*Syst. Biol.* 50(5):628–639, 2001

# Random Sampling of Constrained Phylogenies: Conducting Phylogenetic Analyses When the Phylogeny Is Partially Known

ELIZABETH A. HOUSWORTH[1] AND EMÍLIA P. MARTINS[2]

*Departments of Biology[1,2] and Mathematics,[1] University of Oregon, Eugene, Oregon 97403, USA;*
*E-mail: houswrth@darkwing.uoregon.edu*

*Abstract.*—Statistical randomization tests in evolutionary biology often require a set of random, computer-generated trees. For example, earlier studies have shown how large numbers of computer-generated trees can be used to conduct phylogenetic comparative analyses even when the phylogeny is uncertain or unknown. These methods were limited, however, in that (in the absence of molecular sequence or other data) they allowed users to assume that no phylogenetic information was available or that all possible trees were known. Intermediate situations where only a taxonomy or other limited phylogenetic information (e.g., polytomies) are available are technically more difficult. The current study describes a procedure for generating random samples of phylogenies while incorporating limited phylogenetic information (e.g., four taxa belong together in a subclade). The procedure can be used to conduct comparative analyses when the phylogeny is only partially resolved or can be used in other randomization tests in which large numbers of possible phylogenies are needed. [Branching process; combinatorics; comparative method; phylogenetic analysis; phylogeny.]

As analyses involving phylogenetic trees become more common, the need for computer-generated or "random" phylogenies in statistical randomization tests has increased. For example, one of the main difficulties with applying modern phylogenetic comparative methods (Harvey and Pagel, 1991; Martins and Hansen, 1997) is that they usually require the phylogeny be known with complete accuracy. Most phylogenies, however, are hypotheses describing the proposed relationships among taxa, estimated from a limited set of molecular or morphological data. One suggestion is to generate sets of possible phylogenies under certain conditions and conduct analyses on all of these trees (Losos, 1994; Martins, 1996). Whether the trees are generated from molecular data or generated entirely by a computer, these results can then be combined into confidence intervals and other measures of reliability by using the method described in Martins (1996). However, several problems arise in intermediate situations when some limited phylogenetic information is available (e.g., when taxonomic information tells us that 4 of 10 taxa belong together in one clade). This paper addresses some of the technical difficulties that have been encountered in using computers to generate random trees while also incorporating limited phylogenetic information and proposes a method that overcomes these problems. Although our algorithm should prove useful for randomizations involved in a variety of phylogenetic statistical analyses, we focus on the application to the phylogenetic comparative method.

Several solutions have been proposed for dealing with incomplete or unreliable phylogenetic information in comparative analyses (e.g., Grafen, 1989; Pagel, 1992; Losos, 1994). First, when molecular sequence data are available, we can bootstrap these and conduct our phylogenetic analysis on each of the resulting trees (Felsenstein, 1985). We might also apply Bayesian statistics (Huelsenbeck et al., 2000) to incorporate phylogenetic uncertainty into our comparative analyses. When sequence data are not available, the options are less promising. We can retain polytomies as true information (i.e., hard polytomies), inserting tiny branches between taxa as needed to make the available programs run (Felsenstein, 1985). Alternatively, we can resolve the polytomies artificially, using a variety of algorithms (e.g., Grafen, 1989, 1992; Pagel, 1992). These solutions are not ideal because we are still forced to assume that the final phylogeny is completely accurate before conducting further comparative analyses, and the resulting confidence intervals and hypothesis tests will not reflect our uncertainty in the final tree. Finally, we can develop a set of possible phylogenies (e.g., using computer simulation and a simple branching process model of speciation), conduct analyses on each of these trees, and

TABLE 1. An example of the application of randomization tests with the data and phylogeny from Figure 1. In all cases, Felsenstein's (1985) independent contrasts method was applied to the data and phylogeny, assuming that the character had evolved in a gradual or clocklike manner, the amount of character evolution being directly proportional to time. $\text{Var}_\text{P}$ is the variance in the slope estimate across analyses done on different phylogenies (assumed to be zero when the tree is known or estimated as the average across 1,000 trees), whereas $\text{Var}_\text{S}$ is the residual or sampling variance attributable to the data points not falling exactly on the regression line (estimated from a single data set or as an average of 1,000 estimates).

| Assumption | Slope estimate | S.E. | 95% C.I. | $\text{Var}_\text{P}$ | $\text{Var}_\text{S}$ |
|---|---|---|---|---|---|
| One tree (Fig. 1) is correct | 1.5 | 0.50 | 1.01 to 2.00 | 0.00 | 0.25 |
| 1,000 computer-generated trees of the form | | | | | |
| (A,B,C),(D,E,F),G,H,I | 1.5 | 0.85 | 0.06 to 2.84 | 0.48 | 0.23 |
| (A,B,C,D,E,F),G,H,I | 1.6 | 1.00 | −0.36 to 3.56 | 0.71 | 0.29 |
| (A,B,C),D,E,F,G,H,I | 1.4 | 1.08 | −0.71 to 3.51 | 0.79 | 0.37 |
| Only a branching process model | 1.4 | 1.30 | −1.15 to 3.95 | 1.33 | 0.36 |

combine the results into a broader confidence interval that takes phylogenetic uncertainty into account (Martins, 1996).

Although this last approach is accurate when nothing is known about the phylogeny, it may be too conservative when some limited phylogenetic information (e.g., a taxonomy or phylogeny with soft polytomies) is available but cannot be incorporated into the computer simulation process. Many regression slopes that are significantly different from zero when the analysis is done on a single phylogeny will not be different from zero when the analysis is done on large numbers of computer-generated phylogenies (e.g., Table 1, Fig. 1). Comparative studies are often used as an exploratory tool, to develop ideas for future studies, and are often based on small amounts of data available in the literature. Thus, we would prefer to incorporate any information we might have in the hopes of highlighting possible patterns in the data, rather than to be cautious and miss patterns because we did not include limited but important information. After all, we usually have at least some phylogenetic information, even if only taxonomic groupings based on phenotypic similarity, and we would like to be able to use this information in our hypothesis tests.

