. point out that "discrepancies [between the two types of ASR reconstructions] originate not from misuse of information, but from the difference in the nature of the probabilistic questions asked" [Pupko et al., 2000].

1.8 Uncertainty and Ancestral Reconstruction

Critics of maximum likelihood and emprical Bayesian methods assert that ML and EB approaches fail to account for uncertainty in the parameters which are estimated from the data. In response to this shortcoming, new ASR methods have been introduced to integrate uncertainty about the tree, the branch lengths, and the model parameters. Schultz and Churchill proposed a fully Bayesian (FB) method [Schultz and Churchill, 1999]; Huelsenbeck and Bollback proposed a hierarchial Bayesian (HB) method [Huelsenbeck and Bollback, 2001]. The respective analysis of FB and HB show that these methods – in some cases – yield posterior probability estimates which differ from previous ASR methods. That said, little (or nothing) has been published about their accuracy of FB or HB.

Whereas the FB and HB methods simultaneously consider uncertainty from several sources (the tree, the branch lengths, and the model parameters), a more desirable analysis would consider these sources individually and therefore isolate their effect on ASR. For each source of uncertainty, it would be useful to compare the accuracy of an ASR method which integrates that source of uncertainty to a method which uses the maximum likelihood estimate for that source. In this manuscript, I consider uncertainty about one such source: the phylogeny. Specifically, I address the following question: does integrating phylogenetic uncertainty affect the accuracy of reconstructed ancestral sequences?

In response to this question, the HB method provides one approach to integrate phylogenetic uncertainty into ASR, but only considers cases where the desired ancestor is clearly identifiable. The HB method assumes an a priori outgroup and ingroup for the desired ancestor; on each alternate tree, the set of terminal taxa in the outgroup is assumed to be completely disjoint from the set of terminal taxa in the ingroup. The HB method does not consider alternate phylogenies which violate the assumptions about outgroup and ingroup classification. Essentially, the HB approach assumes the desired ancestor exists, and only considers alternate trees in which this assumption is valid.

Pagel et al. critique the HB method and show that rejecting seemingly irrelevant trees introduces an ASR bias [Pagel et al., 2004]. Instead, Pagel et al. propose a *most-recent-common-ancestor* approach to select ancestral nodes from alternate trees:

> Given a set of species whose common ancestor's ancestral state is of interest, it is straightforward to find a node on each tree in

the MCMC sample that includes those species. In some trees the node will include only the species of interest, whereas in others it will include these species and others; but there will be a node to reconstruct in every tree.

In section 2 of this manuscript, I describe an ASR method which integrates phylogenetic uncertainty, based on the most-recent-common-ancestor approach. I use computational simulations to analyze the accuracy of this ASR method, and present results in section 3.

2 Methods

To investigate the effect of phylogenetic uncertainty on the accuracy of ancestral sequence reconstruction, I developed an empirical Bayesian method which integrates tree uncertainty (with much help from Joe Thornton and Bryan Kolaczkowski). Our method uses a *most-recent-common-ancestor* strategy [Pagel et al., 2004], in which we do not assume *a priori* the existence of the desired ancestor on every topology in a distribution of trees.

2.1 An Algorithm to Integrate Phylogenetic Uncertainty

Our method begins with a sequence alignment, a distribution of putative trees, and a model of character substitution. The distribution of trees can be inferred using MCMC, MC^3, or some other approach, but each tree must have an associated posterior probability. We use the substitution model and the PAML software suite to calculate the marginal ancestral reconstruction for all internal nodes on each tree. Given the output from PAML, the algorithm for integrating phylogenetic uncertainty follows:

1. **Initialization**

   Let T be a set of putative trees $\{t_1, t_2, ..., t_n\}$. Let $PP(t)$ be the posterior probability of some tree t, where $t \in T$. Let D_{in} be the ingroup: the set of descendants for which we desire the most-recent-common-shared ancestral sequence. It should be obvious that D_{in} is a subset of the taxa in the sequence alignment. Let D_{out} be the *proposed* outgroup: a set of descendants we think evolved from lineages more basil than the desired ancestor. Let m be the number of sites in the alignment. Finally, let ε be the set of characters in our state-alphabet (i.e. nucleotides or amino acids).

   Construct a two-dimensional matrix M with dimensions $m \times |\varepsilon|$. In other words: add one column for each site in the sequence alignment and one row for each character in the state-alphabet. For all sites i and states j, where $1 \leq i \leq m$ and $1 \leq j \leq |\varepsilon|$, initialize $M[i, j]$ to zero.

   Construct a three-dimensional matrix A with dimensions $|T| \times m \times |\varepsilon|$ (i.e. the number of putative trees, the number of sites in the alignment, the number of characters in the state-alphabet).

2. For each $t \in T$, find the most-common-shared-ancestor of the taxa in D_{in}. For each site s and each state c, assign to $A[t, s, c]$ the posterior

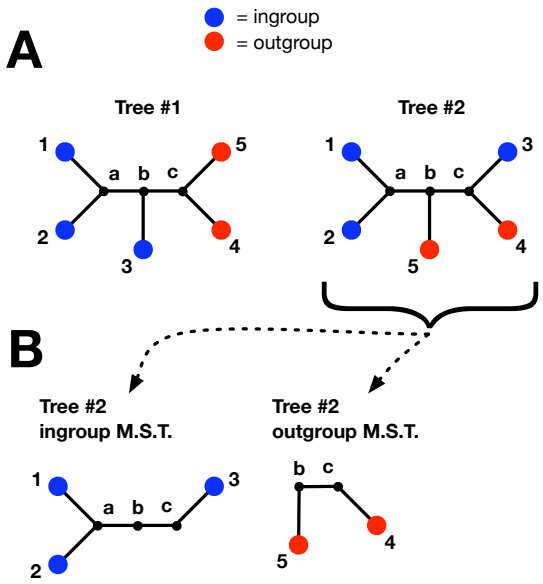


Figure 5: *Figure A shows two putative phylogenies for three ingroup taxa (blue dots) and two outgroup taxa (red dots). Although there exist fifteen possible unrooted phylogenies for five taxa, only two are shown here (for the sake of simplicity). On tree #1, the ingroup taxa form a monophyletic split and the most-common-shared-ancestor of the ingroup is easily identified as node b. On tree #2, the ingroup is non-monophyletic and this raises a phylogenetic dillema: the most-common-shared-ancestor of the ingroup is ambiguous. The algorithm presented in section 2.1 of this manuscript proposes a method to help select the ancestor. In figure B, tree #2 is decomposed into the minimum-spanning tree (MST) containing the ingroup and the MST containing the outgroup.*

probability of observing character c at site s on the most-common-shared ancestor on t. Because we are dealing with unrooted trees, it is possible that the most-common-shared-ancestor will be ambiguous. See figure 5 for an example. In case of ancestral ambiguity, do the following:

a. Find the minimum-spanning tree mst_{in} containing the nodes in D_{in}. Also find the minimum-spanning tree mst_{out} containing the nodes in D_{out}.

b. If mst_{in} and mst_{out} contain completely disjoint sets of nodes, then find the edge e which connects the two subtrees. The most-

common-shared-ancestor of D_{in} on t is the node which is directly connected to e and which is contained in the set D_{in}.

c. If mst_{in} and mst_{out} contain overlapping sets of nodes, then find the set of internal nodes N which are contained in the graph-union of mst_{in} and mst_{out}. Any node in N can plausibly be the most-common-shared-ancestor of D_{in}. This is an unresolvable ambiguity. In this case, my implementation randomly selects a node from N to be the most-common-shared ancestor on t.

