Syst. Biol. 61(5):811–821, 2012
© The Author(s) 2012. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/sys033
Advance Access publication on February 15, 2012

The Probability of Correctly Resolving a Split as an Experimental Design Criterion in Phylogenetics

EDWARD SUSKO[1,*] AND ANDREW J. ROGER[2]

[1]*Department of Mathematics and Statistics; and* [2]*Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7;*
[*]*Correspondence to be sent to: Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; E-mail: susko@mathstat.dal.ca.*

Received 8 November 2011; reviews returned 29 December 2011; accepted 18 January 2012
Associate Editor: Mark Holder

Abstract.—We illustrate how recently developed large sequence-length approximations to probabilities of correct phylogenetic reconstruction for maximum likelihood estimation can be used to evaluate experimental design strategies. The specific criterion of interest is the probability of correctly resolving an a priori defined split of interest in a phylogenetic tree. Design strategies considered include increased taxon sampling and increasing sequence length. Our analyses of specific examples strongly suggest that it is better to sample taxa that connect as close as possible to the split of interest. Assuming this can be done, these examples suggest it is better to sample additional taxa than to add a comparable number of sites for the existing taxa. If the rates of evolution in the added taxa are slow, it is better to choose taxa connecting to a long edge, but if rates are comparable to a sister lineage, it is not necessarily the best strategy to sample taxa connected to a long edge. We also examined deleting taxa while increasing the number of sites. Although deleting a small number of taxa distant from the split of interest can be beneficial, deleting too many or making poor choices as to what should be deleted can lead to smaller probabilities of correct reconstruction than for the original sequence data. [Experimental design; phylogenetics; taxon sampling.]

The effects on phylogenetic accuracy of adding taxa or sites to an existing alignment have long been of interest (Graybeal 1998; Yang 1998; Pollock et al. 2002; Zwickl and Hillis 2002). While the consensus from these studies is that additional taxon sampling improves phylogenetic accuracy, they have usually employed expensive simulation approaches. Consequently, these studies do not provide real-time tools for investigating how adding branches to a tree as an idealized proxy for taxon sampling might improve phylogenetic accuracy.

Experimental design in phylogenetics was pioneered in Goldman (1998) and Geuten et al. (2007). In both studies, the expected, or Fisher, information matrix (the average second derivative matrix of the log likelihood multiplied by -1) was used to determine the criteria by which choices of taxon sampling were to be evaluated, avoiding the need for expensive simulations. Goldman (1998) showed how to increase information about a divergence time in a clock-like phylogeny when adding a sequence and how to select the substitution rate optimally to minimize the variance of edge-length estimation. Geuten et al. (2007) developed strategies for topological estimation. Their approach maximizes criteria, which are calculated as transformations of the information matrix; several transformations are considered, consistent with those used more generally in the statistical theory of experimental design (Kiefer 1959; Atkinson and Donev 1992). The transformations considered in Geuten et al. (2007) lack interpretive value, however. While they can indicate the best location for edge addition, it is not clear how much worse is another location that almost maximized the transformation. Similarly as in the Geuten et al. (2007) approach, our criteria can be considered as transformations of information matrices, but these transformations are probabilities of correct reconstruction. Thus, one can directly consider how far from optimal an alternative strategy is.

In cases where the goal of experimental design is phylogenetic reconstruction, it is more direct to choose the design that maximizes the probability of correct reconstruction. Such approaches have implicitly been considered. Yang (1998) considers the probability of correct reconstruction as a function of evolutionary rates, whereas Graybeal (1998) obtains the probability of correct reconstruction after adding sites or taxa. Both utilize simulation, however. This poses substantial computational challenges for real-time investigation of design strategies when many taxa, complex models, and a variety of settings are involved. For each possible design choice considered, be it adding taxa or a gene with a different rate of evolution, repeated simulation is required. Repeat data sets must be generated and, for each of these, tree estimation is carried out. Tree reconstruction via maximum likelihood (ML) estimation is computationally intensive since the computation of a likelihood requires repeated application of the pruning algorithm of Felsenstein (1981).

In Susko (2011), expressions are given for the probability of correct reconstruction in terms of trivariate normal probabilities that can be computed quickly using numerical integration algorithms like those of Genz (2004). No repeated simulation of large sequences with repeated ML estimation is required. The setting of Susko (2011) requires some restrictions, however. It is assumed that there is a single split of interest that is poorly resolved. We also assume that it is the backbone of the tree without the additional taxon that is of interest. Not imposing this restriction would complicate comparison of designs as it changes the target topologies of interest. A similar restriction is made in Geuten et al. (2007).

Materials and Methods

An example that will be considered presently is the seed plant phylogeny of Geuten et al. (2007) reproduced in Figure 1. Probabilities of correct reconstruction will be calculated for two splits labeled X and Y under various scenarios about where one might be able to sample additional taxa. To consider the merits of design strategies for alternative splits of interest, one can traverse the tree. For illustration, consider the split X under the hypothesis that it is poorly resolved. Due to its statistical consistency properties, the ML tree will eventually correctly estimate the Monocot and "Other" subtrees, or at least that these groups are separated from each other. Thus, if the four subtrees neighboring it are clearly separated from each other, with large sequence lengths, the three trees in Figure 2 will be the only trees with appreciable probability of being estimated. More generally, we assume that the generating tree is Topology 1 of Figure 2 and that groups of taxa in the subtrees 1–4 are separated from each other. It follows that, with large sequence lengths, the topology estimated will be one of the Topologies 1–3. Our design criterion is an approximation to the probability that the correct topology among these three is estimated. To investigate sampling strategies, we can calculate the probability of correctly estimating the topology

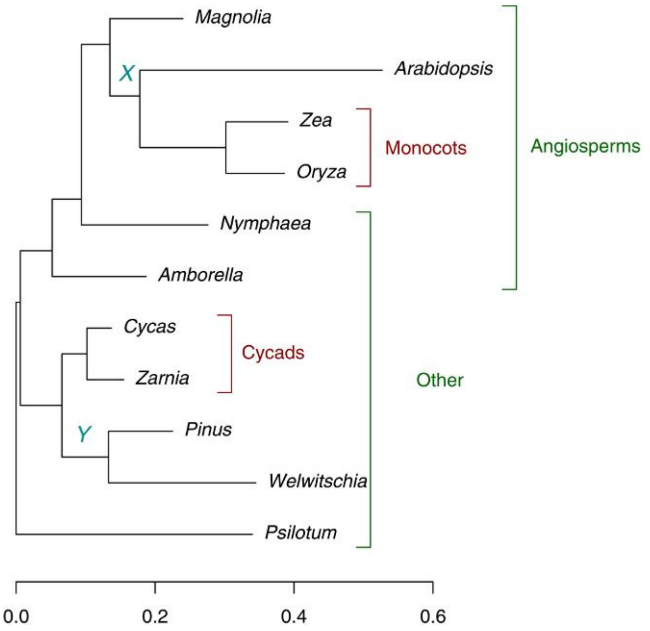

FIGURE 1. The seed plant phylogeny with splits of interest labeled X and Y. A split of interest gives rise to three different possible resolved topologies determined by the four edges nearest to it. For the split X, these have *Arabidopsis* sister to either *Magnolia*, the Monocots, or the group labeled Other. For the split Y, these have *Welwitschia* sister to either *Pinus*, the Cycads, or the remaining group containing *Psilotum* and the *Angiosperms*.