Imagine, for example (Fig. 1), that we have gathered data on nine taxa and are confident that three of these belong together in one subclade, three belong in a second subclade, and the remaining three are outside both of these specified subclades. The fine structure of the two subclades and their relationship to the three remaining taxa, however, are not known. Rather than generate a random set of phylogenies based on only the total number of taxa, we would like to incorporate this subclade information in our generation of random trees. At first thought, we might incorporate this limited phylogenetic information by using a branching process model to generate a tree for each subclade separately and then apply the branching process model a third time to patch the two subclades together into a larger phylogeny (e.g., Losos,

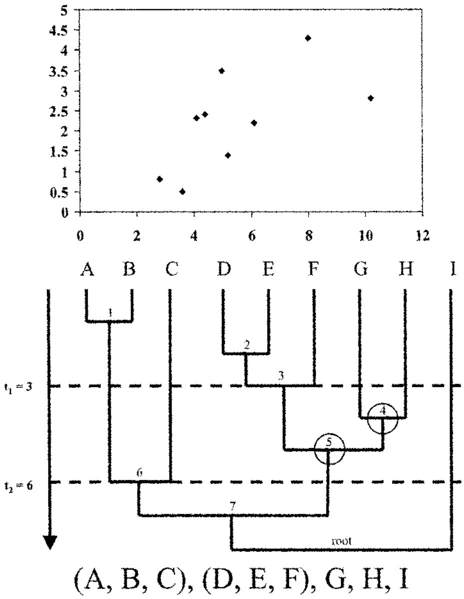

FIGURE 1. A hypothetical example of a comparative analysis used throughout this paper. It includes values for two traits measured in nine taxa and a single phylogeny. Imagine that we have little confidence in the phylogeny, but are fairly confident that the taxa can be divided into two known subclades of three taxa each. We describe this phylogenetic information as: (A, B, C), (D, E, F), G, H, I. Using the definitions in this paper, this tree belongs to a category specified by **P** = (2, 1), **T** = (3, 6), and **C** = (0, 2).

1994). However, how to generate the ordering of the nodes while retaining the standard branching process model of speciation is not obvious (Martins, 1996). In the current study, we discuss this problem in more detail and develop a technique for uniformly sampling from the set of all labeled ordered trees. Use of our technique is equivalent to generating a phylogeny under a random branching process (e.g., Page, 1991).

## The Problem: Random Sampling of Constrained Phylogenies

To use multiple phylogenies to generate confidence intervals around the results of a phylogenetic comparative method (e.g., using the method in Martins [1996]), the set of trees must be either the complete set of all possible phylogenies (e.g., five equally parsimonious trees) or a random sample from a larger population of phylogenies. For example, the standard branching process model, which has been well justified as a model of the speciation process (e.g., Slowinski and Guyer, 1989; Maddison and Slatkin, 1991; Hey, 1992; but see Aldous, 1995), generates a random sample from the set of all possible "ordered topologies". We might thus use a branching process model to generate 1,000 possible phylogenies for our taxa. The branching process model begins at the root of the tree by choosing one of the two daughter taxa at random to divide, forming a total of three taxa. One of these three taxa is then chosen to divide, forming a total of four taxa, and so on. Equivalently, we might start at the tips of the tree, choosing two of the taxa at random to be joined and proceeding towards a final joining event at the root of the tree, as is done when discussing the coalescent (e.g., Kingman, 1982; Hudson, 1990). Branch lengths can be specified in a variety of natural ways (Martins, 1996). For a concrete illustration, we will use an exponential distribution with the speciation rate proportional to the number of existing taxa. The resulting trees form a random sample from the set of all possible "ordered topologies".

An "ordered" topology differs from an "unordered" topology in that the timing of the internal nodes is considered important. For example, two trees with similar pairings of sister taxa are considered to be different "ordered topologies" if the order in which the pairs of sister taxa arise differs for the two trees (Fig. 2). Thus, a random sample from the set of all possible "ordered topologies" will contain more symmetrical pairings [e.g., ((A,B),(C,D))] than will a random sample from the set of all unordered topologies. A set of unordered topologies would be generated, for example, under a "branch-splitting" model (Furnas, 1984), which does not correspond as well to accepted modes of speciation.

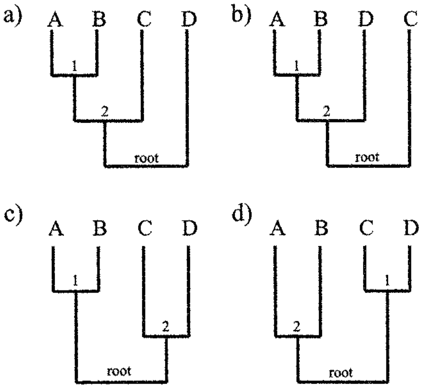

FIGURE 2. The four ordered topologies possible for four taxa under the constraint that taxa A and B be sisters. The standard branching process model produces a random sample from the set of all ordered topologies. A set of unordered topologies for these taxa would not distinguish between trees c and d in the bottom row of this figure. Thus, a set of comparative analyses using a random sample from the unordered topologies would contain fewer symmetrical trees than would analyses conducted using a random sample from the ordered topologies. The suggestion by Losos (1994) of patching together subclades generates a sample from a hybrid population that is neither ordered nor unordered.

When we generate subclades separately and then patch them together into a larger tree, we are forced to ignore the relative order of some nodes and thus do not generate a random sample from the set of all ordered topologies. Suppose, for example, that we are conducting a comparative analysis with four taxa, and we strongly believe that taxa A and B are more closely related to each other than they are to taxa C and D. We can describe this situation as: (A, B), C, D. We might treat the AB clade as a single unit, first creating a set of three-taxon trees and only subsequently resolving the AB branch into two taxa (e.g., Losos, 1994). Larger clades could be treated similarly, generating several subclades independently and then patching them together to form a single large phylogeny.