3. Reduce the matrix A into the matrix M as follows. For each tree $t \in T$, for each site s $(1 \leq s \leq m)$, and for each state $c \in \varepsilon$:
$M[s, c] = M[s, c] + PP(t) \times A[t, s, c]$.

   In other words, weight the posterior probability of observing each state by the posterior probability of its tree. Sum the weighted state PPs.

4. **Termination**

   At this point, M contains the *maximum a posteriori* state distribution for the most-common-shared-ancestor of the taxa in D_{in}. Each column corresponds to a site, and each row corresponds to a possible character assignment. To retrieve the consensus ancestral sequence from M, select the character in each column with the highest posterior probability. In other words, for each site s, find the character c with the maximum value of $M[s, c]$.

I implemented the aforementioned algorithm using the languages Python and C, with MPI parallelization. The resultant software package is called *Lazarus* (formally known as *ART*).

Figure 6 illustrates a very simple demonstration of my algorithm. Here we observe four extant taxa, with amino acid sequences DD, NN, EN, and EE. We used the JTT model [Jones et al., 1991] and PAML [Yang, 2007] to score the likelihood of all fifteen possible phylogenies. We scaled the resultant likelihood scores into posterior probability values. Thirteen of the trees (not shown) have PPs near zero. Two of the trees (shown) have PPs 0.501 and 0.499. The two highest scoring trees tell conflicting evolutionary stories: tree #1 groups taxonB and taxonA as sister lineages, whereas tree #2 groups taxonC and taxonA as sisters. Suppose we want to reconstruct the most-common-shared ancestor of taxonA, taxonB, and taxonC. On tree #1, the ancestral sequence is uncertain: at both sites in the ancestor, the amino acids E and N are inferred with 0.375 posterior probability. On the other hand, the ancestor on tree #2 is inferred with overwhelming support

for E at site 1 and N at site 2. Using our experimental EB algorithm, we integrate the ancestral state distributions from the two trees and thus create a *maximum a posteriori* ancestral sequence. For example, the PP of assigning state E to the MAP ancestral site 1 is calculated as follows: $0.688 = [0.501 \times 0.375] + [0.499 \times 1.0]$.

The example in figure 6 demonstrates how integrating phylogenetic uncertainty can potentially influence ancestral reconstruction. Before the integration, we're forced to choose between two conflicting reconstructions. With phylogenetic integration, much of the previous ambiguity and uncertainty disappears. In this case, the MAP ancestor is EN, albeit not strongly-supported.

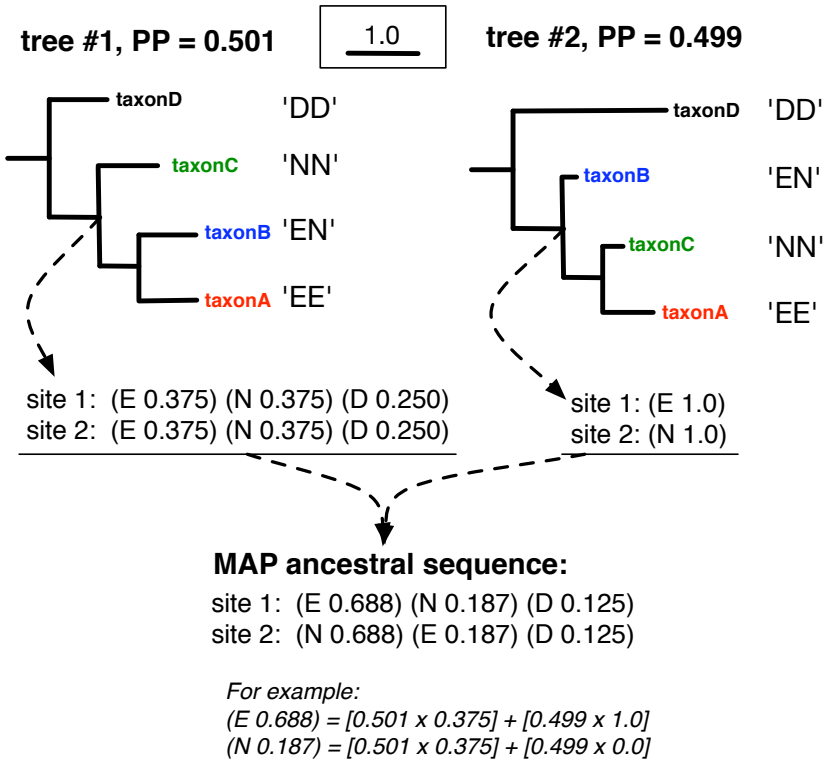


Figure 6: *Using two trees, with four extant taxa each, suppose we want to find the most-common-shared ancestral sequence for taxonA, taxonB, and taxonC. In this example, we calculate the maximum a posteriori (MAP) ancestral sequence as the weighted average of sequences from each tree. See the text in section 2.1 for further explanation.*

2.2 Simulations

I used simulations to compare the accuracy of our ASR method to the accuracy of a method which uses the ML estimate for the phylogeny. Simulations are necessary for measuring accuracy because we need experimental conditions in which the true phylogeny and the ancestral states on that phylogeny are known. My simulations follow these general steps:

1. Generate a rooted phylogeny t. Remember t as the true tree.
2. Generate a random ancestral sequence at the root of t.
3. Using a model of evolution, simulate the random ancestral sequence evolving down the branches of t into a set of terminal sequences. Record these intermediate ancestral sequences.
4. Collect the terminal sequences and arrange them into a sequence alignment.
5. Using MCMC, MC^3, or some other method, infer a distribution of putative phylogenies T from the sequence alignment. Let the tree $\hat{t}$ be the phylogeny with the largest posterior probability.
6. Identify a subset of terminal nodes – an ingroup – for which we would like to infer the most-common-shared-ancestor.
7. Using the algorithm described in section 2.1, reconstruct the *maximum a posteriori* ancestral sequence on the distribution of trees T.
8. Using an ASR algorithm which does not integrate phylogenetic uncertainty, reconstruct the ancestral sequence on tree $\hat{t}$.
9. Using the recorded ancestral sequences (i.e. the true sequences) from step 3, compare the accuracy of the *maximum a posteriori* ancestral sequence to the accuracy of the ancestral sequence on $\hat{t}$.

For simplicity of notation, I will refer to my experimental ASR method as the *EB method*; I will refer to the method of Yang et al. (which ignores phylogenetic uncertainty) as the *ML method* [Yang et al., 1995].

My first set of simulations addresses the question, "how does the accuracy of the EB and ML methods compare as we vary the amount of phylogenetic uncertainty?" In general, phylogenies with short internal branches are more difficult to accurately infer than phylogenies with longer internal

branches [citation]. Therefore, the internal branch length of a tree is a proxy for its phylogenetic uncertainty.

In the first simulation, I evolved an amino acid sequence over a four-taxa phylogeny of variable height and variable internal branch length as shown in figure 7. The topology includes two evolutionary forks, yielding a total of four descendant sequences. Although I could have considered a tree with a single fork (and therefore three descendants), the resultant tree-space would include a single topology; a comparison of the ML and EB methods would be futile because the EB method would not have a distribution of trees over which to integrate.

Using the tree in figure 7, my experimental strategy was to vary the amount of phylogenetic uncertainty in the clade formed by taxon A, B, and C. I varied uncertainty using a combination of two factors, the length of r and the length of h. The true phylogeny is more difficult to infer when the internal branch length r is short, or when the distance h between the ancestral and descendant sequences increases. Thus, I varied the branch length r from 0.01 to 0.200 substitutions per site and I varied the branch length labeled h from 0.250 to 0.750 substitutions per site. I fixed taxon D as an outgroup with branch length 0.75 substitutions per site.