1. with longer sequences.
2. with additional branches added to the tree as a proxy for additional taxon sampling.
3. with taxa deleted but with additional sites.

Examples will be given in the results section for each of these applications. The criteria can also be used as an additional measure of uncertainty in a given phylogenetic analysis. Calculating the probability with a smaller middle edge than the estimated one (for instance, the lower bound of a 95% confidence interval for the middle edge) gives a measure of the certainty of estimation.

In the general setting of Figure 2, the topology estimated will be one of the Topologies 1–3 with Topology 1 being correct. We consider a range of lengths for the middle edge in Topology 1, denoted $t_{0m}^{(n)}$, decreasing to 0 as a function of the sequence length n; m indexes the middle edge. The other edges in the tree remain fixed as a function of sequence length and the edge-lengths leading to the subtrees 1–4 are positive. To ensure that the probabilities of reconstruction are not simply 1 regardless of design considerations, it is necessary to have $t_{0m}^{(n)}$ decrease to 0 but at a rate that is not too slow. It turns out that having $t_{0m}^{(n)} \approx a/\sqrt{n}$, for some constant a, is the appropriate choice to avoid trivial limiting probabilities of correct reconstruction.

Let $P_0(\boldsymbol{d}, A)$ denote the probability that a normal random vector (X, Y, Z) with a mean $\boldsymbol{d}$ and covariance matrix A has all positive elements: $X > 0$, $Y > 0$, and $Z > 0$. The theory developed in Susko (2011) gives the probability that topology j is estimated as

$$P_0(\boldsymbol{d}^{(1)}, A^{(1)}) + P_0(\boldsymbol{d}^{(2)}, A^{(2)}), \quad (1)$$

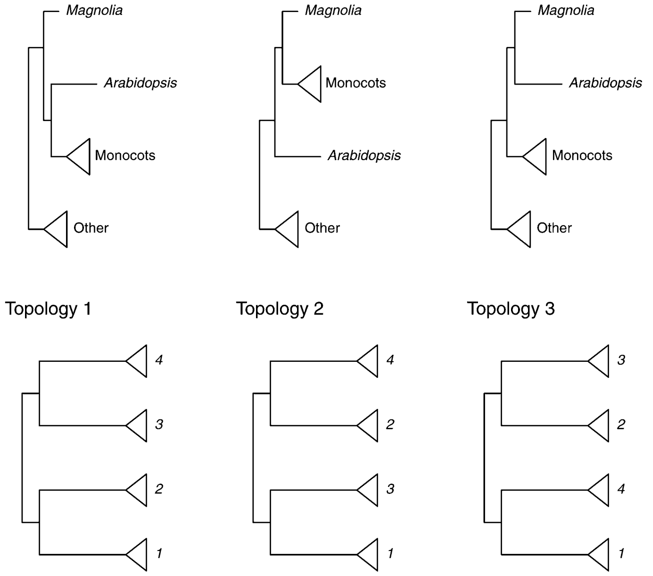

FIGURE 2. The competing topologies when the split of interest X in Figure 1 are not well-resolved. More generally, there may be any number of taxa in the four groups determined by the split of interest and its neighboring edges, leading to three competing topologies labeled 1–3.

for some particular values of $\boldsymbol{d}^{(1)}$, $\boldsymbol{d}^{(2)}$, $A^{(1)}$, and $A^{(2)}$ whose derivation we will discuss presently. Calculation of the limiting probabilities thus requires only the ability to calculate probabilities for a trivariate normal distribution. This can be accomplished using the numerical integration methods of Genz (2004).

The expression (1) arises from approximations to the differences, $\triangle l_j$, between the maximized log likelihood for the jth topology and the maximized log likelihood for that topology but with the middle edge-length set to 0. Up to terms that will be small for large sequence lengths, $\triangle l_j$ is determined by what can be referred to as a standardized score, V^c_{jn}. This quantity is based on the first derivative of the log likelihood with respect to the length of the middle edge, evaluated when that edge-length is set to 0. A correction, described in (3) of Susko (2011), is made for estimation of the other edge-lengths in the tree (which is the reason for the superscript c), and the quantity is standardized to have variance 1. When $V^c_{jn} < 0$, the optimal middle edge-length is 0 and $\triangle l_j$ is 0. Otherwise, $\triangle l_j = [V^c_{jn}]^2/2$, up to terms that will be small for large sequence lengths.

The difference between the log likelihoods for the jth and rth topologies is the same as $\triangle l_j - \triangle l_r$. The reason for this is that the log likelihoods with middle edge-lengths set to 0 are the same for the two trees and cancel in $\triangle l_j - \triangle l_r$, leaving the difference in optimized log likelihoods for the two topologies. Because $\triangle l_j$ is 0 when $V^c_{jn} < 0$, topology j will never be preferred to r if $V^c_{jn} < 0$. Thus, there are two cases where the topology j may be preferred to r: (i) $V^c_{jn} > 0$ and $V^c_{rn} > 0$, in which case the log likelihood difference is $[V^c_{jn}]^2/2 - [V^c_{rn}]^2/2$; or (ii) $V^c_{jn} > 0$ and $V^c_{rn} < 0$, in which case the log likelihood difference is $[V^c_{jn}]^2/2$. In case (ii), the log likelihood difference is always positive, so that topology j is always preferred. In the case (i), since V^c_{jn} and V^c_{rn} are both positive, the approximate log likelihood difference, $[V^c_{jn}]^2/2 - [V^c_{rn}]^2/2$, is positive if and only if $V^c_{jn} > V^c_{rn}$. Since V^c_{jn} is always larger than V^c_{rn} in case (ii), we can succinctly summarize the condition under which j is preferred to r as

$$V^c_{jn} > 0, \quad V^c_{jn} - V^c_{rn} > 0. \tag{2}$$

Topology j is estimated if it is preferred to the two other topologies, say r and s. We can break this into mutually exclusive events (so that the corresponding probabilities can be summed) as (i) topology j is preferred to topology r and $V^c_{rn} > V^c_{sn}$, implying topology r is at least as good as topology s; or (ii) topology j is preferred to topology s and $V^c_{sn} > V^c_{rn}$, implying topology s is at least as good as topology r. Describing these events through (2), we get that the probability that topology j is estimated is

$$\begin{aligned} &P(V^c_{jn} > 0, V^c_{jn} - V^c_{rn} > 0, V^c_{rn} - V^c_{sn} > 0) \\ &+P(V^c_{jn} > 0, V^c_{jn} \quad - V^c_{sn} > 0, V^c_{sn} - V^c_{rn} > 0). \end{aligned} \tag{3}$$