Unfortunately, it is difficult to do this while also taking into account the relative order of all nodes in the phylogeny. For example, when we patch together the four-taxon tree, we generate only three of the four possible trees that we would have generated by using a branching process model (Fig. 2). The problem is that we have ignored the relative order of the AB node with respect to the rest of the phylogeny. As a result, we inadvertently generate a sample from the set of unordered topologies. With larger phylogenies, the result would be a complex combination of ordered (within each subclade) and unordered (between subclades) topologies that would be difficult to justify as having resulted from any realistic speciation process. Alternatively, we might focus on the relative order of nodes and assign these at random to known subclades. For example, in the four-taxon case, subclade AB can join at either the first or the second node above the root of the tree (Fig. 2). If we choose between these two possibilities at random, however, we would create a set of possible trees in which only one-half of the trees join subclade AB at the more recent node (compared with three-fourths of the trees in Fig. 2), and thereby oversample the symmetrical topologies.

Another way of stating the problem is to point out that we cannot easily generate "branching process" branch lengths for a tree that has been patched together as suggested above (Martins, 1996). Branch lengths in a branching process model are calculated as a function of the number of taxa that exist at that particular moment in time. When patching together subclades, we do not know the total number of taxa existing across all subclades at any one point in time, and cannot generate appropriate branch lengths without also knowing the rate of speciation for the phylogeny as a whole. Note that this sampling issue affects the relative frequencies of different topologies, even when we do not intend to use branch lengths in the later comparative analyses (as discussed above).

One solution is to generate a set of possible phylogenies, using a standard branching process model, and then choose only those taxa that conform to our constraints (Martins, 1996). For example, we could generate all 18 of the possible ordered topologies and use only those 4 that include AB as sister taxa (Fig. 2). Unfortunately, the total number of ordered, binary trees for $N$ taxa is given by:

$$\mathrm{M}_N = \binom{N}{2}\binom{N-1}{2}\cdots\binom{3}{2}\binom{2}{2} \qquad (1)$$

and can be quite large (e.g., Felsenstein, 1978). For eight taxa, more than $10^6$ ordered, binary trees can be formed, and the number of possible trees increases rapidly with increasing numbers of taxa. For larger numbers of taxa, we would need to generate exceedingly large numbers of phylogenies to get only a few that correspond to our constraints. For example, only 42,282 out of the 57,153,600 possible trees for nine taxa satisfy the constraint: (A, B, C), (D, E, F), G, H, I (Fig. 1).

## Our Solution

In this paper, we develop a way to generate phylogenies with constraints in such a way as to take the relative order of the nodes into account, thereby generating a random sample from the set of ordered topologies. Our method involves dividing the population of all ordered topologies for a particular set of taxa with known constraints into distinct categories or groups of trees. These categories are uniquely determined by three factors and identify subsets of trees from which we can sample at random without encountering the problems described above. To generate a random sample of constrained trees, we count the number of trees in each category and use this count to assign a relative weight or frequency to the category. We then generate a random sample of constrained trees by sampling the categories according to their weights and constructing a random tree from each category as it is chosen. We can generate many of these trees, repeat our comparative analyses on each tree, and combine the results to form a broad confidence interval that takes into account our limited phylogenetic knowledge as well as our uncertainty.

For counting purposes, we view a phylogeny as a joining process (going from tips to root). We determine tree categories on the basis of the relative timing of "joining" events (the internal nodes of the phylogeny). For example, Eq. 1 (for $\mathrm{M}_N$, the total number of binary ordered trees on $N$ taxa) comes from counting the number of ways in which taxa can be joined. We begin by counting the ways in which any two of the $N$ taxa can be chosen to join first. We then replace those two with their common ancestor, and count the ways in which two of the remaining $N-1$ nodes can be chosen to join second, and so on.

For any phylogeny with $N$ taxa, there will be $N-2$ internal nodes or joining events between the tips and the root of the tree. Again, in a four-taxon tree, there are four ordered topologies that constrain taxa A and B to be sister (Fig. 2). In this simple example, (A, B), C, D, all possible phylogenies must have two internal joining events plus a final joining at the root of the tree. We can further describe these trees by subdividing them into two categories. The first category of trees include all the trees in which taxa A and B are joined together at the first joining time (nearest the tips, as in Fig. 2a–c). Because there are three possible trees in this category, the category has probability 3/4. The second category includes all trees in which the second joining event is assigned to (A, B). There is only one possible tree of this type (Fig. 2d), so the category has probability 1/4. Once we have enumerated the possibilities, we can generate a set of random trees by choosing the first category 75% of the time (and generating one of the three possibilities within this category at random) and choosing the second category 25% of the time.

The situation is more complex with larger and more numerous subclades. However, three sets of combinatorial quantities can be used to divide the population of possible phylogenies into categories. In general terms, we let $A_j^i$ denote the $i$th taxon in subclade $j$, and $B^1, \ldots, B^n$ be taxa that are required only to be outside of all of the specified subclades. We can describe the constraint of having $r$ subclades in the following form:

$$\left(A_1^1, A_1^2, \ldots, A_1^{k_1}\right), \left(A_2^1, A_2^2, \ldots, A_2^{k_2}\right), \ldots,$$
$$\left(A_r^1, A_r^2, \ldots, A_r^{k_r}\right), B^1, B^2, \ldots, B^n$$

For example, we can describe a set of seven taxa including two known subclades of three taxa each and three additional taxa (Fig. 1) as (A, B, C), (D, E, F), G, H, I. We refer to subclade (A,B,C) as subclade 1, whereas (D, E, F) is subclade 2. There are $N-1=8$ joining events for a tree with nine taxa. Counting from the tips to the root of the tree, the last of these occurs at the root, leaving seven to be distributed throughout the rest of the tree. The population of all possible ordered topologies having this constraint can be divided into categories defined by three factors:

