Elicitation for Bayesian Phylogeny

Xiting Yang

Thesis Proposal

October 5, 2007

Abstract

Phylogeny provides important information for evolutionary relationships and is useful in applications. Bayesian phylogenetic inference developed quickly because of its computational feasibility. However, the vague priors usually used in Bayesian phylogenetics do not capture knowledge available from non-genetic sources such as studies of physiology and development. This thesis gives a practical way to elicit a biologist's prior beliefs about phylogeny.

1 Introduction

1.1 Phylogeny

A phylogeny, sometimes called evolutionary tree, "is a branching tree diagram showing the course of evolution in a group of organisms" (Felsenstein, 1983, p. 246). It studies the species relationships by analyzing and finding the time the species split from their most recent common ancestors.

Phylogeny is of importance to many fundamental biological questions such as the evolutionary history, the epidemiology of diseases (Huelsenbeck et al., 2001), the prediction of gene functions (Olken, 2002), etc. Phylogenetic analysis is possibly the first step in any DNA sequence study (Huelsenbeck et al., 2001). In practice, it has been used to determine viral transmission events in a legal case (Metzker et al., 2002), and to identify the origins of wildlife products and to monitor and protect wildlife (Baker and Palumbi, 1994; Baker et al., 2000).

1.2 Notation

The following are copied from (Gascuel, 2005):

A graph is a pair G = (V, E), where V is a set of objects called **vertices** or **nodes**, and E is a set of **edges**, that is, pairs of vertices. A **path** is a sequence $(v_0, v_1, ..., v_k)$ such that for all i, $(v_i, v_{i+1}) \in E$. A **cycle** is a path as above with $k > 2$, $v_0 = v_k$ and $v_i \neq v_j$ for $0 \leq i < k$. A graph is **connected** if each pair of vertices, $x, y \in V$ is connected by a path, denoted p_{xy}. A connected graph containing no cycles is a **tree**.

The degree of a vertex v, $deg(v)$, is defined to be the number of edges containing v. In a tree, any vertex v with $deg(v) = 1$ is called a **leaf**. Other vertices are called internal. In phylogenetic trees, internal nodes have degree 3 or more.

A **metric** is a function with certain properties on unordered pairs from a set. Suppose X is a set. The function $d : X * X \rightarrow R$ (the set of real numbers) is a **metric** if it satisfies:
1. $d(x, y) \geq 0$ for all x, y, with equality if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all x, y.
3. For all x, y and z, $d(x, z) \leq d(x, y) + d(y, z)$.

For the remainder of the proposal, I shall use xy in place of $d(x, y)$.

Phylogenies usually have lengths assigned to each edge. Such lengths are referred to as **branch length**.

For my purposes, I shall reserve the word **topology** to describe the shape of a tree con-

sidering the leaf labels without regard to edge lengths. Different from that in mathematics, topology includes the information not only the shape but also the leaf labels. Different leaf labels on the same "shape" of a tree result different topologies. As shown in Figure 1, the three trees have the same tree "shape" yet different tree topologies.

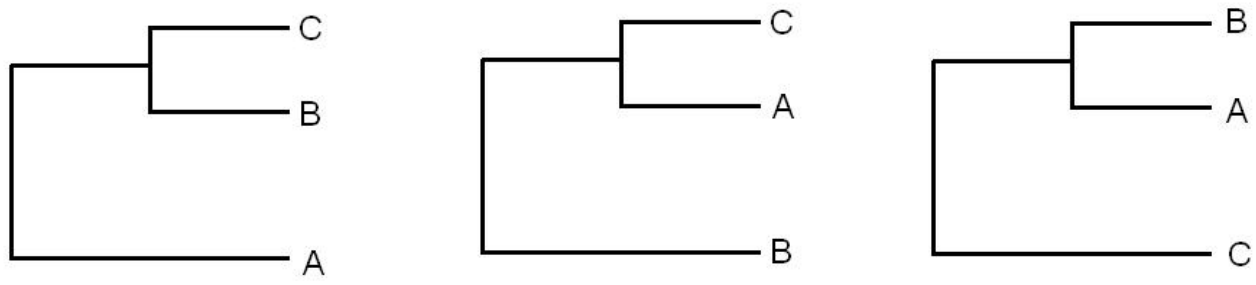

Figure 1: Three tree topologies with the same "shape".

On the other hand, two trees that differ only on left-right order branching are counted as the same tree. For example, the two trees in Figure 2 are considered the same, though the left-right branchings of B and C are different.