In Susko (2011), it is shown that $[V^c_{1n}, V^c_{2n}, V^c_{3n}]$ has an approximate multivariate normal distribution. Expressions for the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ^c are given in that paper and will not be repeated here. Since $[V^c_{jn}, V^c_{jn} - V^c_{sn}, V^c_{rn} - V^c_{sn}]$ is a linear transformation of these normal vectors, it too is normal with a mean and covariance matrix, $\boldsymbol{d}^{(1)}$ and $A^{(1)}$, that can be determined in a relatively straightforward manner from Σ^c and $\boldsymbol{\mu}$. Similarly, $[V^c_{jn}, V^c_{jn} - V^c_{sn}, V^c_{sn} - V^c_{rn}]$ has a normal distribution. Letting $\boldsymbol{d}^{(2)}$ and $A^{(2)}$ denote its mean vector and covariance matrix, we can reexpress (3) as (1).

The exact form of $\boldsymbol{d}^{(1)}$, $\boldsymbol{d}^{(2)}$, $A^{(1)}$, and $A^{(2)}$ are given in equations (10–11) of Susko (2011) and, due to their complexity, are not be repeated here. We can, however, use their properties to explain the constraint that the small middle edge in Topology 1 has $t^{(n)}_{0m} \approx a/\sqrt{n}$ for some constant a. In the approximations, it is only the means $\boldsymbol{d}^{(1)}$ and $\boldsymbol{d}^{(2)}$ that depend on a. When calculation is for the probability of correctly estimating Topology 1, the vectors $\boldsymbol{d}^{(1)}$ and $\boldsymbol{d}^{(2)}$ satisfy that

$$\begin{aligned} \boldsymbol{d}^{(1)} &= a[x_1, y, z], \\ \boldsymbol{d}^{(2)} &= a[x_2, -y, z], \end{aligned}$$

for some constants x_1, x_2, y, and z, independent of a, whose exact form need not be known but which satisfies that x_1, x_2, and z are positive. If $y > 0$, then for a large $\boldsymbol{d}^{(1)}$ will have all large and positive entries. Since a normal random vector (X, Y, Z) is highly likely to have positive entries if its component means are large, $P_0(\boldsymbol{d}^{(1)}, A^{(1)})$ in (1) will be close to 1. Similarly if $y < 0$, one can show that $P_0(\boldsymbol{d}^{(2)}, A^{(2)}) \approx 1$. Thus, the probability of correctly reconstructing the tree is close to 1 for a large. This is true regardless of what design strategy is being considered, making it an uninteresting case. In the other direction, as a gets small, the means get close to 0, and the limiting probability is the same as the probability that Topology 1 is estimated when the generating tree is completely unresolved at the split of interest. Design strategies to increase the probability of estimating Topology 1 are no longer of interest because it is no longer the sole correct tree. One thing to note about the above discussion is that although the constraint $t^{(n)}_{0m} \approx a/\sqrt{n}$ is necessary for the theory, it does not need to be considered from a practical perspective. Simply plugging in values of a and n in the approximation will yield approximations where, if $a/\sqrt{n}$ is too large or small, probabilities will be near 1 or small.

The quantities $\boldsymbol{d}^{(1)}$, $\boldsymbol{d}^{(2)}$, $A^{(1)}$, and $A^{(2)}$ are calculated largely from the expected information matrices for the three trees. Here, the expected information matrix for the jth tree is defined as the expected value of the second derivative matrix of the negative log likelihood fixing the tree for calculation. Derivatives are taken with respect to edge-lengths but with the middle edge-length set to 0. In this sense, the approach is similar to that of Geuten et al. (2007) who consider transformations of the expected information as a design criteria. In a similar vein, the probabilities calculated may be viewed as transformations of the expected information matrices, albeit more directly interpretable values.

As in Geuten et al. (2007), we aim to improve the accuracy of reconstruction of an evolutionary tree connecting predetermined taxa by augmenting it with additional taxa. We do not address whether the evolutionary tree, including the additional taxa, is correctly estimated. While in some cases, this latter question is of interest, the large sequence-length results of Susko (2011) were derived under the assumption of a single poorly resolved edge and thus are not applicable. Moreover, as long as the addition of branches creates a tree with splits that are relatively well resolved (relative to the split of interest), the large sequence-length results still apply; the additional branch will, with large sequence lengths, be placed on the correct side of the split of interest. Difficulties arise when the additional edges are chosen close to an internal node. Here, however, the probabilities might still be viewed as the probabilities of correct reconstruction when the additional edges are constrained to be at their correct location.

Results

While the methods can be applied to more general models, in all the results below, the Jukes–Cantor (Jukes and Cantor 1969) substitution model is used as a generating model. Unless otherwise stated, the number of sites in any simulated alignment is 1000.

Single Taxon Addition versus Increased Sequence Length

The first case we consider is when a single taxon is added to a four-taxon alignment. The setting is illustrated in Figure 3a. The taxon to be added is 5 and connects to the edge leading to 1. Assuming a comparable rate of evolution along the sister lineages, a natural choice for the edge-length leading to 5 is to make it the same as the length of the edge leading to 1. Figure 3b–d give the probability of correct estimation as a function of x for several choices of short edge-length s and long edge-length b. Regardless of the edge-lengths, it is always better to choose the additional branch to be as close as possible to the internal node.

For each choice of x, very similar probabilities were obtained when the additional edge was added to the short (length s) branch. For a given set of s, b, and x, slightly higher probabilities of correct reconstruction were obtained when the edge leading to 5 had length 0; results not shown.