1. **P** is the permutation or relative order of the subclades in terms of the timing of their roots. In general terms, $\mathbf{P} = (p_1, p_2, \ldots, p_r)$, where the subscript on $p$ refers to the order of the nodes from tips to root, whereas the value given to each $p$ identifies a particular clade. For example, the tree in Fig. 2 can be described as a member of the category of trees in which $\mathbf{P} = (p_1, p_2) = (2, 1)$. The first term, $p_1 = 2$, specifies that the root of subclade 2 (D, E, F) occurs first, whereas $p_2 = 1$ indicates that the root of subclade 1 (A, B, C) occurs second (more distant from the tips).
2. **T** is the relative timing of the roots for each subclade with respect to the other internal nodes of the tree. In general terms, $\mathbf{T} = (t_1, t_2, \ldots, t_r)$, $0 < t_1 < t_2 < \cdots < t_r < N-1$, where $t_i$ is the temporal rank for the root of the subclade $p_i$. Because the $t$'s are always in temporal order, we can identify them by counting nodes from the tips to the root of the tree. For example, in Figure 2, the roots of the two specified subclades occur at nodes 3 and 6. Thus, we can describe the tree as belonging to the subset of all possible trees in which $\mathbf{T} = (t_1, t_2) = (3, 6)$. The interpretation of **T** depends on **P**, with $p_1$ providing information about $t_1$, $p_2$ providing information about $t_2$, and so on. For example, $t_1 = 3$ indicates that the earliest subclade to be fully joined does so at node 3. If we also know that $p_1 = 2$, we know that it is subclade 2 (D, E, F) that is fully joined at time 3. In the general situation, **P** also places some restrictions on **T**. For example, because two joining events are required to produce subclade 2 (D, E, F) and because subclade 2 must be fully joined at time $t_1$, $t_1$ must be greater than or equal to 2.
3. **C** is the composition of number of joining events involving external taxa (those outside all specified subclades) between the subclade roots. In general, $\mathbf{C} = (c_1, c_2, \ldots, c_r)$ where $c_i$ is the number of joining events that occur between the

$i - 1$st and the $i$th subclade roots on the unconstrained part of the tree. The subscripts again relate the three types of descriptions (**P**, **T**, and **C**), with $c_1$ providing information about the subclade named in $p_1$. The phylogeny in Figure 2 thus belongs to a set of possible trees with **C**$=$ (0, 2). The number of joining events in the first interval ($c_1$) is the number of joining events occurring among the external taxa between the tips of the tree and the first subclade root. Thus, $c_1 = 0$ specifies there are no joining events involving external taxa before the $p_1$ subclade (D, E, F) is completely joined. In Figure 1, two joining events involving external taxa occur between the roots of the $p_1$ (D, E, F) and $p_2$ (A, B, C) subclades, and we describe this as $c_2 = 2$.

As above, the total number of possible trees defined by **P**, **T**, and **C** is limited by interactions between these factors. For example, two joining events are required to join subclade (D, E, F). Thus, if the root of (D, E, F) were to occur at the second joining event ($t_1 = 2$), the first joining event must also be used in the resolution of subclade (D, E, F), and thus $c_1$ must be 0. More restrictions on **P**, **T**, and **C** can be determined by noting that all the subclades must be fully joined by time $t_r$. There are $k_i - 1$ joining events within each subclade. Thus, the total number of joinings by time $t_r$ must equal the total number of joinings on the unconstrained part of the tree, $c_1 + c_2 + \ldots + c_r$, plus the total number of joinings needed for each subclade, $(k_1 + k_2 + \ldots + k_r - r)$. Thus $t_r = c_r + c_2 + \ldots + c_r + (k_1 + k_2 + \ldots + k_r - r)$. Note that each triplet (**P, T, C**) defines a unique subset or category of possible trees. Given a category defined by specific values of **P, T**, and **C**, we can generate a random branching process tree from within that category (see below).

### *Forming and Counting Constrained Trees*

Specifying taxonomic constraints is equivalent to identifying which of the categories specified by **P, T**, and **C** contain trees, i.e. to identifying triplets **P, T**, and **C** which satisfy the restrictions described above. Different categories defined by **P, T**, and **C** may contain different numbers of trees. To sample randomly from the set of all possible ordered topologies with a specified constraint, we must first enumerate all possible tree categories defined by **P, T**, and **C**, and count the number of trees contained within each category. There are standard algorithms for listing all possible permutations, **P**, and compositions, **C** (e.g., Nijenhuis and Wilf, 1978), and we have developed a similar algorithm for enumerating the temporal rank of the subclade roots, **T** (see Appendix). Using these algorithms, we can list all the possible categories for a phylogeny under the specified constraints. Once we have counted the number of trees in each category (see below), we divide this by the total (summed across categories) to get the relative frequency of each category in the population as a whole. We can then choose one of the categories at random (given its relative frequency in the population) and construct a random phylogeny that is one of many possible phylogenies within that category. We repeat this last step for each random tree desired.

To form a tree within a category given by **P, T**, and **C**, as well as count the number of possible trees in that category,

Step 1. Begin by forming a list of $N-1$ possible absolute times, where $N-1$ is the number of internal nodes on the tree. To create this list, we draw $N-1$ uniform random numbers and transform them by using the inverse of the cumulative distribution function for an exponential distribution with rate $(N - i)s$:

$$m_i = -\ln(x)/[(N - i)s]$$
$$i = 1, 2, \ldots N - 1$$

where $x$ is the uniform random number and $s$ is the speciation rate. (Note that use of this formula with an arbitrary speciation rate [e.g., $s = 1$] gives relative branch lengths on the phylogeny under a standard branching process model. It would also be reasonable to apply a coalescent, to incorporate extinction, or to apply any other model to generate branch lengths. See Martins [1996] for further discussion.) The branch length values correspond to distances between nodes on the phylogeny, which we then sum to obtain distances between each node ($v$) and the tips of the tree:

$$b_v = \Sigma_{i=1}^{v} m_i$$

Step 2. Begin to form the tree, starting at the tips and going to the root of the first subclade to be fully joined (indicated by the ranked time $t_1$).