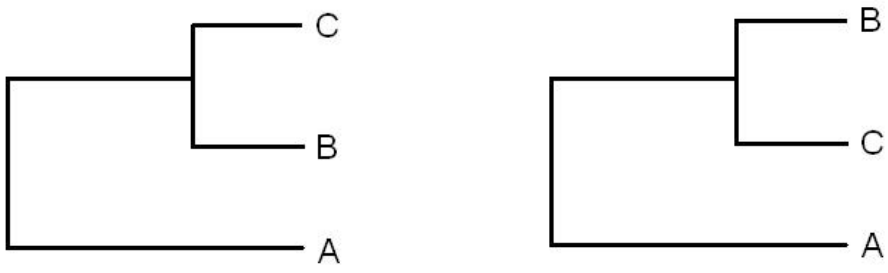

Figure 2: Two trees that differ only on left-right order branching are counted as the same.

In evolutionary studies, phylogenies are drawn as branching trees deriving from a single ancestral species. This species is known as the **root** of the tree (Gascuel, 2005). A tree with

such a root is called a **rooted tree** and that without a root is called an **unrooted tree**. A group of species that have a common ancestor is called a **clade**. A tree such that the degree of each vertex is no more than 3 is a **binary tree**.

In phylogenetic trees, the directly observed species are the leaves; the internal nodes are the common ancestors of the observed species and are not directly observed. Trees in most of the current phylogeny studies are binary trees because the speciation happens when one species splits into two and the simultaneous arising of several species is quite unlikely. Some examples of phylogenetic trees can be found in Larget and Simon (1999) and Larget et al. (2002).

1.3 Statistical Methods in Phylogeny

Many statistical methods have been used to study evolutionary relationships. Both the neighbor joining (NJ) method (Saitou and Nei, 1987) and the unweighted pair grouping method using arithmetic averages (UPGMA) (Sneath and Sokal, 1973, p 230-240) find the best tree by iteratively combining pairs of neighbors. UPGMA assumes that the distance matrix makes the tree a rooted one and combines the pair with minimum single branch length. The NJ does not assume a rooted tree and combines the pair that minimizes the total tree length (Gascuel, 2005). Both NJ and UPGMA are fast and feasible, yet tend to have large errors when the divergence between species is large (Holder and Lewis, 2003). This is because NJ and UPGMA require a distance matrix calculated from the sequences of the species. This distance matrix is supposed to represent the evolutionary distances between the species but that is hard to remain so when many changes occur on the sequences (Holder and Lewis, 2003).

Tree searches in phylogenetics are performed using criteria such as minimum evolution, parsimony and maximum likelihood (Holder and Lewis, 2003). First used by Kidd and Sgaramalla-Zonta (Kidd and Sgaramella-Zonta, 1971), Minimum evolution (ME) methods find the tree with shortest total branch length as the best phylogeny (Felsenstein, 2004). Like NJ, it starts with a distance matrix and thus works well when the sequence divergence is low (Holder and Lewis, 2003). In parsimony methods, an algorithm finds the tree with the minimum number of mutations on gene sequences (Holder and Lewis, 2003; Larget et al., 2005). Different from minimum evolution methods, which are distance-based, parsimony methods put gene sequences onto a tree and count the number of mutations needed for that tree. It is fast for data sets with hundreds of species yet "performs poorly if there is substantial variation in branch lengths" (Holder and Lewis, 2003). Maximum likelihood (ML) phylogeny is widely used and is very "appealing" (Holder and Lewis, 2003). First introduced to phylogeny by Edwards and Cavalli-Sforza (1964), maximum likelihood was first applied to molecular sequences by Neyman (1971). Later Felsenstein (1981) made it practical for

"moderate numbers of sequences" (Felsenstein, 2004, p. 248). It is statistically consistent, which means that as the number of characters calculated in the likelihood function increases, the maximum likelihood estimator converges in probability to the true parameter (Gascuel, 2005). Yet it is also computationally intensive (Alfaro and Holder, 2006). Problems with more than 100 species are "tedious" with ML (Holder and Lewis, 2003). The basic idea of ML is to find the parameters $\hat{\theta}$ in the likelihood function $L(\theta) = P(D|\theta)$ so that the likelihood function is maximized. In ML phylogeny, the parameters θ are the tree topology, the branch lengths, the sequence evolution model, and so on (Gascuel, 2005); the data are the molecular information about the species, such as the DNA sequence. In a DNA sequence, there are many different sites. In calculating the likelihood function, the ML makes an assumption that different sites on a sequence evolve independently. This means that the likelihood function of sequence A evolving to sequence B equals the product of the probability of all sites in A evolving to the corresponding sites in B. This independence assumption effectively simplifies the problem, but is not realistic (Gascuel, 2005). Another problem with ML phylogeny is that defining a confidence interval for the tree picked by the ML is very difficult if not impossible.

