Syst. Biol. 56(4):609–622, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701499563

Experimental Design Criteria in Phylogenetics: Where to Add Taxa

KOEN GEUTEN,[1,5] TIM MASSINGHAM,[2] PAUL DARIUS,[3] ERIK SMETS,[1,4] AND NICK GOLDMAN[2]

[1]*Laboratory of Plant Systematics, and Department of Biosystems,*[3] *K. U. Leuven, Belgium*
[2]*EMBL-European Bioinformatics Institute, Hinxton, United Kingdom*
[4]*Nationaal Herbarium Nederland, Universiteit Leiden Branch, Leiden, The Netherlands*
[5]*Present Address: Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, USA; E-mail: koen.geuten@yale.edu*

Abstract.—Accurate phylogenetic inference is a topic of intensive research and debate and has been studied in response to many different factors: for example, differences in the method of reconstruction, the shape of the underlying tree, the substitution model, and varying quantities and types of data. Investigating whether the conditions used might lead to inaccurate inference has been attempted through elaborate data exploration but less attention has been given to creating a unified methodology to enable experimental designs in phylogenetic analysis to be improved and so avoid suboptimal conditions. Experimental design has been part of the field of statistics since the seminal work of Fisher in the early 20th century and a large body of literature exists on how to design optimum experiments. Here we investigate the use of the Fisher information matrix to decide between candidate positions for adding a taxon to a fixed topology, and introduce a parameter transformation that permits comparison of these different designs. This extension to Goldman (1998. Proc. R. Soc. Lond. B. 265: 1779–1786) thus allows investigation of "where to add taxa" in a phylogeny. We compare three different measures of the total information for selecting the position to add a taxon to a tree. Our methods are illustrated by investigating the behavior of the three criteria when adding a branch to model trees, and by applying the different criteria to two biological examples: a simplified taxon-sampling problem in the balsaminoid Ericales and the phylogeny of seed plants. [A-optimality; D-optimality; E-optimality; experimental design; Fisher information; phylogenetics; taxon sampling.]

Nylander (2001) formulates the problem of taxon sampling as: "Given a sample of taxa and a method of analysis, certain properties of the data could render the reconstruction of the true tree difficult or even impossible. Furthermore, a different sample of taxa could have facilitated the reconstruction, or conversely, even made it more difficult." Examples from the literature illustrate that the addition or deletion of even a single taxon can change the inferred phylogenetic relationships and there is a general consensus on the importance of, and sensitivity to, taxon sampling (e.g., Poe, 2003; Soltis et al., 2004; Martin et al., 2005).

The practice of using genome-scale data to infer the phylogenies of a relatively small number of taxa, originally proposed as a way of ending incongruence (Rokas et al., 2003; Martin et al., 2005), has only increased the importance of judicious taxon sampling (Soltis et al., 2004; Hedtke et al., 2006). However, the main concern of these large-scale studies was gene sampling and the choice of an appropriate evolutionary model, rather than taxon sampling. Several examples illustrate that caution is appropriate when interpreting phylogenies derived from many characters but few taxa (e.g., Naylor and Brown, 1998; Göremykin et al., 2003; cf. Soltis et al., 2004; Rokas et al., 2003; cf. Phillips et al., 2004; Philippe et al., 2005; Hedtke et al., 2006). In the worst imaginable case, the combination of many characters and few taxa could lead to maximum support for incorrect evolutionary relationships, thus only seemingly removing the recurring problem of incongruence between datasets and low support for relationships. Budget constraints also limit taxon sampling in genome-scale phylogeny projects, further highlighting the importance of making a judicious choice of taxa. The goal of experimental design in phylogenetics is to predict where taxa can be added to maximize the improvement in accuracy. Our aim is to develop a methodology to facilitate this goal.

Recent advice on how to design optimal experiments in phylogenetics mainly comes from simulation studies investigating the influence that augmenting a design has on phylogenetic accuracy—adding taxa versus adding characters (Graybeal, 1998; Rosenberg and Kumar, 2001, 2003; Pollock et al., 2002; Zwickl and Hillis, 2002; Hillis et al., 2003; Hedtke et al., 2006). The consensus is that increased taxon addition improves phylogenetic inference, although a debate remains surrounding the importance of this improvement. In molecular phylogenetics, experimentalists seeking advice on their experimental setup are mainly confronted with "what can go wrong when using a certain method" and "what are the most important factors influencing accuracy," but a general theory of how to design their studies is lacking.