(a) We focus on the external nodes first, by randomly choosing $c_1$ of the branch lengths ($b_v$) where $v$ is less than $t_1$. There are $\binom{t_1-1}{c_1}$ ways to choose these branch lengths.

(b) We then randomly join the external nodes in this section of the tree, using those $c_1$ branch lengths. There are $\binom{n}{2}\binom{n-1}{2}\cdots\binom{n-c_1+1}{2}$ possible ways to join these external nodes, where $n$ is the number of external taxa (outside any specified subclade) on the tree as a whole. Note that this formula is like Eq. 1, but we now end when $c_1$ joining events have been made. There is only one way to join the external nodes if $c_1 = 0$ and the product is empty.

(c) We now choose the remaining branch lengths needed to join all the taxa within subclade $p_1$, with $b_{t_1}$ being assigned to the root of that subclade. We need $k_{p_1} - 2$ of the remaining $t_1 - 1 - c_1$ times and there are $\binom{t_1 - 1 - c_1}{k_{p_1} - 2}$ ways to choose them. This leaves $t_1 - c_1 - k_{p_1} + 1$ branch lengths in the top section of the tree (between the tips and time $t_1$) to be used to join other subclades.

(d) We then use the times chosen above to join the taxa in subclade $p_1$. There are $\mathrm{M}_{k_{p_1}}$ ways to form a tree with $k_{p_1}$ taxa (Eq. 1).

Step 3. Move down to the ranked time indicated by $t_2$ at the root of the second subclade to be fully joined.

(a) Again, we focus on the external nodes first, randomly choosing $c_2$ of the branch lengths ($b_v$), where $v$ is strictly between $t_1$ and $t_2$. There are $\binom{t_2 - t_1 - 1}{c_2}$ ways to choose these branch lengths.

(b) We then randomly join the external nodes using those times. There are $n - c_1$ of the original external nodes remaining after the $c_1$ joining events were made in Step 2, and one additional node for the ancestor of subclade $p_1$ has been added. Thus, there are $\binom{n - c_1 + 1}{2}\binom{n - c_1}{2}\cdots\binom{n - c_1 - c_2 + 2}{2}$ possible ways to join these external nodes.

(c) We now choose the remaining times needed to join all the taxa within subclade $p_2$ with $b_{t_2}$ being assigned to the root of that subclade. There were $t_1 - c_1 - k_{p_1} + 1$ times in the top section of the tree left over from Step 2. There are $t_2 - t_1 - 1$ branch lengths between $b_{t_2}$ and $b_{t_1}$, and we used up $c_2$ of them with external nodes. Thus, there are $t_2 - c_1 - c_2 - k_{p_1}$ times remaining. There are $\binom{t_2-c_1-c_2-k_{p_1}}{k_{p_2}-2}$ ways to choose the $k_{p_2} - 2$ of them that we need to join the taxa in subclade $p_2$.

(d) Finally, we use those times to join the taxa in subclade $p_2$. There are $\mathrm{M}_{k_{p_2}}$ ways to form a tree with $k_{p_2}$ taxa (Eq. 1).

Step 4. Repeat Step 3 until the last subclade is fully joined. Restating Step 3 in more general terms, we choose $c_j$ of the branch lengths ($b_v$) between $t_{j-1} + 1$ and $t_j - 1$ and there are $\binom{t_j - t_{j-1} - 1}{c_j}$ ways to choose these lengths. We will join the external nodes (including now the ancestors of subclades $p_1, p_2, \ldots p_{j-1}$), using those times. There are $n - c_1 - c_2 \ldots - c_{j-1} + (j-1)$ external nodes at this stage and thus

$$\prod_{i=1}^{c_j}\left(\frac{n - c_1 - c_2 - \cdots - c_{j-1} + j - i}{2}\right)$$

ways to join the external nodes. We choose $k_{p_j} - 2$ of the remaining branch lengths ($b_v$) with $v \le t_j - 1$. There are $(t_j - 1) - c_1 - c_2 - \cdots - c_j - (k_{p_1} - 1) - (k_{p_2} - 1) - \cdots - (k_{p_{j-1}} - 1) = t_j - c_1 - c_2 -$

$\cdots - c_j - k_{p_1} - k_{p_2} - \cdots - k_{p_{j-1}} +$ $(j-2)$ remaining lengths, and thus $\binom{t_j - c_1 - c_2 - \cdots - c_j - k_{p_1} - k_{p_2} - \cdots k_{p_{j-1}} + (j-2)}{k_{p_j} - 2}$ ways to choose them.

Step 5. We now join the remaining external taxa and the subclade roots at random until the entire tree is resolved. At this point in time, there are $n - c_1 - c_2 - \cdots - c_r + r$ nodes and thus there are $\binom{n-c_1-c_2-\cdots-c_r+r}{2}\binom{n-c_1-c_2-\cdots-c_r+r-1}{2}\cdots\binom{2}{2}$ ways to complete the tree.