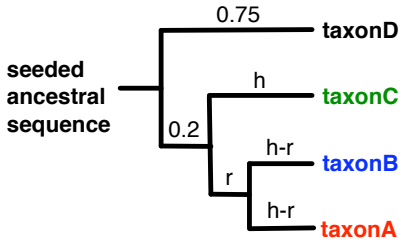


Figure 7: *The four-taxa simulation uses a phylogeny with variable branch lengths h and r. The true phylogeny is more difficult to infer when (1) r is short, (2) h is long, or a combination of (1) and (2).*

At each combination of r and h, I simulated 100 replicates (where each replicate is a set of descendants). To simulate a replicate, I seeded the most-common-shared ancestor of the entire rooted tree with a randomally generated 400 amino acid sequence. This ancestor is labeled *seeded ancestral sequence* in figure 7. Using the Seq-Gen software suite [Rambaut and Grassly, 1997] and the JTT model of evolution [Jones et al., 1991], I simulated the seeded sequence evolving into four terminal sequences: taxon A, B, C, and D.

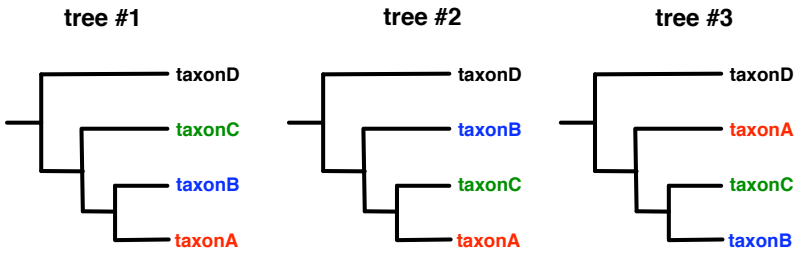


Figure 8: *After simulating descendant sequences on the tree shown in figure 7, I fixed taxon D as the outgroup and considered three possible phylogenies shown here.*

For each replicate, I used *Lazarus* (which uses PAML) to reconstruct the marginal posterior distribution of ancestral states at every internal node on each of the three trees shown in figure 8. I also used Lazarus to reconstruct the EB ancestral sequence – integrating over all three trees – for the most-common-shared ancestor of taxa $\{A, B, C\}$, of taxa $\{A, B\}$, of taxa $\{A, C\}$, and of $\{B, C\}$. Depending on the tree, some of these ancestors are the same. For example, on the tree #1 in figure 8, the most-common-shared ancestor of $\{B, C\}$ is equivalent to the most-common-shared ancestor of $\{A, B, C\}$. For another example, consider the incorrect topology on tree #2: the most-common-shared ancestor of $\{A, C\}$ incorrectly excludes taxon B. Conversely, the most-common-shared ancestor of $\{A, B\}$ on tree #2 spuriously includes taxon C.

My second set of simulations compared the accuracy of the ML and EB methods in more realistic – empirically derived – phylogenetic conditions. Here, I used a phylogeny inferred from an alignment of alcohol dehydrogenase sequences [Thomson et al., 2005]. Using the tree shown in figure 9, I seeded 100 random ancestral sequences at the root node. Each random sequence was 400 amino acids long. I used Seq-Gen [Rambaut and Grassly, 1997] and the JTT model [Jones et al., 1991] to simulate the seeded ancestors evolving across the tree into descendant sequences. Using a technique described by Kolaczkowski et al. [citation for Bryan's EB method?], I used Bayesian Markov Chains to discover a distribution of putative phylogenies and to estimate the posterior probability of each internal node. I used Lazarus to reconstruct the ML and EB ancestral sequences for a large collection of internal nodes, including a mix of uncertain nodes (with PP $<$ 1.0) and certain nodes (with PP $=$ 1.0). Figure 9 shows the reconstructed nodes in relation to the overall phylogeny.

2.3 Criteria for Grouping the Results

I grouped all ancestral inferences from the four-taxon simulation according to two criteria. First, I sorted according to the descendant state pattern of taxa A, B, and C on the phylogeny shown in figure 7. Patterns here include *xxy*, *xyz*, *xxx*, etc. For example if taxa A and B have the state 'p' for the amino acid *proline*, and taxa C has the state 'l' for *leucine*, then we observe pattern *xxy*. I sorted according to this criteria because the difficulty of inferring an ancestral state depends on how well the state has been conserved over evolutionary time.

The second grouping criteria sorts replicates according to the descendant membership of the ancestor. Here I group ancestral inferences into three bins: (i) the membership is correct, (ii) the membership spuriously includes an extra taxon, and (iii) the membership incorrectly excludes a taxon. As an example, consider the case where the true tree is (((A,B),C),D), but the maximum likelihood tree is (((B,C),A),D). Using the ML tree, our inference of the most-common-shared-ancestor of taxa $\{B, C\}$ will incorrectly exclude taxon A (bin iii). Furthermore, our inference of the last-shared ancestor of $\{A, B\}$ will spuriously include taxon C (bin ii).

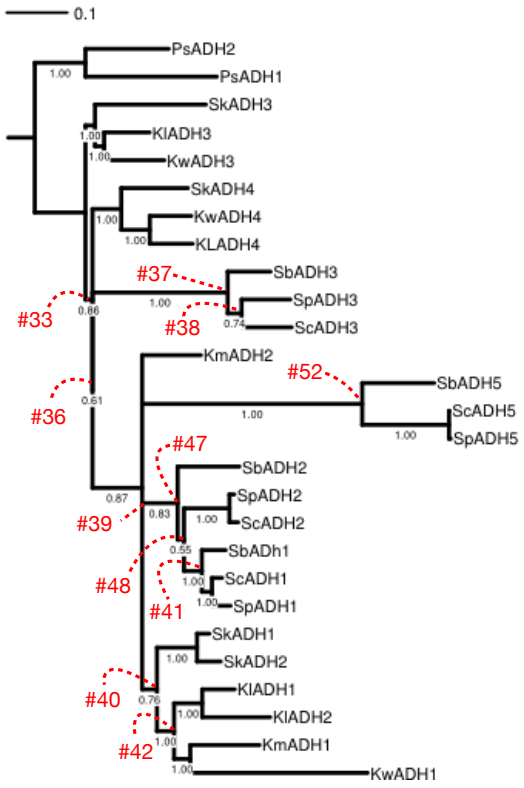


Figure 9: *A phylogeny of alcohol dehydrogenases. The terminal labels correspond to the published taxon labels [Thomson et al., 2005]. I used this tree to simulate the evolution of 400 amino acid sequences. I used the resultant descendants to infer a distribution of putative trees and reconstruct the ancestors labeled in red.*

3 Results

3.1 The ML and EB methods infer the same state at almost all ancestral sites

In both my simulations, the ML and EB methods reconstructed the same state at more than 99% of sites (see Table 1). This observation appears to be true regardless of the descendant state pattern and the membership of the descendant clade. The ML and EB reconstructions are especially similar in situations when the ancestral state is well conserved (such as state pattern xxx), and also when the taxa membership of the descendant clade is correct. On the other hand, the ML and EB reconstructions are slightly more divergent in situations when the membership of the descendant clade spuriously contains an extra taxon, and also when the ancestral state is not well conserved.