The horizontal lines in Figure 3 give the probabilities that the correct tree is reconstructed with $n = 1250$ for the four-taxon tree without taxon 5. The number of sites is 1250 which was chosen to so that the total number of nucleotides in the alignment was the same as for the five-taxon designs. Adding a taxon is better than

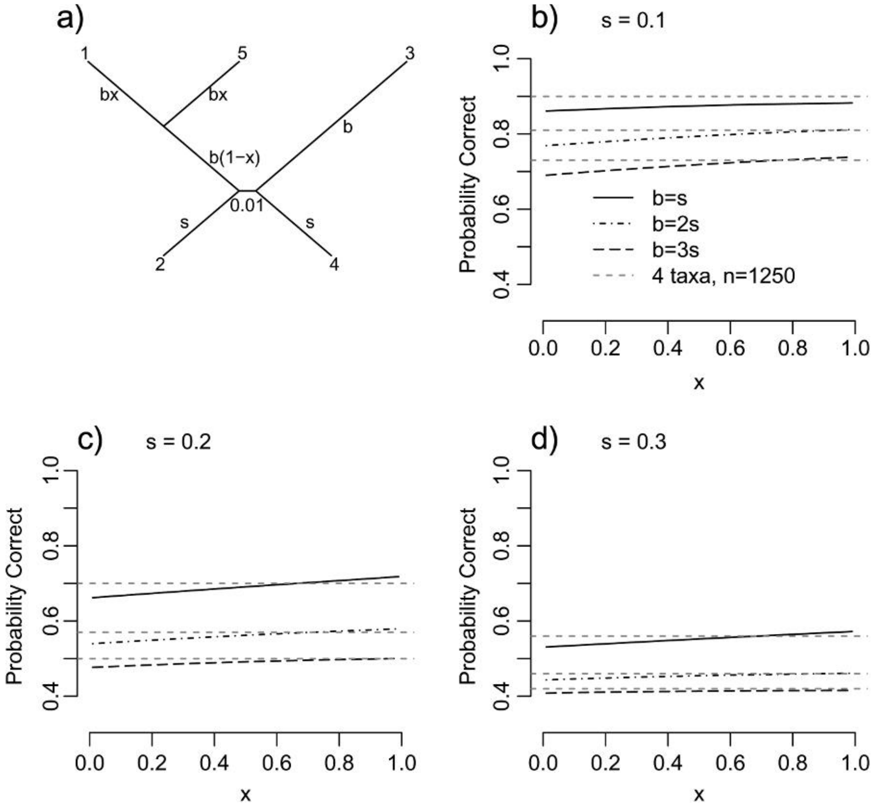

FIGURE 3. The split 12|34 in (a) is of interest. Taxon 5 is added to the alignment, and (b–d) give the probability of correct estimation of this split as a function of x. The horizontal lines give the probability of correct reconstruction with the original four taxa when sequence length is increased from 1000 to 1250.

TABLE 1. Probabilities of correct reconstruction when two taxa are added

e_5	1			2			3			4				
x	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	y	e_6
A	0.56	0.57	0.58	0.56	0.57	0.58	0.57	0.58	0.59	0.56	0.58	0.59	0.25	1
		0.57	0.58	0.57	0.58	0.59		0.60	0.61	0.58	0.60	0.62	0.50	
			0.58	0.58	0.58	0.59			0.63	0.59	0.61	0.63	0.75	
				0.55	0.56	0.58				0.56	0.58	0.59	0.25	2
					0.57	0.58					0.60	0.62	0.50	
						0.58						0.64	0.75	
B	0.56	0.58	0.60	0.57	0.59	0.60	0.59	0.62	0.65	0.58	0.61	0.63	0.25	1
		0.58	0.60	0.59	0.60	0.61		0.66	0.71	0.61	0.64	0.68	0.50	
			0.60	0.61	0.61	0.62			0.77	0.64	0.68	0.73	0.75	
				0.56	0.57	0.59				0.58	0.60	0.62	0.25	2
					0.57	0.59					0.63	0.66	0.50	
						0.59						0.71	0.75	

Notes: The tree with taxa 1–4 in Figure 4 is of interest. Taxon 5 is added to the edge e_5, and Taxon 6 is added to the edge e_6. In Table 1A, the lengths of new edges are x and y times the length of the edge they have connected to. In Table 1B, these lengths are 0. Because of the symmetry of the problem, not all possible edges for connection have results listed. For instance, adding taxa to 1 and 4 gives the same table as for 2 and 3. Here, $s=0.2$, $b=0.4$, and $a=0.01$

increasing sequence length, as long as that taxon can be chosen so that the additional branch connects near an internal node. If instead, for instance, it connects at the half length of the terminal edge, the probabilities are comparable.

Adding two taxa.—Table 1 gives the probabilities of correct reconstruction when two taxa are sampled. Consistent with results when a single taxon is added, for each pair of branches that the taxa can be added to, one can see that probabilities of correct reconstruction are always larger when the taxa are added closer to an internal edge.

In Table 1A, it is marginally better to connect the two edges to the short edges 2 and 4. Since estimates of longer edges have larger variance, this may seem surprising, but because the connecting edges are of length x and y times the length of the edge they connect to, they will also be shortest when connected to 2 and 4. In Table 1B, the connecting edges are of length 0, and the optimal edges to connect them to are the long edges 1 and 3.

Adding three or four taxa.—The probabilities of correct reconstruction when adding three taxa are considered in Table 2. The first two taxa are added at the optimal locations from Table 1: attached to the short edges 2 and 4 as deeply as possible. Rather than adding additional edges to the short edges, it is best to add the third taxon to one of the long edges, again as deeply as possible.

Table 3 gives the results adding a fourth taxon. The optimal location for addition is the remaining long edge 3 and, again, as deeply as possible. The associated optimal probability is 0.72. With 4 additional taxa and 1000 sites, the total number of additional sites is 4000. Thus, the comparable four-taxon design that does not add taxa but lengthens the 4 sequences is one with 2000 sites. The probability of correct reconstruction with this design is only 0.63.

Optimal Rates

Townsend (2007) defines an optimal rate of evolution for a four-taxon molecular clock tree as the rate which maximizes the probability that a character experiences at least one change along a middle edge but remains unchanged along its tips. For a rooted four-taxon tree with two taxa on each side of the root, he shows that the optimal rate is

$$\lambda = \frac{1}{t_0}\log\frac{4T-t_0}{4T-2t_0}, \tag{4}$$

where T is the total elapsed time from root to tips and t_0 is the total length of the root branch. Figure 5 gives the probability of correct reconstruction as a function of the rate λ for a four-taxon tree with equal terminal edge-lengths, $t_0=0.1$ and $T=1$, 2, and 3. Indicated as well with vertical lines are Townsend's optimal rates from (4). These are close to, but consistently overestimate, optimal rates using maximized probability of reconstruction as the optimality criterion.

TABLE 2. Probabilities of correct reconstruction when a third taxon is added

Edge	x	Probability	Edge	x	Probability
1	0.25	0.65	3	0.25	0.65
1	0.50	0.67	3	0.50	0.67
1	0.75	0.68	3	0.75	0.68
2	0.25	0.64	4	0.25	0.64
2	0.50	0.65	4	0.50	0.65
2	0.75	0.66	4	0.75	0.66

Notes: The tree with taxa 1–4 in Figure 4 is of interest. The first two taxa are added at the optimal location in Table 2: They are attached to edges 2 and 4 as deeply as possible: $x=y=0.75$. The probabilities are listed as a function of the edge that the third taxon is added to and x; As with other design settings when an edge is added to an edge of length l, it is of length lx and is connected distance lx from the internal node.