Starting in the last century, Bayesian inference of phylogeny (Sinsheimer et al., 1996; Larget and Simon, 1999; Rannala and Yang, 1996) quickly developed (Alfaro and Holder, 2006; Huelsenbeck et al., 2001). The posterior sampling obtained with Bayesian phylogenetics gives ways to characterize the uncertainties of topologies and branch lengths (Alfaro and Holder, 2006).

As the two problems in the Bayesian analysis, the explanation of the posterior distributions and the use of proper prior distributions are the foci of Bayesian Phylogenetics (Alfaro and Holder, 2006). The former concerns the biological explanation of the tree samplings obtained from the posterior and the characterization of the uncertainties. The latter focuses on finding a way to express and to use proper priors in the Bayesian phylogeny. This thesis studies both problems.

1.4 Priors in Bayesian Phylogenetic

Priors used in Bayesian phylogenetic inference include the priors for the topology, for the branch lengths, and for any other related parameters (Alfaro and Holder, 2006).

In most of the applications of Bayesian phylogenetics, the priors on topology are set to be uniform, which means all trees are equally likely (Huelsenbeck et al., 2001). This vague prior obviously does not reflect knowledge about species relationships as almost all species have been studied by biologists. A lot of information is lost in using the vague prior.

Some initial efforts have been made using informative priors on the topologies. The idea

of using the posterior probability of a first study as the prior for a second study is "not practical" due to the computer memory limit and the lack of precise probability density (Alfaro and Holder, 2006). Putting a prior distribution on a subset of possible tree topologies instead of all tree topologies is another option. For example, Larget et al. (2005) tried to put a uniform prior on a group of "biologically plausible trees" instead of on all possible trees (Larget et al., 2002) and got better results.

Though not the primary focus of phylogenetic analysis, branch lengths are closely related to the topology (Yang and Rannala, 2005; Alfaro and Holder, 2006). Like maximum likelihood methods, nearly all Bayesian approaches "treat each branch length as an independent parameter" (Alfaro and Holder, 2006). This is problematical. For example, a set of branch lengths treated as independent might not meet the requirement of a rooted tree, where the distances of all leaves to the root are equal (Alfaro and Holder, 2006).

The focus of this thesis is to find a way to express effectively a biologist's opinion on the evolutionary history of the species involved in a study. This means having a proper informative joint prior on the topology and branch lengths.

2 Proposed Work

2.1 Elicitation

"Elicitation is the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a (joint) probability distribution for those quantities" (Garthwaite et al., 2005, p680). Coming in different formats, elicitation has been applied to medicine, the nuclear industry, agriculture, and business, etc. (O'Hagan et al., 2006). Dealing with tree spaces, the main aim of this thesis is to elicit an expert's knowledge on phylogeny to provide better priors in order to improve Bayesian inference of phylogeny.

2.2 Tree Topology

2.2.1 Four Point Condition

The four Point Condition is the condition that for every four (not necessarily distinct) elements $w, x, y, z \in T$, where T is the leaf set, two of the three terms in the following list are equal and greater than or equal to the third: $wx + yz$, $wy + xz$, $wz + xy$, where wx represents the total length of the unique path between w and z in T and similarly with yz, xz and xy (Gascuel, 2005). Buneman (1971) proved that the four point condition is the necessary and sufficient condition for a metric to be realized by a (unrooted) tree.

2.2.2 Ultrametrics and Three Point Condition

A metric (X,d) is said to satisfy the **Three Point Condition (3PC)** if for any three leaves x, y and z in X, two of xy, xz and yz are equal and no less than the third. The three point condition is also called the **ultrametric inequality**. It is a necessary and sufficient condition for a metric to be realized by a rooted tree (Barthelemy and Guenoche, 1991; Gascuel, 2005). In other words, for any tree metric that satisfies the **3PC**, there is a unique tree (with specific branch lengths) and vice versa. Such a metric is called an **ultrametric**.

2.3 Distribution Over Tree Space

2.3.1 Joint Distribution and Conditional Distribution

As we are interested in the evolutionary relationship in a time scale, our work uses rooted trees to represent phylogenies. So at any time, the tree we have should meet the three point condition (**3PC**). The **3PC** gives the unique correspondence between a rooted tree and an ultrametric. The unweighted pair grouping method using arithmetic averages (UPGMA) (Sneath and Sokal, 1973) can be easily applied to reconstruct a rooted tree from an ultrametric. So to describe the distribution on tree space, with characters of both the tree topology and the branch lengths, all we care about is the joint distribution of all the distances between all species. A set of distances satisfying the three point condition specifies a unique tree topology with specific branch lengths. Those sets of distances not satisfying the three point condition have probability of 0 in the tree space.