Indeed, a method for ASR which integrates phylogenetic uncertainty rarely reveals a novel ancestral state which was not already revealed as a plausible state by the ML distribution. In the four-taxon simulation, the EB state assignment appears as a plausible state within the ML distribution – with posterior probability greater than 0.2 – at more than 99.8% of sites (see table 2).

extant state pattern	clade correct	mem. +	mem. -	all
all	0.9968	0.9933	0.9951	0.9961
xxx	0.9975	0.9949	0.9960	0.9969
xxy	0.9971	0.9940	0.9951	0.9964
xyx	0.9960	0.9939	0.9950	0.9955
yxx	0.9961	0.9936	0.9953	0.9955
xyz	0.9952	0.9926	0.9943	0.9946

Table 1: The proportion of ancestral sites at which the EB and ML methods assigned the same state. *For every replicate, I sorted the ancestral state inferences according to two criteria: (1) the descendant state pattern of taxa A, B, and C (where patterns include **xxy**, **xyz**, etc.) and (2) the membership of the descendant clade (where **clade correct** means that the ancestor was inferred using all the correct ingroup taxa, **mem +** indicates the descendant membership spuriously included an extra taxon; **mem -** means the membership omitted a taxon that should have been included). Each cell reports the proportion of ancestral sites at which the EB and ML methods assigned the same state. Each row corresponds to a unique state pattern; the top row expresses data counted over all descendant state patterns. Each column corresponds to a membership pattern; the right-most column corresponds to data counted over across all membership patterns.*

extant state pattern	clade correct	mem. +	mem. -	all
all	0.0010	0.0011	0.0019	0.0011
xxx	0.0006	0.0008	0.0007	0.0007
xxy	0.0009	0.0012	0.0013	0.0010
xyx	0.0011	0.0011	0.0011	0.0011
yxx	0.0012	0.0012	0.0011	0.0012
xyz	0.0016	0.0016	0.0015	0.0016

Table 2: The proportion of ancestral sites at which the maximum a posteriori EB state is different from the ML state and is not found as an alternate (with PP ≥ 0.2) within the ML state distribution. *I sorted the ancestral state inferences from every replicate according to the same criteria in table 1. Each cell reports the proportion of ancestral sites at which the most likely EB state assignment is not found within the ML distribution of state assignments (with posterior probability greater than 0.2). The top row expresses the proportion across all descendant state patterns. The right-most column express the proportion across all membership patterns.*

3.2 The ML and EB methods infer disagreeing states at poorly-supported sites on poorly-supported trees.

Although the ML and EB methods infer the same state at more than 99.8% of sites, I wanted to determine what conditions cause the methods to disagree at the remaining 0.2% of sites.

The ML and EB reconstructions disagree only at sites which are already ambiguous (see figure 10). In other words, the two methods agree about ancestral states with strong support. On the four-taxon simulation, the ML and EB reconstructions agree at sites with posterior probability greater than 70%. On the ADH simulation, two methods agree at sites with PP greater than 76%.

The ML and EB reconstructions disagree more often in phylogenetically uncertain situations (see figure 11). When the posterior distribution of trees is peaked – such that the ML tree has posterior probability 1.0 – the ML and EB reconstructions do not differ because there is no phylogenetic uncertainty over which to integrate. On the other hand, when the posterior distribution of trees is less-peaked, the ML and EB methods are able to reconstruct different states because there exists more phylogenetic uncertainty over which to integrate.

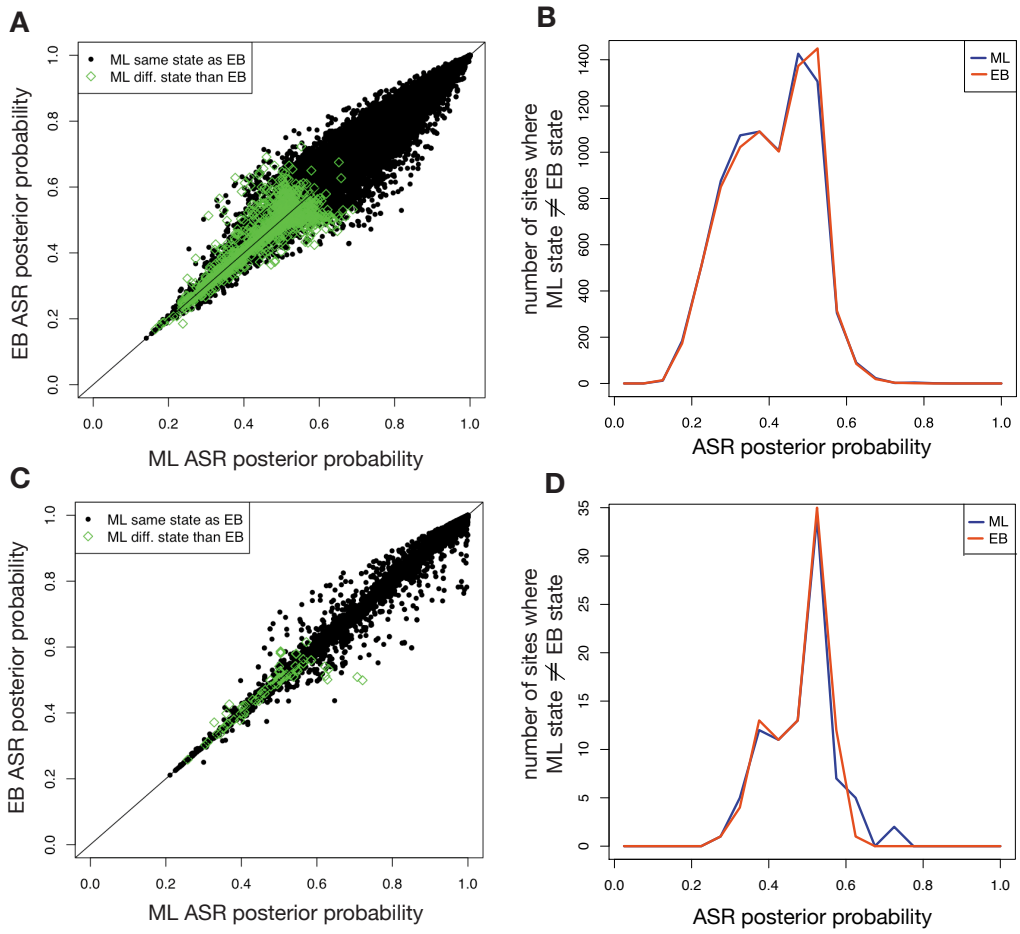


Figure 10: How similar are the EB and ML reconstructions? *I plotted the ML posterior probability of every reconstructed ancestral site versus its corresponding EB posterior probability; figure A shows results from the four-taxon simulation, figure C shows results from the ADH simulation. Sites at which ML and EB yield disagreeing states are highlighted in green. I also plotted the frequency of sites at which the ML and EB reconstructions disagree – i.e., the number of green points; figure B shows results from the four-taxon simulation, figure D shows results form the ADH simulation. In B and D, each ASR inference was binned into 5% groups according to its posterior probability.*

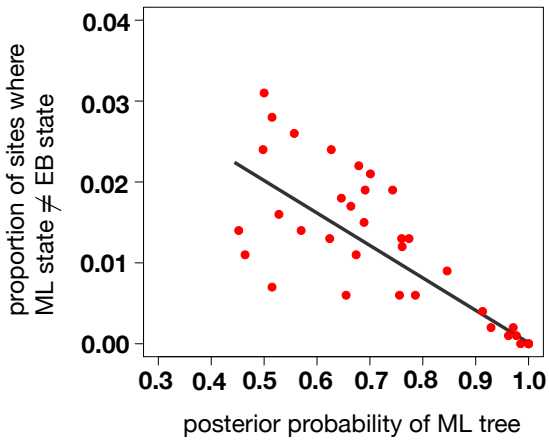


Figure 11: The proportion of sites where the ML reconstruction is different from the EB reconstruction. *Here, each red point represents a group of replicates with the same r and h value (see figure 7 in the methods section for a description of r and h). The horizontal position of each point represents the average posterior probability of the replicates' ML trees; the vertical position represents the fraction of ancestral sites with disagreeing ML and EB state assignments.*

3.3 When ML and EB reconstructions differ, EB is usually less accurate

Although the ML and EB reconstructions are rarely different, the ML method is overall more accurate than the EB method (see table 3.3). ML is especially more accurate than EB when the membership of the descendant clade is correct. In this case, the ML tree is the true tree: integrating phylogenetic uncertainty only serves to introduce error.