TABLE 3. Probabilities of correct reconstruction when a fourth taxon is added

Edge	x	Probability	Edge	x	Probability
1	0.25	0.68	3	0.25	0.70
1	0.50	0.69	3	0.50	0.71
1	0.75	0.69	3	0.75	0.72
2	0.25	0.68	4	0.25	0.68
2	0.50	0.68	4	0.50	0.69
2	0.75	0.69	4	0.75	0.70

Notes: The tree with taxa 1–4 in Figure 4 is of interest. The first three taxa are added in the optimal locations from Tables 2 and 3: The first two are added to short edges 2 and 4 with $x = y = 0.75$ and the the third is added to 1 with $x = 0.75$. The probabilities are listed as a function of the edge that the fourth taxa is added to and x; as with other design settings when an edge is added to an edge of length l, it is of length lx and is connected distance lx from the internal node.

Seed Plant Phylogeny

Figure 1 gives the seed plant phylogeny used as an illustrative example in Geuten et al. (2007). The edge-lengths are the same as those of Figure 6 in Geuten et al. (2007), which are ML estimates based on a 5544-site alignment; the alignment and tree are available at TreeBASE (http://www.treebase.org) under the study ID S1811. Using these edge-lengths, and a GTR model, the optimal location for taxon addition was found to be at the node connecting to the terminal edge leading to *Arabidopsis*, unanimously by A-, D-, and E-optimality criteria of Geuten et al. (2007). Our design strategy in a setting such as this becomes more complex because it requires a split of interest, the choice of which is unclear here, and hypothesizes a smaller-than-estimated edge-length for this split.

For our probabilistic criterion, a reasonable design investigation would consider a number of putative splits of interest that might potentially be wrong. Once the split of interest is chosen, since it is considered uncertain, one should consider the three possible resolved relationships of the four edges closest to the split. Fixing these choices and considering a number of hypothetical middle edge-lengths, the question becomes: where is it best to add an edge to the tree? It is possible that multiple answers will arise in this case and that additional criteria will be required to decide which choice to pursue. We illustrate this approach by considering two splits of interest. These are labeled as X and Y in Figure 1 that correspond to relatively small edge-lengths in the Geuten et al. (2007) tree.

For the split X in Figure 1, there are three possible resolved relationships, which are given at the top of Figure 6; similarly, for split Y in Figure 1, the three resolved relationships are given in Figure 7. For each of these resolved relationships, we considered several hypothetical small middle edge-lengths. We fixed the sequence length at 5544, as for the original data, and used a Jukes–Cantor model in the results reported below; similar results were obtained with a GTR model. A different optimal location for addition may be found for each of the fixed choices of split of interest, resolved relationship and middle edge-length. So, we calculated probabilities of correct resolution after adding a new edge to edges of the tree by varying all these conditions.

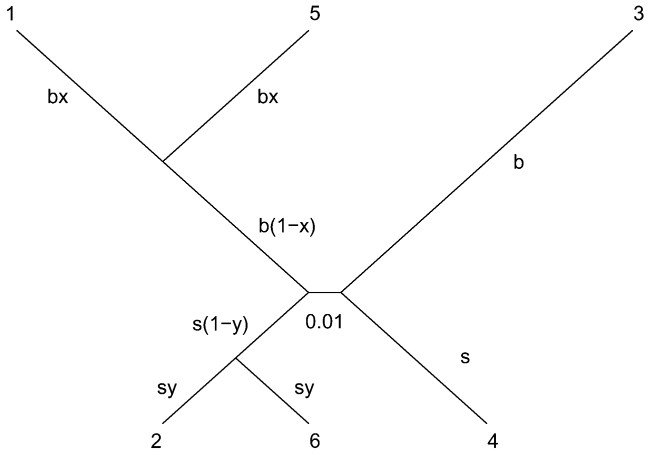

FIGURE 4. The design for two-taxon addition. The split 12|34 is of interest and the two of the taxa 5 and 6 are to be added, possibly to the same edge. In the illustrative example, 5 is added to the edge leading to 1 and 6 is added to edge leading to 4.

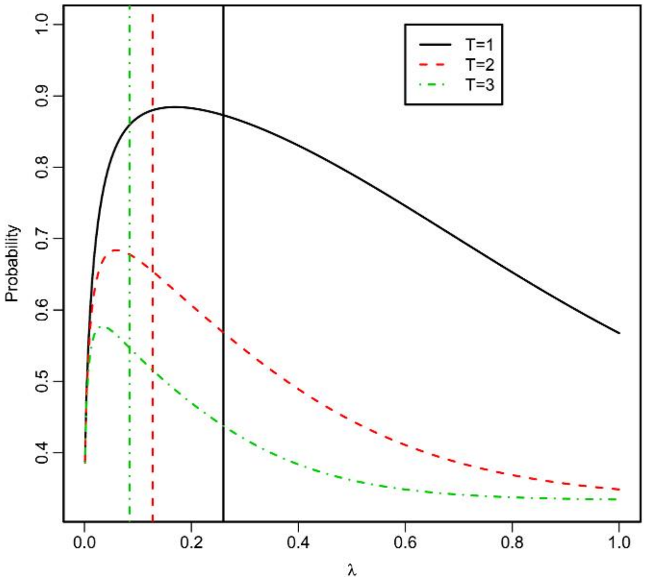

FIGURE 5. The probability of correct estimation as a function of an overall rate multiplier λ when the rooted generating tree has equal terminal edge-lengths, total distance T from root to tip and middle edge-length 0.01. Vertical lines indicate the optimal λ values (4) defined in Townsend (2007).