Nevertheless, experimental design is well treated in the statistical literature, dating back to the work of R. A. Fisher (1926, 1935) and with many extensions after that. The information matrix has a central role in modern experimental design. The observed information matrix describes the precision with which parameters have been measured in a given experiment and is closely related to classical confidence intervals (the higher the information about a parameter, the smaller its confidence interval). Indeed, the observed information has been used by many phylogenetics packages to establish the standard error of model parameter and branch length estimates (e.g., PAML: Yang, 1997). In comparison to the actual precision achieved, the expected information (also known as the Fisher information) describes what precision we would expect to achieve before any data are collected.

Goldman (1998) showed how to calculate the Fisher information matrix in the context of molecular systematics and developed theory to address experimental design questions in phylogenetics directly. These original

methods are only applicable to certain questions, however, and Goldman illustrates their use in two examples. The first shows how to calculate the increase in information about a single node in a clock-like phylogeny, when either augmenting the phylogeny with an additional sequence or by extending the sequence lengths. The second addresses the question of how to choose which relative evolutionary rate would be most informative about the phylogeny, or different branches within the phylogeny, when considering a new gene to sequence. Goldman's (1998) theory, as originally described, is not applicable to an important category of design problems: choosing where in the topology to add a new taxon to optimally improve our understanding of the phylogeny. Scientists are regularly confronted with this practical problem when trying to improve the design of an experiment in phylogenetics, which may explain why the topic has drawn some attention. That Goldman's original theory is not applicable can easily be shown with the following thought experiment: consider augmenting a tree with a sequence very similar to one already chosen; this creates two very short branches whose lengths can be estimated with high precision, considerably increasing the total information about the tree because it includes information about these two new branches. The closer the new sequence is to the original sequence, the more information we have about these two branches, leading to the absurd conclusion that it is best to augment the tree with a sequence identical to one already chosen. This apparent paradox arises for two reasons: each taxon addition creates branches that are not comparable between augmentations, and it is incorrect to consider the variance of parameters that have been fixed by the experimental design under consideration. These issues do not arise in Goldman (1998: e.g., the first example avoids them by considering only clock-like trees and calculating information relating only to internal nodes present in the original tree) but are of considerable interest in practice.

Given the present lack of general guidelines for biologists to design better phylogenetic experiments despite the numerous case studies indicating potential problems with one approach or another, we extend the approach of Goldman (1998) to tackle the addition of arbitrary branches (taxa) to clock-free trees. The information matrices from augmenting a phylogeny each represent information about a different topology and not all their parameters are directly comparable; we introduce a transformation that allows different experiments to be compared by expressing the information in terms of that relating to the original tree. The information matrix describes how much we learn about each parameter and how much knowing something about one parameter tells us about another; in order to compare experiments we need some criterion that describes what trade-off between the different parameters we are prepared to accept. Several optimality criteria based on the Fisher information matrix have been proposed, denoted the "alphabetical optimality criteria" after their names (for example, A, D, and E; Kiefer, 1959; Atkinson and Donev, 1994), and this paper also investigates the behavior of these different criteria.

METHODS

Fisher Information

We briefly reintroduce the concept of Fisher information in the context of phylogenetic inference (for the original explanation, see Goldman, 1998). In maximum likelihood (ML) inference, a parameter or vector of parameters θ of interest is estimated by the value ($\hat{\theta}$) that maximizes the probability of the observed data D given parameter θ, $L(\theta \mid D)$ (Edwards, 1972). The precision with which θ is estimated can be measured by the curvature of the log-likelihood function at the maximum value, high curvature meaning that the likelihood surface is sharply peaked and so the parameter is estimated confidently. This precision, or *observed information*, is defined as minus the second derivative of the likelihood function evaluated at its maximum, which is where θ is equal to the maximum likelihood estimate $\hat{\theta}$:

$$J(\hat{\theta}) = -\left.\frac{d^2 \ln L(\theta|D)}{d\theta^2}\right|_{\theta=\hat{\theta}}$$

The standard error of parameter estimates is related to this observed information and can be used to produce confidence intervals (see Yang et al., 1995; Felsenstein, 2004, for applications in phylogenetics).

When multiple parameters are being estimated, the expression for the observed information is more complicated because θ is a vector and the information must take into account correlations between parameters: how errors in one affect estimates of the others. The observed information matrix is minus the matrix of second-order partial derivatives of the log-likelihood function evaluated at $\hat{\theta}$; thus, the (i, j) entry of the observed information matrix is:

$$J_{ij}(\hat{\theta}) = -\left.\frac{\partial^2 \ln L(\theta|D)}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}}$$