Step 6. The total number of trees in a category determined by **P**, **T**, and **C** is the product of all the combinatorial terms above. Because **P** is a permutation, this product always contains a factor of $\mathbf{M}_{k_1}\ \mathbf{M}_{k_2} \cdots \mathbf{M}_{k_r}\ \mathbf{M}_n$ which cancels out of the equation when determining the relative weight of the category. Thus we define the relative weight of a category determined by **P**, **T**, and **C** to be what remains of the product after this common factor is removed. This weight can be written as:

$$W_{P,T,C} = \binom{t_1-1}{c_1}\binom{n-c_1+1}{2}\binom{t_1-1-c_1}{k_{p_1}-2} \times \prod_{j=2}^{r}\left[\binom{t_j-t_{j-1}-1}{c_j}\binom{n-\sum_{i=1}^{j}c_i+j}{2} \times \binom{t_j-\sum_{i=1}^{j}c_i-\sum_{i=1}^{j-1}k_{p_i}+(j-2)}{k_{p_j}-2}\right] \tag{2}$$

To sample at random from the population of all trees with a specified constraint, we should choose a tree from a particular category with probability given by the weight of that category divided by the sum of weights across all categories: $W_{\text{TOTAL}} = \Sigma W_{\mathbf{P,T,C}}$.

Minor modifications to our approach can also be used to extend its usefulness. For example, we might be interested in a set of taxa grouped as: $((A_1, A_2, \ldots, A_k), B_1, B_2, \ldots, B_m), C_1, C_2, \cdots, C_n$. To do so, we can apply the above procedure recursively. We begin by considering $(A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_m), C_1, C_2, \cdots, C_n$ and choose the times to use in joining $(A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_m)$ together. We then apply the above procedure to redistribute those times to form a tree constrained as $(A_1, A_2, \ldots, A_k), B_1, B_2, \cdots, B_m$. Combining the two trees gives us a final tree with a nested constraint. The above procedures have been implemented in COMPARE 4.3 (Martins, 2000).

## An Example

Once a random sample from the set of all possible constrained, yet ordered, topologies is available, we can use these topologies to incorporate phylogenetic uncertainty into our comparative analyses. For example, we might generate 1,000 possible trees by using a branching process model and conduct a phylogenetic comparative analysis (e.g., Felsenstein [1985] contrasts) on each of the possible trees to obtain 1,000 estimates of statistical parameters (e.g., correlation coefficient, regression slope) for our data. We can then use the procedure outlined in Martins (1996) to combine these estimates into a single number with a broad confidence interval that assumes the branching process model is a reasonable description of the speciation process (as opposed to assuming that any single tree is entirely correct).

For example, consider the relationship between two hypothetical traits measured in nine taxa (Fig. 1). When the independent contrast analysis is done on a single phylogeny that is assumed to be correct, we estimate a regression slope of 1.5 (SE = 0.50) which is significantly different from zero (Table 1). In the terms used in Martins (1996), the phylogenetic variance (that due to uncertainty in the phylogeny) is assumed to be zero ($\text{Var}_\text{P} = 0$). Thus the residual or sampling variance ($\text{Var}_\text{S} = 0.50^2 = 0.25$, a result of the points not falling exactly on the regression line) alone describes our uncertainty in the results.

Alternatively, we can use the method in Martins (1996) and assume that no phylogenetic information is available. We calculate the standard error by averaging across the residual variances for analyses on all 1,000 trees ($\text{Var}_\text{S} = 0.36$), adding this to the phylogenetic variance (variance in regression slopes calculated across the 1,000 trees, $\mathrm{Var}_P = 1.33$), and taking the square root. The result is a similar estimate of the regression slope (1.4; Table 1), but with a larger standard error (1.30), reflecting our lack of information regarding the phylogeny. Note that the larger standard error primarily reflects an increase in the phylogenetic variance ($\mathrm{Var}_P$). By this analysis, the regression slope is not significantly different from zero, even though a quick look at the raw data (Fig. 1) suggests that the relationship is quite strong. Thus, without further phylogenetic information, we are forced to conclude that the evidence of a relationship is not very strong.

The method described in the current paper allows us to incorporate some limited information about the topology. Imagine, for example, that we know only that A, B, and C belong together in one clade, that D, E, F form a second clade, and that the three remaining taxa fall outside of these subclades: (A, B, C), (D, E, F), G, H, I. We can apply the above procedure to generate 1,000 possible trees with this constraint under a branching process model of speciation. The result is 1,000 regression slopes, the mean of which (1.5) is again quite similar to the regression slopes obtained above. We can calculate the residual variance ($\mathrm{Var}_S = 0.231$), as above, by taking the average across all 1,000 analyses, and estimate the phylogenetic variance ($\mathrm{Var}_P = 0.479$) as the variance among slopes calculated for the 1,000 trees. As expected, the phylogenetic variance for the constrained phylogenies is intermediate between that calculated for the single-tree analysis ($\mathrm{Var}_P$ assumed to equal 0) and the analysis on completely random phylogenies ($\mathrm{Var}_P = 1.33$). Summing these and taking the square root, we find that the result is an estimate of the regression slope that is once again significantly different from zero.

The magnitude of $\mathrm{Var}_P$ will depend on the amount and quality of the phylogenetic information that is incorporated (Table 1). Often, we also expect $\mathrm{Var}_S$ to increase slightly. The analysis on the single tree produces a $\mathrm{Var}_S$ that might be viewed as a single sample from the distribution of $\mathrm{Var}_S$ for the 1,000 trees generated without constraints. Because variances tend to be skewed to the right, any single element of that distribution is likely to be less than the mean of the distribution as a whole. Thus, $\mathrm{Var}_S$ for a single tree analysis is likely to be smaller than $\mathrm{Var}_S$ for 1,000 trees generated without constraints. $\mathrm{Var}_S$ for the 1,000 constrained trees is likely to be more similar to $\mathrm{Var}_S$ for the single-tree analysis if the single tree exhibits the same constraints as the computer-generated sample. $\mathrm{Var}_S$ for any of these analyses may also be small if the data match the phylogenies on which the analyses were done.

All of the above analyses were conducted by using COMPARE 4.3 (Martins, 2000). On a standard desktop computer, this took 2 min to generate 1,000 trees with the specified constraints, and an additional 30 s to calculate Felsenstein contrasts on those 1,000 trees.