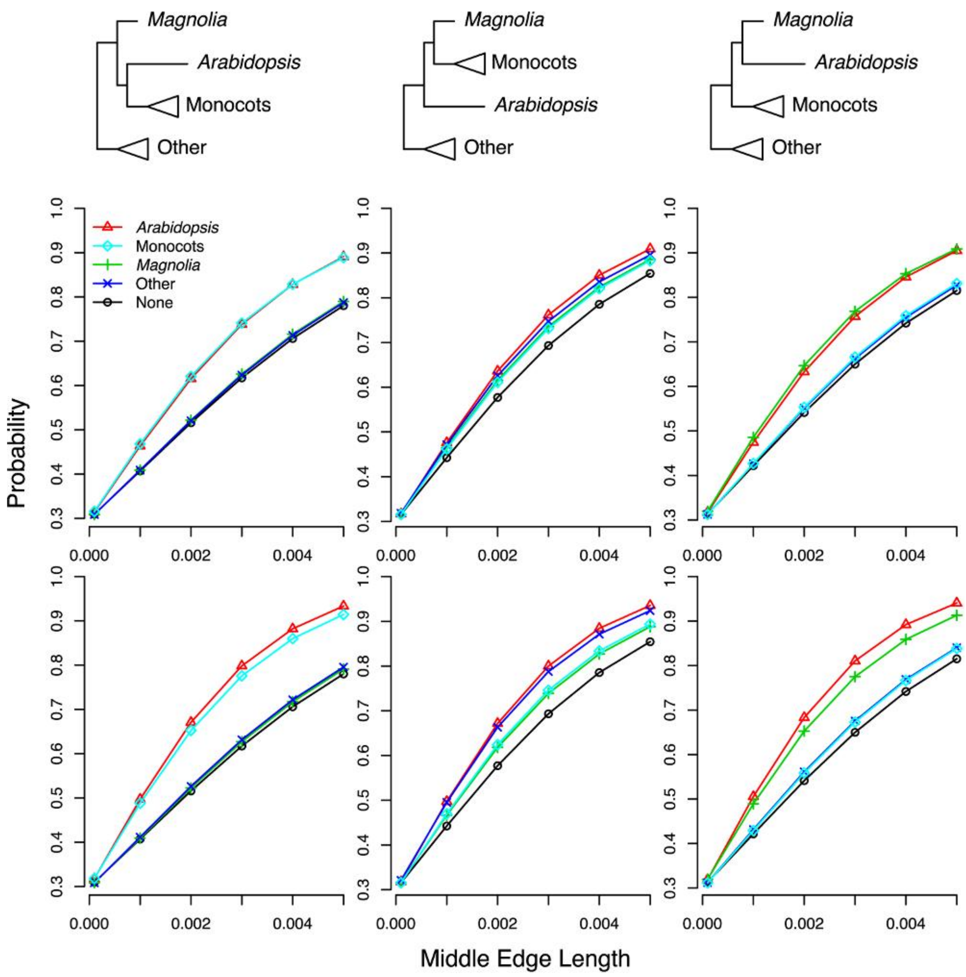

FIGURE 6. The probability of correct reconstruction is plotted against middle edge-length when taxon addition is for the split X in Figure 1. Each set of points plotted is for addition to an edge leading to a particular subtree in Figure 1; *None* indicates that no addition was done. The subtree at the top of a column indicates the assumed resolved relationship in calculating probabilities of reconstruction after taxon addition. The first row of plots corresponds to the variable edge length scheme and the lower row to the fixed edge length scheme.

We considered additions to all edges of the tree. For each edge, we also considered additions along the edge at several locations. As a fraction of the total distance, these were one-fourth, one-half, and three-fourth of the total edge-length. Each choice of taxon addition also requires a length for the terminal edge to be added. To allow for the possibility that sampling closer to the split of interest comes with the cost that the added edge will be longer, we considered both a variable edge-length addition scheme and a fixed edge-length addition scheme. The fixed edge-length scheme simply involves adding an edge of length 0.05. Considering the tree as rooted at the split of interest, under the variable edge-length scheme, fixing the location of the edge to be added, the length was taken as the average length from root to tip for the subtree rooted at that location.

In the case of a middle edge-length 1.0e^{-4}, probabilities and differences in probabilities of correct reconstruction were small; in fact, for a number of the edges, they were no better than without taxon addition. This case serves mainly to indicate that if, in fact, the middle edge-length is sufficiently small, additional taxon selection, for a given sequence length, will not aid reconstruction.

For middle edge-length larger than 1.0^{-4}, for each edge, we always found that probability of correct resolution was optimal after adding at the location closest to the split of interest. In addition, we found that the best edge to add the new taxon to was one of the four closest to the split of interest. Since for all choices of middle edge-length larger than 1.0e^{-4}, the best location for addition was one of the four edges closest to the split of interest, in what follows we report results only for these four edges and for the location along these four edges closest to the split of interest.

The results for the split labeled X in Figure 1 are given in Figure 6, and for the split Y, results are given in Figure 7. Not surprisingly, additions to neighboring taxa for the same middle edge-length give similar probabilities of correct resolution. This is because optimal edge-length addition is closest to the split of interest; differences arise because this is still one-fourth of the way

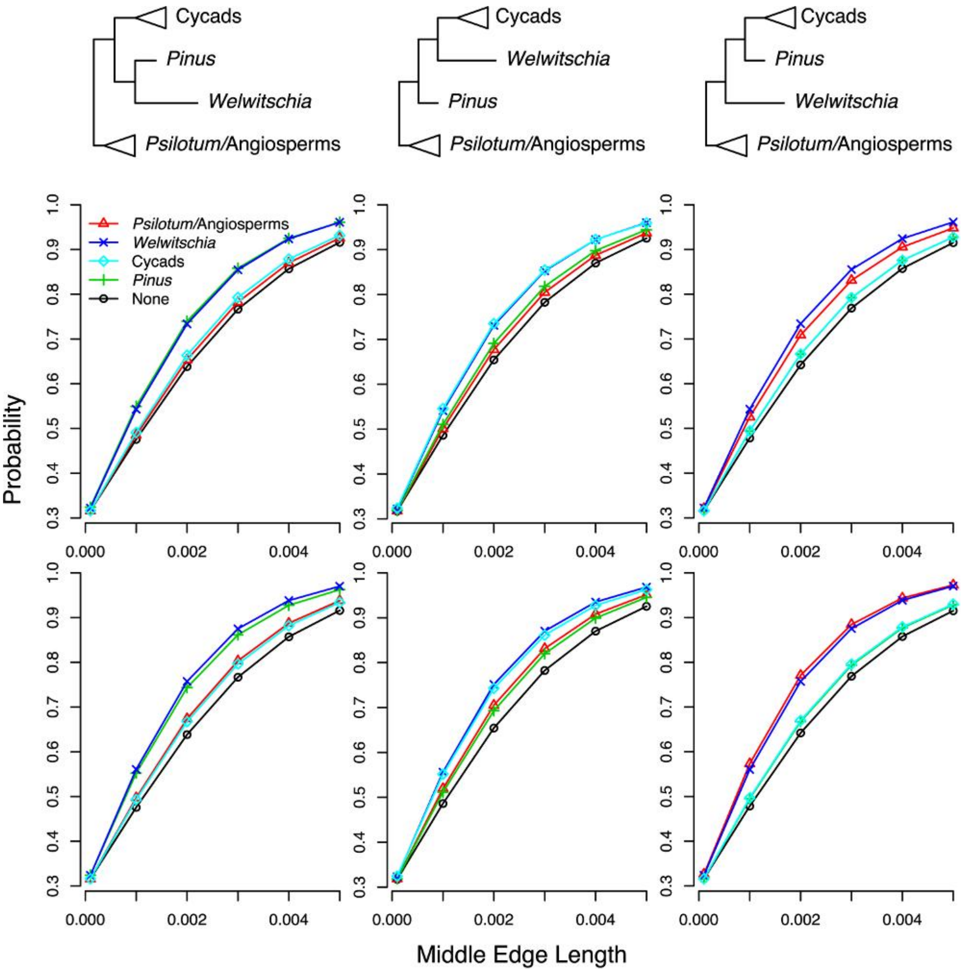

FIGURE 7. The probability of correct reconstruction is plotted against middle edge-length when taxon addition is for the split Y in Figure 1. Each set of points plotted is for addition to an edge leading to a particular subtree in Figure 1; *None* indicates that no addition was done. The subtree at the top of a column indicates the assumed resolved relationship in calculating probabilities of reconstruction after taxon addition. The first row of plots correspond to the variable edge length scheme and the lower row to the fixed edge length scheme.

along the edge. Considering the split labeled X first, it is almost always best to add the new edge to the long edge leading to *Arabidopsis*. In the cases where the neighbor of *Arabidopsis* is not the *Other* group, it appears that no substantial improvement is obtained by adding a new edge to the edge leading to the *Other* group or its neighbor. In contrast, when the *Other* group is the neighbor of *Arabidopsis*, the probability of correct resolution is improved by addition to any of the edges.