Instead of considering the observed data, one can also ask how much extra information would be expected to be obtained by observing an additional site. This is achieved by considering all possible observations: the "patterns" of characters observable at the leaves of the tree. The expected amount of information for this unknown site is the sum of the observed information for every possible pattern that may occur, weighted by the probability that it is observed. If the (i, j) entry of the information matrix for observing pattern b is $J_{ij}(\theta|b)$ and this pattern is observed with probability p_b, then the (i, j) entry of the expected, or Fisher, information is:

$$I_{ij}(\theta) = \mathbb{E}J_{ij}(\theta|b) = -\sum_b p_b \frac{\partial^2 \ln L(\theta|b)}{\partial\theta_i \partial\theta_j}$$

$$= \sum_b \frac{1}{p_b} \frac{\partial p_b}{\partial\theta_i} \frac{\partial p_b}{\partial\theta_j}, \quad (1)$$

because the following relations must hold:

$$\sum_b p_b = 1; \qquad \sum_b \frac{\partial p_b}{\partial \theta_i} = 0;$$

$$\sum_b \frac{\partial^2 p_b}{\partial \theta_i \partial \theta_j} = 0; \qquad \frac{\partial \ln p_b}{\partial \theta} = \frac{1}{p_b}\frac{\partial p_b}{\partial \theta}. \tag{2}$$

The (*total*) *expected information* for an experiment sampling n independent sites equals $nI(\theta)$.

Information and Covariance Matrices, and the Confidence Ellipsoid

Because altering an experiment to improve the precision with which one parameter may be estimated can adversely effect the precision of another parameter, an additional criterion describing the trade-off we are willing to make must be used in addition to the Fisher information matrix in order to rank different designs and so pick the most desirable. In this paper we compare three different criteria for designing experiments, known as the A-, D-, and E-criteria in the experimental design literature. Although each of the three criteria is correctly considered as a function of the Fisher information matrix, we motivate them in a more concrete form by considering a hypothetical confidence ellipsoid. Intuition gained from considering such ellipsoids is transferable because of the close relationship between the inverse of the Fisher information matrix and the covariance matrix of the ML estimator.

The Fisher information has an important connection to the Cramér-Rao lower bound, which places a limit on the performance of any estimator: the (i, i) element of the inverse of the total expected information matrix is a lower-bound on the variance of any unbiased estimator for parameter θ_i (e.g., Pawitan, 2001), and so designs based on the Fisher information represent the best precision that can be achieved in any experiment using an unbiased estimator. Biased estimators are subject to a similar bound but may appear to outperform unbiased ones in certain situations (Yang, 1996a). As the number of observations increases, the covariance of the ML parameter estimates tends to the inverse of the total expected information matrix (i.e., the ML estimate is asymptotically efficient); standard results (e.g., Pawitan, 2001) justify the representation of confidence intervals as ellipsoids, as illustrated in Figure 1.

The "alphabetic" design criteria already mentioned are the determinant criterion (D-optimality), the average-variance criterion (A-optimality), and the smallest-eigenvalue criterion (extreme optimality or E-optimality). These are illustrated in Figure 1. The D-criterion seeks to minimize the area (volume, for three or more parameters) of the confidence ellipse. Although being an intuitively reasonable thing to do, the D-criterion has an undesirable property: it may result in a long and thin ellipse, estimating one direction (linear combination of parameters) well at the expense of producing a very poor estimate of another.