The joint distribution of the distances between species can be factored into conditional distributions. For example, suppose we have a set of leaves named A, B, C, D and E and the tree is recorded as Tree ABCDE. The distance between species A and B is recorded as AB and similarly with any pair of species. Then,

$$
\begin{aligned}
& P(TreeABCDE) && (1)\\
= \; & P(AB, AC, BC, AD, BD, CD, AE, BE, CE, DE) && (2)\\
= \; & P(AB)P(AC, BC|AB)P(AD, BD, CD|AB, AC, BC) \\
& P(AE, BE, CE, DE|AB, AC, BC, AD, BD, CD) && (3)\\
= \; & P(AB)P(AC, BC|TreeAB)P(AD, BD, CD|TreeABC) \\
& P(AE, BE, CE, DE|TreeABCD) && (4)
\end{aligned}
$$

In general, starting with a probability distribution on trees with n species (satisfying the 3PC), one must specify the joint distribution of the distance between a $(n+1)^{st}$ species and

each of the original n (satisfying the 3PC) in order to get a probability distribution on trees with (n+1) species.

2.3.2 Introducing a New Species

The distribution of new trees can be obtained by expanding the distribution of existing trees. For example, for any species A, B and C, if the distribution of the branch length of AB is known and we want to add a new species called C. Let's say that AC is no bigger than BC. So there are only two possibilities: $AC < BC$ and $AC = BC$. If $AC < BC$, then by **3PC**, $BC = AB$, then $P(AC, BC|AB) = P(AC|AB)$; if $AC = BC$, obviously $P(AC, BC|AB) = P(AC|AB)$. Overall, we have

$$\begin{aligned} & P(AB, AC, BC) && (5)\\ = \; & P(AC, BC|AB)P(AB) && (6)\\ = \; & P(min(AC, BC)|AB)P(AB) && (7)\\ = \; & P(AC|AB)P(AB) && (8) \end{aligned}$$

In general, with the requirements of the three point condition, when the distribution of an n-species tree (with species $A_1, A_2, ..., A_i, ..., A_n$) is known and we would like to add a new species A_{n+1}, the distribution of the minimum distance among the branches of A_iA_{n+1}, where $i \in 1...n$, would define the distribution of the new tree. For example, if the expert gives the minimum as a path between A_s and A_{n+1}, then for any other species j ($A_s \neq A_j$), we know the distance of A_sA_j from the previous n-species tree, and the distance of A_sA_{n+1} as new information. By the three point condition, A_jA_{n+1} equals to A_sA_j. So the distribution of the minimum of the newly added branches specifies the distribution of the whole tree.

$$\begin{aligned} & P(TreeA_1A_2...A_{n+1}) && (9)\\ = \; & P(A_1A_2, A_1A_3, ..., A_1A_n, A_2A_3, ...A_{n-1}A_n, A_1A_{n+1}, A_2A_{n+1}, ..., A_nA_{n+1}) && (10)\\ = \; & P(A_1A_{n+1}, A_2A_{n+1}, ..., A_nA_{n+1}|A_1A_2, A_1A_3, ..., A_1A_n, A_2A_3, ...A_{n-1}A_n) \\ & P(A_1A_2, A_1A_3, ..., A_1A_n, A_2A_3, ...A_{n-1}A_n) && (11)\\ = \; & P(min(A_1A_{n+1}, A_2A_{n+1}, ..., A_nA_{n+1})|A_1A_2, A_1A_3, ..., A_1A_n, A_2A_3, ...A_{n-1}A_n) \\ & P(A_1A_2, A_1A_3, ..., A_1A_n, A_2A_3, ...A_{n-1}A_n) && (12)\\ = \; & P(A_sA_{n+1}|TreeA_1A_2A_3...A_n)P(TreeA_1A_2...A_n) && (13) \end{aligned}$$

2.4 Distribution of A Single Branch Length

An expert would have a possible range for the time at which two species have a most recent common ancestor. So using a Beta distribution to represent the probability of such an event is a good choice. This requires the expert give a range plus two quantiles. I recommend using the 33% quantile and 67% quantile plus the lower and the upper bound of the range to specify a shifted and scaled Beta distribution.

The elicitation is done by asking about the possible range, and two quantiles of the distance between A and B. For example, a biologist might think that the time A and B split falls into the range of 40 - 90 million years ago, with a 33% quantile as 50 and 67% quantile as 70. This information gives us a shifted and scaled Beta distribution as $P(AB) \sim 40 + 50 * Beta(0.57, 0.82)$. Parameters of α and β used in the $Beta(\alpha, \beta)$ distribution are solved numerically.