## Discussion

In this paper, we describe a method for generating random phylogenies with constraints. Although we focus our discussion on how such phylogenies can be used to incorporate even small amounts of systematic information into phylogenetic comparative analyses, it may also be useful in other statistical randomization tests. Briefly, in the comparative method, this procedure provides a way to increase the variance about a parameter estimate to take our uncertainty in the phylogeny into account while also not ignoring small amounts of taxonomic or other information in which we are confident. As shown in our example, even a little phylogenetic information can make a big difference in a comparative hypothesis test. Computer-generated phylogenies and randomization tests may also be useful in a variety of other contexts, whenever a distribution of possible trees is needed to conduct hypothesis tests or provide confidence intervals. For example, distributions of trees may be useful in comparing phylogenetic hypotheses derived from different genes or in using phylogenies to test for cospeciation.

Of course, whenever a real phylogeny is available, it should be used. For comparative analyses, if molecular data are available, we advocate using Felsenstein's suggestions of bootstrap trees in combination with the statistical method in Martins (1996) for taking this information into account in calculation of confidence intervals, or perhaps newer Bayesian procedures (e.g., Huelsenbeck et al., 2000). The method described in the current paper is intended primarily for those difficult situations when only a taxonomy is available or when available phylogenies are mostly unresolved.

Our method is superior to other polytomy reduction procedures suggested for comparative analyses in that it incorporates phylogenetic uncertainty into our final confidence intervals and does so in a way that is statistically explicit. Other suggestions that reduce polytomies but do not incorporate that variation into the final parameter estimates (e.g., Grafen, 1992; Pagel, 1992) may still provide useful heuristic tools, for example, if one type of reduction seems more probable than the others. Purvis and Garland's (1993) suggestion that we reduce the degrees of freedom associated with hypothesis tests in phylogenetic comparative analyses to account for polytomies in the tree is also vaguely similar to our method. The test for whether a regression slope is significantly different from zero depends on the mean square error, which is estimated as the variance for the slope divided by the degrees of freedom. Thus, reducing the degrees of freedom (as suggested by Purvis and Garland, 1993) has an effect similar to that of increasing the variance (our suggestion). However, Purvis and Garland (1993) suggest only that we consider a range of possible degrees of freedom, without offering an explicit formula for the correct degrees of freedom. In contrast, we give an explicit way of estimating the actual increase in variation associated with phylogenetic uncertainty.

### *Possible Concerns*

Although our method works well with as many as four subclades and tens of taxa, we have found that increasing the number of constraints or the number of taxa beyond this is computationally difficult. Alternatively, for large trees with many nonnested constraints, a Markov chain technique such as the Metropolis–Hasting algorithm (Gelman et al., 1995; Kuhner et al., 1995) might be developed to generate random constrained trees. To apply a Markov chain technique of this sort, we would not have to calculate $W_{TOTAL}$ but would instead only calculate the relative weight of a given category, $W_{P,T,C}$, using Eq. 2 above.

Another possible worry is that our method focuses entirely on the branching process model of speciation, whereas real phylogenies may not conform to this model (e.g., Abouheif, 1998). If the true phylogeny is highly unusual or if the speciation process is different from how it is currently envisioned by most evolutionary biologists, the branching process model will not be appropriate and the above method may be biased. We do not see this as a serious concern because the branching process model has been well justified in the evolutionary literature as a model of speciation (e.g., Slowinski and Guyer, 1989; Maddison and Slatkin, 1991; Hey, 1992; Losos, 1994; Martins, 1996; Chu and Adami, 1999) and seems eminently reasonable. Our method can also easily be extended to include extinction (e.g., as in Harvey and Rambaut, 1998).

Finally, in this paper, we ignore many other possible sources of error in a phylogenetic comparative method (Martins and Hansen, 1997). Note in particular that the method above does not explicitly vary branch lengths in a way that conforms to models of character evolution and does not calculate confidence intervals that take uncertainty in the model of character evolution into account. The branch lengths from the above procedure are reasonable estimates of the relative amount of time that has occurred between speciation events on a tree. Most comparative methods, though, require a phylogeny with branch lengths in units of the expected amount of character change (not in units of time). Each trait may undergo more or less evolutionary change depending on the types of evolution to which it is subjected (Hansen and Martins, 1996). For example, under Hansen's (1997) model of adaptive evolution, we expect traits to change as an exponential function of time rather than the usual linear (gradual) function assumed by most comparative analyses (e.g., Felsenstein contrasts). Given phylogenies and branch lengths in units of time (e.g., as a result of applying our method for generating random trees with constraints), we must still then apply a transformation to the branch lengths to reflect this evolutionary model.

## Acknowledgments

We thank the referees, Joseph Slowinski and Junhuong Kim, for their helpful comments. We are also grateful to Xingxing Wu for writing the computer code necessary to incorporate our method into COMPARE. This work was supported by National Science Foundation grants

DMS-0075143 (E.A.H.) and DEB-9720641 (E.P.M.).

## References

Abouheif, E. 1998. Random trees and the comparative method: A cautionary tale. Evolution 52:1197–1204.

Aldous, D. 1995. Darwin's log: A toy model of speciation and extinction. J. Appl. Prob. 32:279–295.

Chu, J., and C. Adami. 1999. A simple explanation for taxon abundance patterns. Proc. Natl. Acad. Sci. USA 96:15017–15019.

Felsenstein, J. 1978. The number of evolutionary trees. Syst. Zool. 27:27–33.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Furnas, G. W. 1984. The generation of random, binary unordered trees. J. Classif. 1:187–233.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian data analysis. Texts in statistical science. Chapman & Hall, London.

Grafen, A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. London B: Biol. Sci. 326:119–157.