The ML reconstruction is slightly less accurate than the EB reconstruction in the rare case that the descendant membership spuriously includes an extra taxon and the ancestral state is poorly conserved among the terminal branches (state pattern xyz). In this case, we have very little information from which to make an accurate ancestral reconstruction; the ML method uses the branch lengths and the model to essentially guess between states x, y, and z. On the other hand, the EB method integrates reconstructions from alternate trees which do not necessarily include the spurious terminal branch with state z. Therefore, the EB method is slightly more accurate in this tough situation because integrating over trees helps to eliminate state z from the set of possible ancestral states.

extant state pattern	**clade correct**	**mem. +**	**mem. -**	**all**
all: ML	**0.871**	0.804	**0.800**	**0.854**
all: EB	0.859	**0.807**	0.798	0.845
xxx: ML	0.990	0.996	0.996	0.991
xxx: EB	0.990	0.996	0.996	0.991
xxy: ML	**0.852**	0.877	0.823	**0.852**
xxy: EB	0.809	0.877	0.823	0.818
xyx: ML	0.889	0.898	0.898	0.892
xyx: EB	0.889	0.898	0.898	0.892
yxx: ML	**0.902**	0.840	0.849	**0.886**
yxx: EB	0.900	0.840	0.849	0.885
xyz: ML	**0.619**	0.564	**0.574**	**0.602**
xyz: EB	0.596	**0.574**	0.567	0.588

Table 3: The proportion of accurately reconstructed sites. *I sorted the ancestral state inferences from every replicate according to the same criteria in table 1. Here, each cell reports two values: (top) the proportion of sites accurately reconstructed by the ML method and (bottom) the proportion accurately reconstructed by the EB method.*

3.4 The ML and EB methods yield slightly different posterior probability values, but the overall accuracy of those values is nearly the same

Although the ML and EB methods usually infer the same ancestral state, the two methods do not necessarily support those states with the same posterior probability support values. Overall, EB posterior probabilities are are slightly less accurate than ML support values. On the four-taxon simulation, EB PPs are slightly inflated for PPs greater than 0.5 (see plot 12A). On the ADH simulation, the ML and EB methods appear to produce PPs with similar – if not identical – average accuracy (see plot 12B). According to a chi-square test, there is a small advantage to EB PPs when the ML tree is incorrect; there is a bigger advantage to ML PPs when the ML tree is correct (see table 4). The latter case is more frequent, and the ML PPs are therefore overall more accurate than EB PPs. This observation is corroborated by the ADH simulation: the ML PPs are more accurate than EB PPs on nodes which are strongly-supported to be true (see table 5).

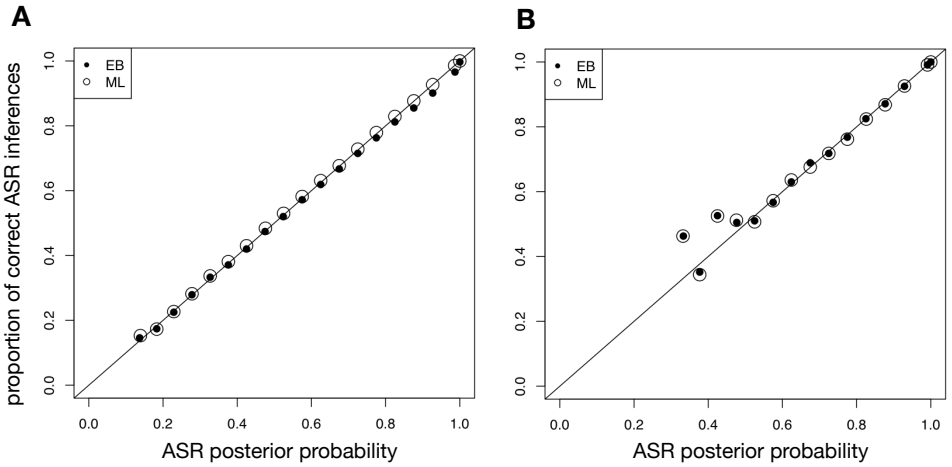


Figure 12: The accuracy of ASR posterior probability values. *I grouped all ancestral state inferences – according to their posterior probability support value – into 5%-sized bins. Specifically, all sites with PP ranging from 0.00% to 4.99% were grouped into one bin, all inferences with support ranging from 5.00% to 9.99% were grouped into another bin, and so on. I excluded bins with fewer than fifty members. Within each bin, I kept the ML inferences segregated from the EB inferences. I counted the fraction of inferences which correctly inferred the true state. If PP was a perfect measure of accuracy proportions, then we would expect the fraction of correct inferences in each bin to equal the mean PP of the inferences in that bin. For example, in our four-taxon simulation, the average ML inference in the bin ranging from 0.900% to 0.949% has posterior probability 0.927. Therefore, we expect 92.7% of the ML state inferences within this bin to be accurate. However, we observe that 92.6% of sites in this bin are correct, indicating that the ML method is slightly overconfident with support values in this range from 0.900 to 0.949. The proportion of correct inferences was counted for each bin.* ***Figure A*** *plots the proportion of correct inferences for the four-taxa simulation;* ***figure B*** *plots these proportions for the ADH simulation.*

extant state pattern	clade correct	mem. +	mem. -	all
all: ML	**10.201**	4.823	3.344	**9.439**
all: EB	17.762	**2.191**	**2.532**	14.896
xxx: ML	**0.051**	0.002	0.001	**0.037**
xxx: EB	0.275	**0.001**	0.001	0.141
xxy: ML	**2.014**	0.081	0.081	**3.022**
xxy: EB	14.210	**0.013**	**0.024**	17.835
xyx: ML	1.809	**0.209**	0.464	**1.676**
xyx: EB	**1.565**	0.464	**0.274**	2.109
yxx: ML	2.645	**0.062**	0.214	**4.766**
yxx: EB	**2.630**	0.081	**0.205**	5.355
xyz: ML	**10.347**	4.475	**2.333**	**7.839**
xyz: EB	19.703	**1.597**	2.568	15.317

Table 4: χ^2 statistics for the four-taxon simulation. *I calculated a weighted chi-square statistic to measure the fit between the function $f(x) = y$ and the points in figure 12. The chi-square calculation is weighted because the bins (along the X axis) each contain different numbers of inferences; some bins contain more than 10,000 state predictions, while other bins contain less than 100 predictions. I calculated the weighted chi-square statistic as follows: $\chi^2 = \sum_{i=1}^{n} \frac{B_i(O_i - E_i)^2}{E_i}$, where n is the number of bins, B_i is the number of inferences within bin i, O_i is the observed proportion of correct inferences for bin i, and E_i is the expected proportion of correct inferences for bin i. Lower χ^2 scores correspond to more accurate posterior probability values. In this table, I sorted the ancestral state inferences from every replicate according to the same criteria in table 1. The top row expresses χ^2 values across all descendant state patterns. The right-most column express χ^2 values across all membership patterns.*