In many other cases, an expert would have an idea of the relationship based on the current tree topology regardless of the branch length. For example, conditional on the current topology of ABC as shown in Figure 3, an expert thinks that the new species D relates most closely to the most recent common ancestor of A and C. As shown in Figure 3, the length of the branch on which AC are not separate is a-b. Consider the time D splits from AC has a lower and upper bound of $\frac{1}{4}(a-b)$ and $\frac{4}{5}(a-b)$ with 33% and 67% quantiles as $\frac{1}{3}(a-b)$ and $\frac{2}{3}(a-b)$. This gives a tree ABCD as shown in Figure 4.

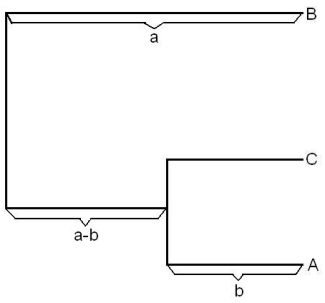

Figure 3: An example of Tree ABC: $d_{AB} = a, d_{AC} = b$.

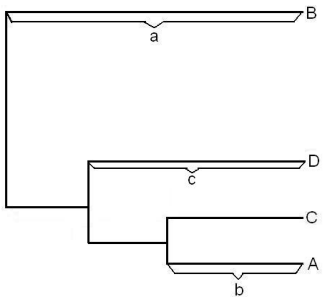

Figure 4: An example of Tree ABCD: $d_{AB} = a, d_{AC} = b, d_{AD} = c$.

2.5 Proposed Elicitation

2.5.1 New Species Relates to a Single Species

When adding more species, always consider the question that "Among the previous species, is there one which is the closest to the new incoming one?" If that question can be answered with a particular one species, then this defines a specific topology. As mentioned in Equation 13, if the distribution of the length of the new shortest distance is set, the distribution of the new tree is set as well.

2.5.2 New Species Relates to a Group of Species

If the question "Among the previous species, is there one which is the closest to the new incoming one?" can be answered as "several of those previous species are equally related to the incoming species as a clade", say, A_{s_1}, A_{s_2} and A_{s_3}. Then we know that A_{n+1} has a most recent common ancestor with the most recent common ancestor of A_{s_1}, A_{s_2} and A_{s_3}. For any $A_j \neq A_{s_1}, A_{s_2}, A_{s_3}$, by the three point condition, $A_j A_{n+1} = A_{s_1} A_j = A_{s_2} A_j = A_{s_3} A_j$. Equation 13 still applies.

When the species A_{s_1}, A_{s_2} and A_{s_3} are considered equally related to the incoming species, yet these species do not compose a clade, the species A_{s_1}, A_{s_2} and A_{s_3} can be treated as if they are several groups, as in section 2.5.3.

2.5.3 New Species Relates to Several Groups of Species with Different Probabilities

Much of the time, an expert wouldn't be able to specify a single species or a single group of species for a new species to connect to. More often, she would say something like "I believe there is a 20% chance the new species is connected to group ABD, and 50% of the chance it is directly related to group AB, and 30% chance it's most closely related to the species E". In that case, the elicitation will be done in a hierarchical manner: each possibility will be handled separately and each gives a distribution, and the overall distribution is the mixture of these distributions with corresponding probabilities.

2.5.4 Elicit the Branch Length

The question "Among the previous species, is there one which is the closest to the new incoming one?" elicits knowledge about the topology. Questions like "Condition on the current topology and consider the length of the branch to which the new species can be added is of length "l", please give us the possible range of the time you think the split happens." together with questions like "What's the 33% and 67% of the quantile the time of the split happens." will specify the parameters α and β used in the Beta distribution $Beta(\alpha, \beta)$ to describe the distribution of the branch lengths. The distribution will be shifted and scaled by the previous branch lengths. The joint distribution of these previous lengths can be found in the distributions of the existing trees, as explained in Section 3.

3 Example and Preliminary Results

3.1 Trees Involving Only Two Species

Consider starting from a relationship involving only two species: A and B. If the expert tells us that the time that A and B separate from each other falls into the range of 80-100 million years ago, the 33% quantile of that time is 90 million years ago and the 67% quantile is 95 million years ago, then we know that $P(AB) \sim 80 + 20 * Beta(1.94, 1.26)$, which means that

$$\begin{aligned} P(treeAB) &= P(AB = a) \\ &= \frac{1}{B(1.94, 1.26)} \left(\frac{a-80}{20}\right)^{0.94} \left(1 - \frac{a-80}{20}\right)^{0.26} \end{aligned} \tag{14}$$

where $80 < a < 100$. This gives the topology in Figure 5.