Grafen, A. 1992. The uniqueness of the phylogenetic regression. J. Theoret. Biol. 156:405–424.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. Evolution 40:1404–1417.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford Univ. Press, Oxford, England.

Harvey, P. H., and A. Rambaut. 1998. Phylogenetic extinction rates and comparative methodology. Proc. R. Soc. London B 265:1691–1696.

Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46:627–640.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. 7:1–44.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288:2349–2350.

Kingman, J. F. C. 1982. The coalescent. Stochastic Proc. Appl. 13:235–248.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. Genetics 140:1421–1430.

Losos, J. B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. Syst. Biol. 43:117–123.

Maddison, W. P., and M. Slatkin. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution 45:1184–1197.

Martins, E. P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. Evolution 50:12–22.

Martins, E. P. 2000. COMPARE: For the statistical analysis of comparative data. Version 4.3. (Computer programs distributed by the author at http://compare.indiana.edu). Indiana Univ. Bloomington.

Martins, E. P., and Hansen, T. F. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into analysis of interspecific data. Am. Nat. 149:646–667 (erratum. Am. Nat. 153:448).

Nijenhuis, A., and H. Wilf. 1978. Combinatorial algorithms for computers and calculators. Academic Press, New York.

Page, R. 1991. Random dendrograms and null hypotheses in cladistic biogeography. Syst. Zool. 40:54–62.

Pagel, M. 1992. A method for the analysis of comparative data. J. Theor. Biol. 156:431–422.

Purvis, A., and T. Garland, Jr. 1993. Polytomies in comparative analyses of continuous data. Syst. Biol. 42:569–575.

Slowinski, J. B., and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: An improved null model. Am. Nat. 134:907–921.

*Received 26 July 2000; accepted 23 January 2001*
*Associate Editor: R. Olmstead*

## Appendix

The algorithm for generating the rank of the roots for the subclades, **T**, assumes that a permutation, **P**, has already been chosen. Algorithms for generating permutations are readily available (Nijenhuis and Wilf, 1978.) We then generate the last entry in **T**, $t_r$, as a number at least $k_1 + k_2 + \cdots + k_r - r$ and no more than $k_1 + k_2 \cdots + k_r + n - 2$. Once $t_r$ is chosen, we have from the formula $t_r = c_1 + c_2 + \cdots + c_r + k_1 + k_2 + \cdots + k_r - r$ that $c = c_1 + c_2 + \cdots + c_r = t_r - (k_1 + k_2 + \cdots + k_r - r)$ is fixed. We generate a composition of $c$, $\mathbf{C} = (c_1, c_2, \ldots, c_r)$ where $c = c_1 + c_2 + \cdots + c_r$, by a standard algorithm (Nijenhuis and Wilf, 1978). Not every composition represents a possible tree. If $c$ is greater than $n - 1$, where $n$ is the number of taxa external to the subclades, then the composition $(c, 0, 0, \ldots, 0)$ is not possible, for instance, because we cannot join $n$ taxa more than $n - 1$ times. The irrelevant possibilities for **C** may be eliminated by initializing **C** in such a way as to avoid them. In particular, if $c$ is no more than $n - 1$, we initialize **C** to be the standard $(c, 0, 0, \ldots, 0)$, but if $c$ is greater than $n - 1$, we initialize **C** to be $(n - 1, 1, 1, \ldots, 1, 0, 0, \ldots, 0)$ where the sum is still required to be $c$.

We then use **P, C**, and $t_r$ to generate all possibilities for the ranks of the roots $\mathbf{T} = (t_1, t_2, \ldots, t_r)$ with $t_r$ fixed and **P** and **C** given, taking into account the interdependence of **P, C**, and **T** via the following subroutines:

[We generate the initial possibility of **T** which makes each $t_i$ as small as possible. We generate the final possibility of **T** which makes each $t_i$ as large as possible and helps us determine when to increment a new section of **T** and when to stop because we have generated all the possibilities. We generate a global reference **T** which will help us renew **T** when we begin incrementing a new section of **T**. To begin with, the reference **T** is just the initial **T**.] Subroutine Initial_T called with **P, C**, and $t_r$

```
$t_1 = k_{p_1} - 1 + c_1$;
reference_$t_1 = t_{1;}$
final_$t_r = t_r$;
final_$t_{r-1}$ = final_$t_r - c_r - 1$;
For $i = 2$ to $r - 1$ {
    $t_i = t_{i-1} + k_{p_i} - 1 + c_i$;
    reference_$t_i = t_i$;
    final_$t_{r-i}$ = final_$t_{r-i+1} - c_{r-i+1} - 1$;
    }
  Return T = $(t_1, t_2, \ldots, t_r)$;
```

Subroutine Next_T called with **P, C**, and **T**, where **T** was the last set of ranked root times generated, which we now want to update.

```
  \* We have generated all the possible root rankings
    and therefore want to stop when $t_1$ is as large as
    possible, which forces all the other $t_i'$s to be as large
    as possible too.*/
  If ($t_1$ == final_$t_i$){quit;}
  \* We find the last position in T we have not finished
    cycling through.*/
h = $r - 1$;
While ($t_h$ == final_$t_h$){$h = h - 1$;}
\* Then we need to increment T at that position.*/
  $t_h = t_h + 1$;
\* If $h = 1$, we need to update the reference T as
  above.*/
If ($h == 1$) {For $i = 2$ to $r - 1$
          If (reference_$t_i$ < reference_$t_{i-1} + c_i + 1$)
      {reference_$t_i$ = reference_$t_{i-1} + c_i + 1$;}
  }         }
\* We need to reset the positions in T after position $h$,
  keeping the constraints in mind.*/
For $i = h + 1$ to $r - 1$ {
        $t_i$ = reference_$t_i$;
        If ($t_i < t_{i-1} + c_i + 1$){$t_i = t_{i-1} + c_i + 1$;}
            }
Return T = $(t_1, t_2, \ldots, t_r)$;
```