ADH node	χ^2	**PP(node)**
33: ML	6.939	0.86
33: EB	**2.989**	
36: ML	**2.536**	0.61
36: EB	3.463	
38: ML	2.477	0.74
38: EB	**2.063**	
39: ML	**2.827**	0.87
39: EB	2.921	
40: ML	1.909	0.76
40: EB	**1.385**	
47: ML	**6.129**	0.83
47: EB	6.137	
48: ML	4.496	0.55
48: EB	**3.377**	
37: ML	**7.682**	1.00
37: EB	8.871	
41: ML	**3.606**	1.00
41: EB	3.725	
42: ML	**3.989**	1.00
42: EB	4.281	
52: ML	**7.255**	1.00
52: EB	7.644	

Table 5: χ^2 statistics for the ADH simulation. *I calculated χ^2 values – as described for table 4 – for the state assignments at ancestral nodes in the ADH phylogeny. Lower χ^2 scores correspond to more accurate ASR posterior probability values. Unlike previous tables in this manuscript, the χ^2 values reported here are for all descendant state patterns and descendant membership patterns. The left-most column lists node numbers corresponding to phylogenetic labels in figure 9. The right-most column lists the posterior probability (PP) of the corresponding node.*

3.5 Phylogenetic uncertainty is not correlated with ASR accuracy

It might seem paradoxical that integrating phylogenetic uncertainty yields an ancestral reconstruction which is nearly identical to the ML reconstruction. Here, I explain this nonintuitive result.

Phylogenetic uncertainty correlates with internal branch length (see figure 13A), but internal branch length does not correlate with ASR accuracy (see figure 13B). Therefore, phylogenetic uncertainty does not correlate with ASR accuracy. This result can be understood in light of previous results which show that ASR is more accurate on star-like trees (i.e. trees with short or zero-length internal branches) [Blanchette et al., 2004, Lucena and Haussler, 2005]. Although ASR is more accurate on star-like trees, star-like trees are less accurately inferred (see figure 13A). It seems that a trade-off exists between phylogenetic accuracy and ASR accuracy. In my four-taxon simulation, as I increased the internal branch length, I created a situation in which we more easily identified the true phylogeny but less easily identified the true ancestral state. Conversely, as I decreased the internal branch length, I created a situation where we can less easily identified the true phylogeny but more easily identified the ancestral state.

The conditions that produce phylogenetic uncertainty result in ancestral states on the alternate trees which are very similar – and often identical – to the ancestral states on the most probable tree. When the internal branch is short, the evolutionary distance between the true ancestor and the incorrectly inferred ancestor is small. As an illustration of this phenomenon, consider figure 14. Suppose we want to infer the last-shared ancestor of taxa A and B. On the true tree, taxa A and B are sisters, neighbored with the lineage to taxon C. However, in this example, the shortness of the branch connecting the most-common-shared-ancestor (MCSA) of $\{AB\}$ to the ancestor of $\{ABC\}$ causes uncertainty in our phylogenetic inference. Consequently, our maximum likelihood tree incorrectly joins taxa B and C as sisters. Although our ML reconstruction of MCSA(AB) spuriously includes taxon C, the accuracy of the reconstruction will not suffer for this error: the distance between the true ancestor and our incorrect ancestor is short. This example is relevant to our four-taxa simulation, where the distance between MCSA(AB) and MCSA(ABC) is exactly r. The resultant error from including taxon C will be proportional to r. When r is very short (for instance, 0.01 substitutions per site) and our ancestral sequence is only 400 sites long, it is possible that MCSA(AB) will be identical to the MCSA(ABC).

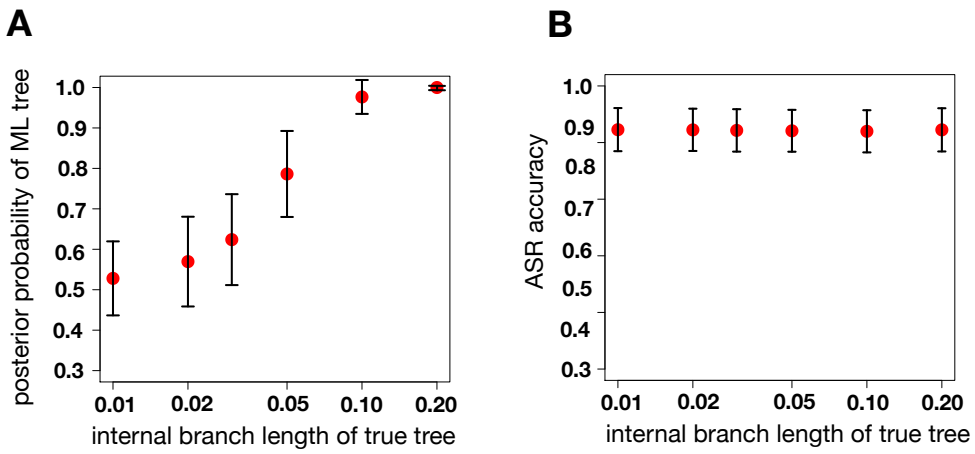


Figure 13: Phylogenetic uncertainty is not correlated with ASR accuracy. *Each point represents replicates from our four-taxon simulation, grouped according to the length of their internal branch. Figure A plots the Bayesian PP of the true phylogeny as a proxy for phylogenetic certainty. Figure B plots the fraction of sites which were correctly inferred for all ancestral reconstructions for replicates with the given internal branch length.*

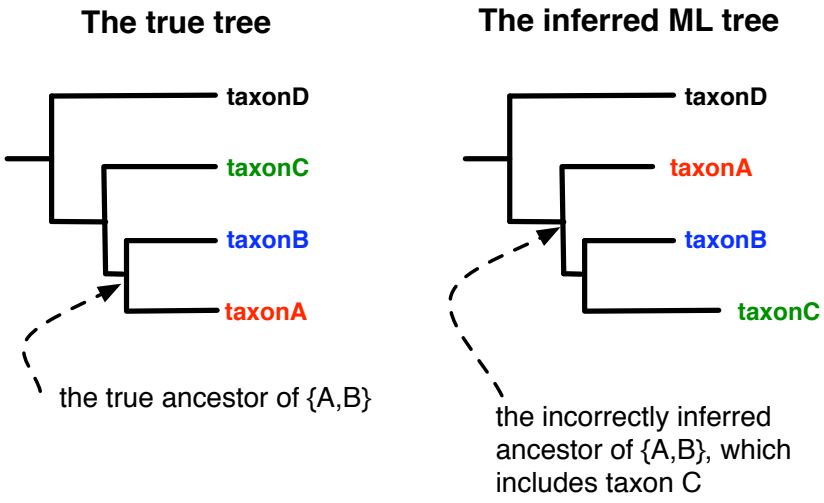


Figure 14: An example in which phylogenetic error is not significantly deleterious to ASR accuracy. *In this example, suppose we want the most-common-shared ancestor of taxonA and taxonB. Also suppose our ML tree is incorrect, and consequently we spuriously include taxon C in our reconstruction. The distance between our incorrect ancestor and the true ancestor is short. Therefore, our phylogenetic error will not seriously impact the accuracy of the reconstruction.*

4 Conclusions

In many cases, there exists uncertainty about the true phylogeny for an alignment of molecular sequences. In this project, my analysis showed phylogenetic uncertainty is not correlated with the accuracy of reconstructed ancestral sequences. The conditions which produce phylogenetic uncertainty result in ancestral sequences on alternate trees which are similar (and often identical) to the ancestral sequence on the maximum likelihood tree. Consequently, I claim integrating phylogenetic uncertainty does not significantly affect the accuracy of reconstructed ancestral sequences.