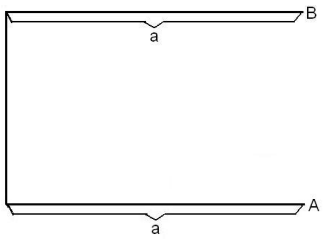

Figure 5: An example of Tree AB: $d_{AB} = a$.

3.2 Adding the Third Species

For the question "Among the previous species, is there one which is the closest to the new incoming one?", let's say the expert chooses species A. And the range is given as $\frac{a}{4}$ to $\frac{4a}{5}$ with a 33% quantile as $\frac{a}{3}$ and 67% as $\frac{a}{2}$. Then the topology would be the one in Figure 3. The conditional distribution $P(AC, BC|AB = a) = P(AC|AB = a) = \frac{a}{4} + \frac{11a}{20} * Beta(0.65, 1.25)$.

So $P(AC = b|AB = a) = \frac{1}{B(0.65,1.25)} \left(\frac{b-\frac{a}{4}}{\frac{11a}{20}}\right)^{-0.35} \left(1 - \frac{b-\frac{a}{4}}{\frac{11a}{20}}\right)^{0.25}$, where $\frac{a}{4} < b < \frac{4a}{5}$. The distribution for the tree ABC is

$$\begin{aligned} P(treeABC) &= P(AC, BC|AB = a)P(AB = a) \\ &= P(AC = b|AB = a)P(AB = a) \\ &= \frac{1}{B(0.65, 1.25)} \left(\frac{b - \frac{a}{4}}{\frac{11a}{20}}\right)^{-0.35} \left(1 - \frac{b - \frac{a}{4}}{\frac{11a}{20}}\right)^{0.25} \\ &\quad * \frac{1}{B(1.94, 1.26)} \left(\frac{a - 80}{20}\right)^{0.94} \left(1 - \frac{a - 80}{20}\right)^{0.26} \end{aligned}$$

where $80 < a < 100$ and $\frac{a}{4} < b < \frac{4a}{5}$.

3.3 Adding the Fourth Species

Condition on the current tree of A, B and C, the expert thinks that species D has a 50% chance being most closely related to group AC with a range of $b+\frac{(a-b)}{4}$ to $b+\frac{a-b}{2}$ and a 33% and 67% quantiles of $b+\frac{a-b}{3}$ and $b+\frac{5(a-b)}{12}$, plus a 50% chance of being most closely related to B directly with a range of $\frac{a}{2}$ to $\frac{7a}{8}$ and 33% and 67% quantiles of $\frac{5a}{8}$ and $\frac{6a}{8}$. The former gives a $\frac{a-b}{4}+\frac{a-b}{4}*Beta(0.91,0.91)$ and a topology in Figure 4.

$$P(AD=c|treeABC)=\frac{1}{B(0.91,0.91)}\left(\frac{c-b-\frac{a-b}{4}}{\frac{a-b}{4}}\right)^{-0.19}\left(1-\frac{c-b-\frac{a-b}{4}}{\frac{a-b}{4}}\right)^{-0.19} \tag{15}$$

where $b+\frac{(a-b)}{4}<c<b+\frac{(a-b)}{2}$.

The latter gives a $\frac{b}{2}+\frac{3*b}{8}Beta(1.14,1.14)$ and a topology in Figure 6.

$$P(BD=d|treeABC)=\frac{1}{B(1.14,1.14)}\left(\frac{d-\frac{a}{2}}{\frac{3a}{8}}\right)^{0.14}\left(1-\frac{d-\frac{a}{2}}{\frac{3a}{8}}\right)^{0.14} \tag{16}$$

where $\frac{a}{2}<d<\frac{7a}{8}$

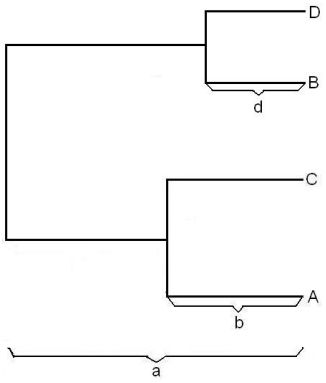

Figure 6: An example of Tree ABCD: $d_{AB}=a, d_{AC}=b, d_{BD}=d$.