Similar results were obtained for split Y. The best location for addition was along the longest edge, leading to *Welwitschia* or its neighbor. While the improvement was obtained through taxon addition regardless of which of the four edges it was to, in the cases where the *Psilotum/Angiosperms* group was not the neighbor of *Welwitschia*, improvements were relatively small when adding to one of the branches leading to the *Psilotum/Angiosperms* group or its neighbor.

For the two splits interest, we get different choices for where it is best to add taxa. One could choose the placement, which maximizes the gain in performance. For most of the resolved splits, and a fixed middle edge-length, the largest difference in probabilities of correct resolution for the best addition and no taxon addition (the difference between the largest and smallest y-axis value) is greatest for the split X suggesting that, overall, adding to the node closest to *Arabidopsis* is the optimal choice.

Table 4 gives probabilities of correct reconstruction of the splits X and Y when sequence length is doubled but taxa are deleted. These are of interest in cases where it may be possible to sample additional sites for some taxa. In such cases, it is of interest to know which taxa would be the best or worst to exclude from the additional sampling. As might be anticipated, it is better to delete taxa positioned farther away from the split of interest. What may be surprising are the rates of change in the probabilities. Considering the choices of best deletion, probabilities of reconstruction are almost constant when up to three taxa are deleted. They decrease precipitously if 6 or 7 taxa are deleted but are always better than when no taxa are deleted and the original number of sites is used. For the worst deletions and split Y, the probabilities decrease more rapidly with the number of taxa deleted. The worst deletion of 4 taxa is worse than the best deletion of 7 taxa and gives only a slightly better reconstruction probability than with half as many sites and no taxa deleted. Similarly for the split X, probabilities for the worst deletions are small and with 6–7 taxa deleted, worse than with half the sites and no taxa deleted.

Software Implementation

Software for the tools presented here is available at http://www.mathstat.dal.ca/~tsusko.

Programs are provided that add and delete branches from a tree. The main program, `pr4design`, uses a control file similar to PAML (Yang 1997, 2007) and returns probabilities calculated under a number of widely used nucleotide and amino acid substitution models.

Discussion

By utilizing the large sequence-length results of Susko (2011), we have developed a fast way of evaluating the merits of design strategies that add taxa or sites. An alternative approach is to estimate the probability of correct reconstruction via the corresponding proportion in repeated simulated sequence alignments. Because the theory implies that only the three competing topologies in Figure 2 need to be considered, full tree searching is not necessary in simulations. To investigate the computational savings due to using the theoretical calculations, we obtained the central processing unit (CPU) times required for the several of the results by comparison with those required to obtain estimated trees in 1000 simulations using TREE-PUZZLE 5.2 (Schmidt et al. 2002). For the four-taxa with n=1000 and n=1250, the average CPU time was 0.002 s for the theoretical calculations and 33 s for simulation-based approximations. For the five-taxa examples at the two extremes, with $x = 0.01$ and $x = 0.99$, the average CPU time was 0.005 s for theoretical calculations and 3 m 29 s for simulation-based approximations. Considering the X split for the seed plant data with 11 taxa, and a single edge-length setting from Figure 6, the theoretical calculations required 2 m 6 s and simulation-based approximations 15 m 44.312 s. The reason that theoretical calculations require more time for 11 taxa is that calculating the expected information matrices (expected negative second derivative matrix of the log likelihood) requires summing over all possible site patterns. These expectations can be approximated reasonably well by the negative second derivative matrix of the log likelihood for long sequences generated from the tree of interest. Using such an approximation with 100,000 sites in the 11 taxa seed data example gave a CPU time of 7.5 s.

TABLE 4. Probabilities of correct reconstruction for the two splits X and Y in the seed plant tree when the number of sites is doubled but the taxa are deleted

Split = X	Split = Y
Taxa Deleted	Taxa Deleted
0.607 None	0.761 None
0.607 Zar	0.761 Zea
0.606 Wel Zar	0.760 Ara Zea
0.606 Wel Zar Pin	0.759 Ara Ory Zea
0.605 Wel Zar Pin Psi	0.754 Nym Ara Ory Zea
0.599 Wel Zar Cyc Pin Psi	0.739 Zar Nym Ara Ory Zea
0.574 Wel Zar Cyc Pin Psi Nym	0.720 Zar Nym Mag Ara Ory Zea
0.533 Zea Wel Zar Cyc Pin Psi Nym	0.671 Zar Psi Nym Mag Ara Ory Zea
1: 0.547 Ory	0.736 Cyc
2: 0.542 Ory Nym	0.717 Cyc Psi
3: 0.533 Ory Amb Nym	0.686 Cyc Psi Amb
4: 0.530 Ory Psi Amb Nym	0.658 Psi Amb Nym Mag
5: 0.526 Ory Pin Psi Amb Nym	0.611 Cyc Psi Amb Nym Mag
6: 0.516 Wel Zar Cyc Pin Amb Nym	0.564 Psi Amb Nym Mag Ory Zea
7: 0.493 Ory Wel Zar Cyc Pin Amb Nym	0.516 Cyc Psi Amb Nym Mag Ory Zea

Notes: Listed in the top portion of the table are the best choices for deletion and in the bottom panel, the worst choices. The middle edge-length is set to 0.002 in both cases. For the original number of sites, and with no taxa deleted, the probabilities of correct reconstruction are 0.516 and 0.638.