My result should not be interpreted as an endorsement for sloppy phylogenetics. ASR practitioners should continue to use rigorous statistical practices, including model-fitting methods, to infer the best possible phylogeny for their data. That said, integrating phylogenetic uncertainty into reconstructed ancestral sequences is computationally demanding and statistically unnecessary.

References

[Adachi and Hasegawa, 1996] Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Biology*, 42(4):459–468.

[Adachi et al., 2000] Adachi, J., Waddell, P. J., Martin, W., and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. *Journal Molecular Evolution*, 50:348–358.

[Adey et al., 1994] Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H., and III, C. A. H. (1994). Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proceedings of National Academy of Science*, 91:1596–1573.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281.

[Aldrich, 2007] Aldrich, J. (2007). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176.

[Alfaro and Holder, 2006] Alfaro, M. E. and Holder, M. T. (2006). The posterior and prior in bayesian phylogenetics. *The Annual Review of Ecology, Evolution, and Systematics*, 37:19–42.

[Anisimova and Gascuel, 2006] Anisimova, M. and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 4:539–552.

[Armougom et al., 2006] Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006). Espresso: automatic incorporation of structural information in multiple sequence alignments using 3d-coffee. *Nucleic Acids Research*, 34:W604–W608.

[Azencott, 1992] Azencott, R. (1992). *Simulated Annealing: Parallelization Techniques*. John Wiley and Sons.

[Bairoch and Boeckmann, 1992] Bairoch, A. and Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucleic Acids Research*, 20:2019–2022.

[Baltimore, 1970] Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumor viruses. *Nature*, 226:1209–1211.

[Blanchette et al., 2004] Blanchette, M., Green, E. D., Miller, W., and Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Research*, 14:2412–2423.

[Brauer et al., 2002] Brauer, M. J., Holder, M. T., Dries, L. A., Zwickl, D. J., Lewis, P. O., , and Hillis, D. M. (2002). Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. *Molecular Biology and Evolution*, 19(10):1717–1726.

[Brent, 1972] Brent, R. P. (1972). *Algorithms for Minimisation without derivatives (automatic computation)*. Prentice Hall.

[Bridgham et al., 2006] Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 307:97–101.

[Bryant et al., 2005] Bryant, D., Galtier, N., and Poursat, M.-A. (2005). *Mathematics of Evolution and Phylogeny*, chapter 2, pages 33–62. Oxford University Press.

[Burks et al., 1992] Burks, C., Cinkosky, M. J., M.Fischer, W., Gilna, P., E.-D.Hayden, J., M.Keen, G., Kelly, M., Kristofferson, D., and Lawrence, J. (1992). Genbank. *Nucleic Acids Research*, 20:2065–2069.

[Camin and Sokal, 1965] Camin, J. and Sokal, R. (1965). A method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326.

[Cavalli-Sforza and Edwards, 1967] Cavalli-Sforza, L. and Edwards, A. (1967). Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, 19(3):233–257.

[Cedric Notredame, 2000] Cedric Notredame, Desmond G. Higgins, J. H. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

[Chang et al., 2002] Chang, B. S. W., Jonsson, K., Kazmi, M. A., Donoghue, M. J., and Sakmar, T. P. (2002). Recreating a functional ancestral archosaur visual pigment. *Molecular Biology and Evolution*, 19(9):1483–1489.

[Chenna et al., 2003] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Research*, 31(13):3497–3500.

[Collins et al., 1994] Collins, T. M., Wimberger, P. H., and Naylor, G. J. P. (1994). Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Systematic Biology*, 43(4):482–496.

[Cunningham et al., 1998] Cunningham, C. W., Omland, K. E., and Oakley, T. H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution*, 13(9):361–366.

[Dantzig, 1963] Dantzig, G. (1963). *Linear Programming and Extensions*. Princeton University Press.

[Dayhoff et al., 1978] Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352.

[Diaconis and Efron, 1983] Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 249:116–130.

[Do et al., 2005] Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005). Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340.

[Edgar, 2004] Edgar, R. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

[Edgar and Batzoglou, 2006] Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, 16:368–373.

[Edwards and Cavalli-Sforza, 1963] Edwards, A. and Cavalli-Sforza, L. (1963). The reconstruction of evolution. *Annals of Human Genetics*, 27:105–106.

[Efron, 1979] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26.

[Efron and Gong, 1983] Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jacknife, and cross-validation. *The America Statistician*, 37:36–48.

[Efron et al., 1996] Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of National Academy of Science*, 93(14):7085–7090.

[Farris, 1970] Farris, J. S. (1970). Methods for computer Wagner trees. *Systematic Zoology*, 19(1):83–92.

[Farris, 1977] Farris, J. S. (1977). Phylogenetic analysis under dollo's law. *Systematic Zoology*, 26:77–88.

[Felsenstein, 1978] Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 227:401–410.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

[Felsenstein, 1985] Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791.

[Felsenstein, 2004] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinaur Associates, Inc.

[Fenchel, 2002] Fenchel, T. (2002). *The Origin and Early Evolution of Life*. Oxford University Press, illustrated edition.

[Fischer, 1922] Fischer, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368.

[Fitch, 1971] Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416.

[Fitch and Margoliash, 1967] Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760):279–284.

[Gaucher et al., 2003] Gaucher, E. A., Thomson, J. M., Burgan, M. F., and Benner, S. A. (2003). Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, 425(18):285 – 288.

[Geyer, 1991] Geyer, C. J. (1991). Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. Technical Report 568, School of Statistics, University of Minnesota.

[Geyer and Thompson, 1995] Geyer, C. J. and Thompson, E. A. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of American Statistical Association*, 90(431):909–920.

[Golub and Loan, 1996] Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. The Johns Hopkins University Press, 3 edition.

[Guindon and Gascuel, 2003] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

[Hartigan, 1973] Hartigan, J. A. (1973). Minimum mutation fits to a given tree. *Biometrics*, 29:53–65.

[Hasegawa et al., 1985] Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal Molecular Evolution*, 21:160–174.

[Hastings, 1970] Hastings, W. (1970). Monte carlo samping methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

[Hillis and Bull, 1993] Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2):182–192.

[Huelsenbeck and Bollback, 2001] Huelsenbeck, J. P. and Bollback, J. P. (2001). Empirical and heirarchical bayesian estimation of ancestral states. *Systematic Biology*, 50(3):351–366.

[Jermann et al., 1995] Jermann, T. M., Opitz, J. G., Stackhouse, J., and Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, 374:57–59.

[Jones et al., 1991] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1991). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282.

[Jukes and Cantor, 1969] Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In Munro, M., editor, *Mammalian Protein Metabolism, Volume III*, pages 21–132. Academic Press, New York.

[Kidd and Sgaramella-Zonta, 1971] Kidd, K. and Sgaramella-Zonta, L. (1971). Phylogenetic analysis: concepts and methods. *American Journal of Human Genetics*, 23(3):235–252.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

[Kluge and Farris, 1969] Kluge, A. G. and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, 18(1):1–32.

[Knoll, 2004] Knoll, A. H. (2004). *Life on a Young Planet*. Princeton University Press, 2004, illustrated edition.

[Kolaczkowski and Thornton, 2007] Kolaczkowski, B. and Thornton, J. W. (2007). Effects of branch length uncertainty on bayesian posterior probabilities for phylogenetic hypotheses. *Molecular Biology and Evolution*, 24(9):2108–2118.