The overall conditional distribution is

$$P(AD,BD,CD|AB,AC,BC)=0.5*P(AD|AB,AC,BC)+0.5*P(BD|AB,AC,BC)$$

The distribution of tree ABCD is:

$$
\begin{aligned}
P(treeABCD) &= 0.5P(AD|treeABC)P(treeABC) + 0.5P(BD|treeABC)P(treeABC) \\
&= 0.5 * (\frac{1}{B(0.91, 0.91)} \left(\frac{c - b - \frac{a-b}{4}}{\frac{a-b}{4}}\right)^{-0.19} \left(1 - \frac{c - b - \frac{a-b}{4}}{\frac{a-b}{4}}\right)^{-0.19} \\
&\quad + \frac{1}{B(1.14, 1.14)} \left(\frac{d - \frac{a}{2}}{\frac{3a}{8}}\right)^{0.14} * \left(1 - \frac{d - \frac{a}{2}}{\frac{3a}{8}}\right)^{0.14}) \\
&\quad * \frac{1}{B(0.65, 1.25)} \left(\frac{b - \frac{a}{4}}{\frac{11a}{20}}\right)^{-0.35} \left(1 - \frac{b - \frac{a}{4}}{\frac{11a}{20}}\right)^{0.25} \\
&\quad * \frac{1}{B(1.94, 1.26)} \left(\frac{a - 80}{20}\right)^{0.94} \left(1 - \frac{a - 80}{20}\right)^{0.26}
\end{aligned}
$$

where $80 < a < 100$, $\frac{a}{4} < b < \frac{4a}{5}$, $b + \frac{(a-b)}{4} < c < b + \frac{(a-b)}{2}$, and $\frac{a}{2} < d < \frac{7a}{8}$.

4 Discussion and Future Work

4.1 Discussion

This elicitation is based on conditional distribution of previous trees. The size of tree space increases exponentially when the number of species increase. With n species, the number of binary rooted trees is $(2n-3)!! = (2n-3) * (2n-5) * ... * *3 * 1$ (Schroder, 1870). So the elicitation proposed here works well when many trees have a probability of 0 and thus only a small number of trees need to be studied. When the number of the trees to be studied is large, the system becomes unwieldy.

The zero probability priors put on many tree topologies makes it impossible for any of those topologies to show in the posterior. So the choice of which topologies have zero probability is vital in the elicitation and its application in Bayesian phylogeny.

If the biologist is very sure that some species should exist as a clade, we can use this fact to help to handle a larger number of species. For example, for a study on species A, B, C, D, E, and F, if the biologist thinks that species A, B and C exist as a clade, we can call the clade including A, B and C "Cl". Do elicitation on Cl, D, E, and F as if Cl is a general species. Let's say this 4 species elicitation gives a tree shown on the left of Figure 7. The clade "Cl" splits with species D at time "s". Then focus on the species within the clade and do an elicitation involving the three species A, B, and C. The length from the root to the

present time "r" is no bigger than the length of "s". Overall, the whole tree on the six species is shown on the right of Figure 7. Similarly, if two or more clades exist, each of the clade can be considered as a group in the first level elicitation and more detailed elicitation can be made within each clade on the second level elicitation. This iterative elicitation reduces the complexity of the elicitation and makes it possible to handle larger number of species.

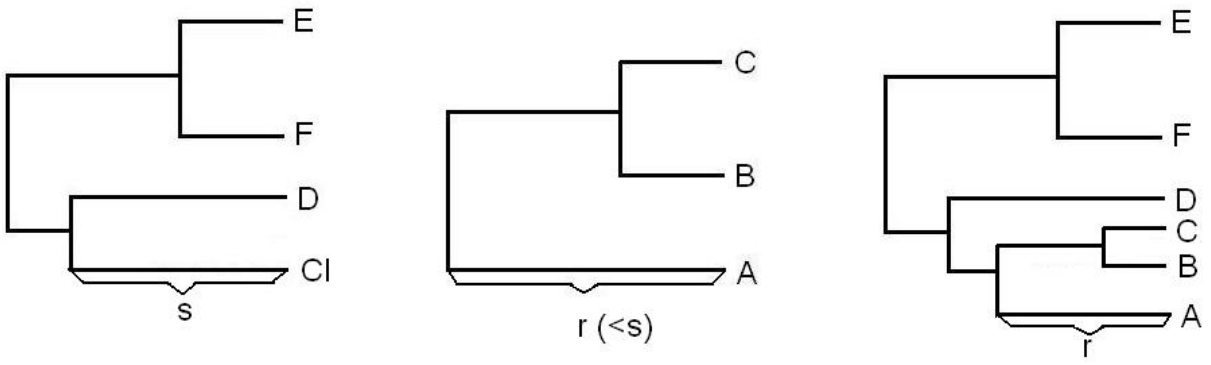

Figure 7: When some species exist as a clade, the iterative elicitation can help to handle larger number of species.

The branch length of the very first pair of species plays an important role in the whole elicitation. As can be seen in section 3, once the distribution of the first branch length is set, all other branch lengths are considered proportional to it. In our example, once the distribution of branch length AB, a, is set, other branch lengths, b, c and d, are considered to be proportional to a.