To check how the theoretical approximations agreed with simulation-based approximations, we calculated simulation-based approximations for some of the settings in Figure 3. The results are given in Figure 8. One can see that for most settings, theoretical and simulation-based approximations coincide up to the level of uncertainty inherent in the simulation-based approximations. For some of the larger probabil-

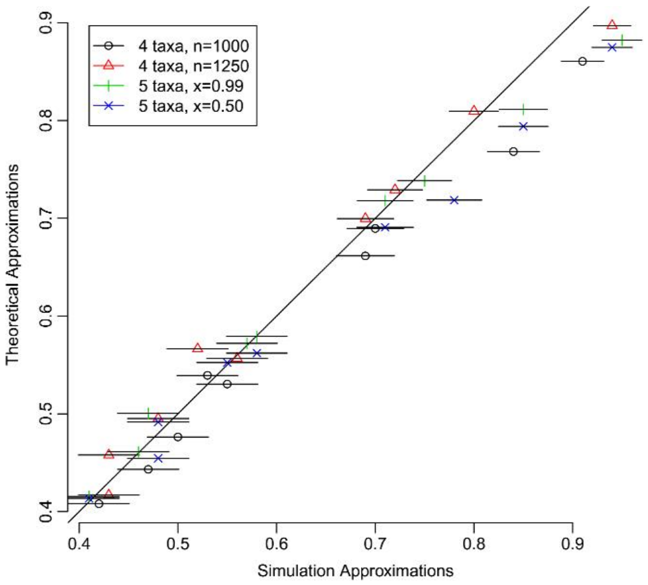

FIGURE 8. The theoretical approximations for some of the Figure 3 settings plotted against simulation-based approximations. Simulation-based approximations were determined as the proportion of correct estimations over 1000 simulations. Error bars indicate 95% confidence intervals for the true probabilities of correct reconstruction.

ities of reconstruction, the theoretical probabilities are smaller than the simulation approximation. It may be the case that for such settings, larger sequence lengths are required for the theoretical approximations to be accurate. In any case, the relative rankings of the probabilities whether via simulation or theory remain roughly the same.

In general, we found that it is usually better to add taxa that connect near an internal node for the split of interest, if they can be so chosen, and that it is better to sample additional taxa than to add a comparable number of sites to the data set of interest. The findings here are largely consistent with Geuten et al. (2007) who similarly found that it is almost always better to add taxa that connect near an internal node for the split of interest. (An exception arises with one of their design criteria, E-optimality, which sometimes prefers connection farther along long edges, but the other two criteria, A and D optimality, favor deep placement.) Geuten et al. (2007) found it best to place the additional taxa along a long edge, which appears to differ somewhat from the findings reported here. In our study, however, we assumed additional lineages would have the same evolutionary rate as sister lineages. Thus, the edge-lengths of taxa connecting to long edges are longer than those connected to short edges. If this is the case, it is sometimes better to sample taxa connecting to a short edge. In contrast, Geuten et al. (2007) usually added fixed-length edges. For comparison, we sometimes added zero-length edges and, unsurprisingly, found that it was better to attach to long edges.

In the larger seed plant phylogeny example, the two different splits of interest gave two different optimal edges for addition that were almost always near the split of interest. In this case, other information may be used to decide which of these two choices to follow up on, such as which region of the tree is of greater interest and what taxa are available to add. Choosing the edge that gave the biggest improvement in probability of correct resolution suggested that addition to *Arabidopsis* is best, which is consistent with the findings in Geuten et al. (2007). In any case, similar results were obtained for the each of these, consistent with results for smaller numbers of taxa. For instance, it was best to add to the longest edge near the split of interest. Interestingly, when the *Other* group was not a neighbor of the longest edge, the probability of correct resolution did not increase appreciably if an additional taxon was added to it, particularly under the variable edge length scheme, where the new edge is likely to be longer. This is likely a consequence of the *Other* group containing the largest number of taxa.

The criteria of Geuten et al. (2007) were transformations of the expected information matrix of the edge-length parameters for the true topology. As shown in Susko (2011), the probability of reconstruction is heavily dependent upon the expected information matrix but also requires knowledge of the correlations of key quantities across the three competing topologies of interest.

It is important to keep in mind that for all the examples considered here, the phylogenetic position of the taxon to be added is known in advance. Even if this is the case, it is possible that with real data, the *estimated* position of the additional taxon may be uncertain, leading to increased apparent uncertainty in the topology near the split of interest (i.e., lowered bootstrap support values). Although this may seem a counterproductive strategy to increase resolution, the problem might be remedied by constraining the topology after taxon addition so that the new taxon is forced to remain in its known position during tree searching.

Finally, we also investigated the case where additional sampling of sequence data is performed but some of the original taxa are not included. This situation would occur when the investigator is interested in gathering more data but cannot for logistical or financial reasons sample all taxa in the original matrix. In this case, we showed that unsurprisingly the optimal taxa to "leave out" in additional sampling effort are often distant to the split of interest. Unexpectedly, quite a large number of such taxa can be left out without drastically affecting the probability of correct reconstruction by doubling the number of sites to the original matrix. Those taxa whose deletion most negatively impact on the probabilities of reconstruction are much closer to the split of interest, as expected. Interestingly, the latter taxa are not the same taxa next to which sampling of additional edges is expected to most improve probabilities of reconstruction. That is, while our analyses suggest that addition of taxa next to *Arabidopsis* in the seed plant data set is the optimal choice to maximize probability of correct reconstruction of branch *X*, we found that the worst taxon to delete when sampling new sites is actually *Oryza* (*Arabidopsis* was not even among the worst seven taxa to delete). These methods will be very helpful aiding the selection of taxa for additional sequence sampling. Furthermore, as data matrices grow larger and larger, they can be used by researchers to choose subsets of taxa from their data matrices for computationally intensive analyses that would otherwise be impossible with the full taxon set.

Supplementary Material

The alignment and Newick format treefile used in the Seed Plant Data example and considered as well in Geuten et al. (2007) are available at Tree-BASE (http://www.treebase.org) under the study ID S1811.

Funding

This research was supported by Discovery grants awarded to A.J.R. and E.S. by the Natural Sciences and Engineering Research Council of Canada.

Acknowledgments

A.J.R. acknowledges support from the Canada Research Chairs Program and the Tula Foundation. E.S. and A.J.R. are members of the Centre for Comparative Genomics and Evolutionary Bioinformatics.

References

Atkinson G., Donev A. 1992. Optimum experimental design. Oxford: Oxford University Press.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376

Genz A. 2004. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. Stat. Comput. 14:251–264.

Geuten K., Massingham T., Darius P., Smets E., Goldman N. 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56:609–622.

Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B Biol. Sci. 265: 1779–1786.

Graybeal A. 1998. Is it better to add taxa or characters to a difficulty phylogenetic problem? Syst. Biol. 47:9–17.

Jukes T.H., Cantor C.R. 1969. Evolution of Protein Molecules. In Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kiefer J. 1959. Optimal experimental design. J. R. Stat. Soc. Series B Stat. Methodol. 21:272–319.

Pollock D.D., Zwickl D.J., McGuire J.A., Hillis D.M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664671.

Schmidt H.A., Strimmer K., Vingron M., von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.

Susko E. 2011. Large sample approximations of probabilities of correct topological estimation and biases of maximum likelihood estimation. Stat. Appl. Genet. Mol. Biol. 10(1):Article 10.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Yang Z. 1997. PAML: a program for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Yang Z. 2007. PAML 4: a program for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.