[Kolaczkowski and Thornton, 2008] Kolaczkowski, B. and Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, 25(6):1054–1066.

[Koshi and Goldstein, 1996] Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal Molecular Evolution*, 42:313–320.

[Krishnan et al., 2004] Krishnan, N. M., Seligmann, H., Stewart, C.-B., de Koning, A. J., and Pollock, D. D. (2004). Ancestral sequence reconstruction in primate mitochondrial dna: Compositional bias and effect on functional inference. *Molecular Biology and Evolution*, 21(10):1871–1883.

[Larget and Simon, 1999] Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16:750–759.

[Lewis, 1998] Lewis, P. O. (1998). A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution*, 15(3):277–283.

[Li, 1996] Li, S. (1996). *Phylogenetic tree construction using Markov Chain Monte Carlo*. PhD thesis, Ohio State University, Columbus.

[Lucena and Haussler, 2005] Lucena, B. and Haussler, D. (2005). Counterexample to a claim about the reconstruction of ancestral character states. *Systematic Biology*, 54(4):693–695.

[Maddison and Maddison, 1992] Maddison, W. P. and Maddison, D. (1992). *MacClade: Analysis of Phylogeny and Character Evolution*. Sinaur Associates, Inc.

[Malcolm et al., 1990] Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F., and Wilson, A. C. (1990). Ancestral lyzozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, 245:86–89.

[Marinari and Parisi, 1992] Marinari, E. and Parisi, G. (1992). Simulated tempering: a new monta carlo scheme. *Europhysics Letters*, 19(6):451–458.

[Maritz, 1970] Maritz, J. (1970). *Empirical Bayes Methods*. Methuen's monographs on applied probability and statistics. Methuen.

[Matsuda, 1996] Matsuda, H. (1996). Protein phylogenetic inference using maximum likelihood with a genetic algorithm. In Hunter, L. and Klein, T., editors, *Biocomputing: Proceedings of the 1996 Pacific Symposium*, Singapore. World Scientific Publishing Co.

[Mau, 1996] Mau, B. (1996). *Bayesian phylogenetic inference via Markov chain Monte carlo methods.* PhD thesis, University of Wisconsin, Madison.

[Mau and Newton, 1997] Mau, B. and Newton, M. (1997). Phylogenetic inference for binary data on dendrograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6:122–131.

[Mau et al., 1999] Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics*, 55:1–12.

[Metropolis, 1949] Metropolis, N. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341.

[Metropolis, 1987] Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1091.

[Newton et al., 1999] Newton, M. A., Mau, B., and Larget, B. (1999). Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees for aligned molecular sequences. *Statistics in Molecular Biology*, 33:143–162.

[Osborn and Smith, 2005] Osborn, A. M. and Smith, C. J. (2005). *Molecular Microbial Ecology*. Taylor and Francis.

[O'Sullivan et al., 2004] O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3dcoffee: Combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2):385–395.

[Ott et al., 2007] Ott, M., Zola, J., Aluru, S., and Stamatakis, A. (2007). Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*.

[Pagel et al., 2004] Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53(5):673–684.

[Pauling and Zuckerkandl, 1963] Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.*, A(17):S9–S16.

[Pupko et al., 2000] Pupko, T., Shamir, I. P. R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, 17(6):890–896.

[Rambaut and Grassly, 1997] Rambaut, A. and Grassly, N. C. (1997). Seqgen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238.

[Rannala and Yang, 1996] Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43:304–311.

[Ridley, 2004] Ridley, M. (2004). *Evolution*. Blackwell Publishing.

[Rodrigo, 1993] Rodrigo, A. G. (1993). Calibrating the bootstrap test of monophyly. *International Journal for Parasitology*, 23(4):507–514.

[Rogers, 1997] Rogers, J. S. (1997). On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic Biology*, 46(2):345–357.

[Ronquist and Huelsenbeck, 2003] Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

[Schopf, 2006] Schopf, J. W. (2006). Fossil evidence of archaean life. *Philosophical Transactions of the Royal Society*, 361:869–885.

[Schultz and Churchill, 1999] Schultz, T. R. and Churchill, G. A. (1999). The role of subjectivity in reconstructing ancestral character states: A bayesian approach to unknown rates, states, and transformation asymmetries. *Systematic Biology*, 48(3):651–664.

[Simmons et al., 2004] Simmons, M. P., Pickett, K. M., and Miya, M. (2004). How meaningful are bayesian support values? *Molecular Biology and Evolution*, 21(1):188–199.

[Sokal and Sneath, 1963] Sokal, R. R. and Sneath, P. H. (1963). *Principles of numerical taxonomy*. W.H. Freeman, San Francisco.

[Stackhouse et al., 1990] Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P., and Benner, S. A. (1990). The ribonuclease from an extinct bovid ruminant. *Federation of European Biochemical Societies*, 262(1):104–106.

[Stamatakis et al., 2007] Stamatakis, A., Blagojevic, F., Nikolopoulos, D., and Antonopolous, C. (2007). Exploring new search algorithms and hardware for phylogenetics: RAxML meets the IBM Cell. *The Journal of VLSI Signal Processing*, 48(3):271–286.

[Stamatakis et al., 2005] Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

[Subramanian et al., 2008] Subramanian, A. R., Kaufmann, M., and Morgenstern, B. (2008). Dialign-tx: greedy and progressive approaches for

segment-based multiple sequence alignment. *Algorithms for Molecular Biology*, 3(6).

[Swofford, 2003] Swofford, D. L. (2003). Phylogenetic analysis using parsimony (and other methods) 4.0 b10. Sineer Associates, Inc.

[Swofford and Maddison, 1987] Swofford, D. L. and Maddison, W. P. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, 87:199–229.

[Swofford and Maddison, 1992] Swofford, D. L. and Maddison, W. P. (1992). Parsimony, character-state reconstructions and evolutionary inferences. In R.L.Mayden, editor, *Systematics, Historical Ecology, and North American Freshwater fishes*, chapter 5. Stanford University Press.

[Swofford et al., 1996] Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). *Phylogenetic Inference*, chapter 11. Sinaur Associates, Inc., 2 edition.

[Tavare, 1986] Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathmatics in Life Sciences*, 17:57–86.

[Temin and Mizutani, 1970] Temin, H. M. and Mizutani, S. (1970). Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of rous sarcoma virus. *Nature*, 226:1211–1213.

[Thompson et al., 1994] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680.

[Thomson et al., 2005] Thomson, J. M., Gaucher, E. A., Burgan, M. F., Kee, D. W. D., Li, T., Aris, J. P., and Benner, S. A. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics*, 37(6):630–635.

[Thornton, 2004] Thornton, J. W. (2004). Resurrecting ancient genes: Experimental analysis of extinct molecules. *Nature*, 5:366–375.

[van der Vaart, 2000] van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

[Wheeler et al., 1995] Wheeler, W. C., Gatesy, J., and DeSalle, R. (1995). Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Molecular Phylogenetics and Evolution*, 4(1):1–9.

[Whelan and Goldman, 2001] Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–699.

[Yang, 1994] Yang, Z. (1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology*, 43(3):329–342.

[Yang, 1997] Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556.

[Yang, 2007] Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

[Yang et al., 1995] Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650.

[Yang and Rannala, 1997] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo metho. *Molecular Biology and Evolution*, 14:717–724.

[Zharkikh and Li, 1995] Zharkikh, A. and Li, W.-H. (1995). Estimation of confidence in phylogeny: The complete-and-partil bootstrap technique. *Molecular Phylogenetics and Evolution*, 4(1):44–63.