4.2 Future Work

A program needs to be written to help form a quick realization of the elicitation. It will be done in R because of the convenience of the many existing statistical packages in R.

The program will be tested on a biologist. A feasible maximal number of species will be explored. As mentioned before, the system complexity increases greatly as the number of species increases. In practice, when the number of species to study is too big, the elicitation method proposed here would be very hard to carry out. I am interested in gaining experience about how many species can practically be used in an elicitation.

Also as presented, this elicitation method depends on the order in which species are introduced. To what extent do the answers vary depending on this order? One way to check

this is to generate trees from the distributions obtained from the elicitation and then check with the biologist to see which distribution represent the best of his belief.

The elicitation results will be incorporated into MCMC and Bayesian phylogeny.

What's more, it would be interesting to use this method to summarize the output from an MCMC. As we can get probability information on topology and branch length from a biologist, we can find out these from the posterior sample. Thus the elicitation made with the help of a biologist can be used in the same way to summarize the information in the posterior. For example, we can get the range and quantiles of the time that species A and B separate from the MCMC sampling. This gives us the shifted and scaled Beta distribution for tree AB, as described in section 3.1.

In modern phylogeny, the molecular information about the species are the data used in the analysis. The likelihood implied by the data can be studied by comparing the frequencies of the same clades in the prior and the posterior.

"Multidivtime" (Thorne, 2003) and "BEAST" (Drummond and Rambaut, 2006) are two programs focusing on getting the branch length when the topologies are known. For example, "BEAST" takes uniform, normal, lognormal, exponential, or gamma priors (Drummond et al., 2007) and then gives the posterior distribution. Using the elicitation proposed in this proposal, we can get the elicited priors on the branch length conditioned on a specific topology. The prior in "BEAST" (or "Multidivtime") and that obtained through elicitation lead to different branch length posteriors. It will be interesting to see the differences in real data settings and check the effect of elicited priors on the posteriors.

Finally, the elicitation discussed above is based on a time scale. How to relax the molecular clock assumption and get trees with edge lengths on the expected number of substitutions scale is another problem to study.

References

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

Baker, C. and Palumbi, S. (1994). Which whales are hunted? a molecular genetic approach to monitoring whaling. *Science*, 265(5178):1538–1539.

Baker, C. S., Lento, G. M., Cipriano, F., and Palumbi, S. R. (2000). Predicted decline of protected whales based on molecular genetic monitoring of Japanese and Korean markets. *Proceedings: Biological Sciences*, 267(1449):1191–1199.

Barthelemy, J.-P. and Guenoche, A. (1991). *Trees and proximity representations*. Wiley, New York.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In Hodson, F. R., Kendall, D. G., and Tautu, P., editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387– 395, Edinburgh. Edinburgh University Press.

Drummond, A. J., Ho, S. Y., Rawlence, N., and Rambaut, A. (2007). *A Rough Guide to BEAST 1.4*. Available from http://beast.bio.ed.ac.uk/.

Drummond, A. J. and Rambaut, A. (2006). BEAST v1.4, Available from http://beast.bio.ed.ac.uk/.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. In Heywood, V. H. and McNeill, J., editors, *Phenetic and Phylogenetic Classification*.

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

Felsenstein, J. (1983). Statistical inference of phylogenies. *Journal of the Royal Statistical Society. Series A (General)*, 146(3):246–272.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701.

Gascuel, O. (2005). *Mathematics of Evolution and Phylogeny*. Oxford University Press.

Holder, M. and Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews. Genetics*, 4(4):275–284.

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

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

Larget, B. and Simon, D. L. (1999). Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–759.

Larget, B., Simon, D. L., and Kadane, J. B. (2002). Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of The Royal Statistical Society Series B.*, 64(4):681–693.

Larget, B., Simon, D. L., Kadane, J. B., and Sweet, D. (2005). A Bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution*, 22(3):486495,.

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

Neyman, J. (1971). Molecular studies of evolution: A source of novel statistical problems. In Gupta, S. S. and Yackel, J., editors, *Statistical Decision Theory and Related Topics*.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley and Sons Ltd.

Olken, F. (2002). Phylogenetic tree computation tutorial.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3):304–311.

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.

Schroder, E. (1870). Vier combinatorische probleme. *Zeitschrift fr Mathematik und Physik*, 15:361–376.

Sinsheimer, J. S., Lake, J. A., and Little, R. J. A. (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, 52(1):193–210.

Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy*. W.K. Freeman and Company, San Francisco, CA.

Thorne, J. L. (2003). Multidivtime Software v9/25/03, Available from http://statgen.ncsu.edu/thorne/multidivtime.html.

Yang, Z. and Rannala, B. (2005). Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology*, 54(3):455–470.