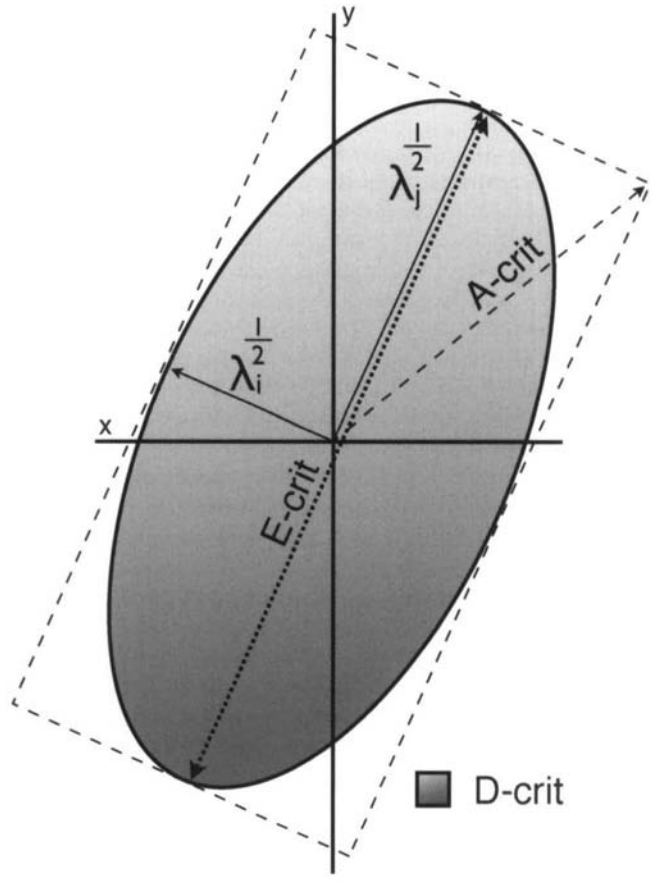

FIGURE 1. Example confidence ellipse for two parameters, with its smallest bounding box (dashed lines). The orientation of the coordinate axes (x, y) corresponds to the parameter estimates. The orientation of the major and minor axes of the ellipse correspond to the eigenvectors of the covariance matrix, and the accompanying eigenvalues (λ_i, λ_j) correspond to the square of the length of the ellipse axes. The experimental design criteria have the following effects on this confidence ellipse: the D-criterion is equivalent to minimizing the area of the ellipse (gray gradient area), the E-criterion minimizes the largest diameter (dotted arrow), and the A-criterion minimizes the length of the diagonal of the bounding box (dashed arrow).

The A-criterion minimizes the diagonal of the smallest possible box that contains the confidence ellipse; it may also be thought of as minimizing the average variance, although the variances in question are those along the axes of the ellipse and not those of the parameters. The E-criterion minimizes the largest diameter of the confidence ellipse and so ensures that the worst case of the experiment is improved. The formulae for these three criteria are shown in Table 1, expressed in terms of the eigenvalues of the Fisher information matrix, λ_i. The eigenvectors of the covariance matrix describe the axes of the confidence ellipse (a linear combination of the parameters of interest) and the corresponding eigenvalues are equal to the squares of the axis lengths. Given the assumptions above about the relationship between the covariance matrix and the Fisher information matrix, the eigenvalues of the covariance matrix are approximately reciprocal to those of the Fisher information matrix.

TABLE 1. Various experimental design criteria. The criteria are expressed in terms of the eigenvalues ϵ_i of the Fisher information matrix and the properties of the confidence ellipsoid they are related to. The lengths of the corresponding axes of the confidence ellipsoid are (for large samples) the square roots of the inverse of these eigenvalues: $\lambda_i = 1/\epsilon_i$. The total number of branches (parameters) is m. These criteria are the maximized over all designs of interest to find the optimal design, minimizing the stated property of the confidence ellipsoid.

Criterion	Formula	Property of confidence ellipsoid
A	$(\frac{1}{m}\sum_i \frac{1}{\epsilon_i})^{-1}$	Bounding box diagonal
D	$(\prod_i \epsilon_i)^{\frac{1}{m}}$	Volume
E	$\min_i \epsilon_i$	Length of longest axis

Experimental Design in the Context of Phylogenetic Inference

In the context of phylogenetic inference, there are several different types of parameters to consider: those describing the underlying substitution model, the branch lengths of the tree, and the topology of the tree. The first two types are ordinary parameters in the sense already discussed but, for simplicity, we ignore substitution model parameters in this paper and assume they are fixed at a known value (estimates derived from other data, for example). Designing experiments to increase the accuracy of substitution model parameter estimates is covered by Goldman (1998) and can be readily incorporated into the framework we describe here, as can incorporating the increase in the variance of branch length estimates caused by having to estimate model parameters from the same data. The derivatives needed in Equation (1) to calculate the entries of the information matrix corresponding to model parameters can be calculated analytically using the formulas given in Schadt and Lange (2002).

In addition to branch lengths and substitution model parameters, it is also necessary to specify a topology relating the sequences under study. The topology of the tree is a more complex parameter than the others (Yang et al., 1995)—changing the topology changes the set of parameters that are to be estimated as new branches appear and others are removed—and so does not fit into the framework outlined above. When interest is in topology rather than branch lengths themselves, we use the information about the branch lengths as a surrogate for information about the tree; this substitution is reasonable because the topology of a tree can be retrieved correctly from its pairwise distances (Zaretskii, 1965; Sempel and Steel, 2003) and from estimates of these distances so long as estimation errors are sufficiently small (Atteson, 1999; Huson et al., 1999; Mihaescu et al., 2006). Branch lengths are not, however, a perfect surrogate for topology, as discussed later.

Adding taxa to a phylogenetic inference problem increases the total number of pairwise distances available to distance matrix-based tree-building algorithms, and can increase the accuracy of inferred distances between pairs of sequences already present in the data when sophisticated estimation methods are used (Ranwez and Gascuel, 2002; Tamura et al., 2004).

Although all the theory described assumes that better designs are those that estimate parameters better, it is important to remember what the proposed experiment is hoping to achieve. Here we aim to improve the accuracy of reconstruction of an evolutionary tree connecting predetermined taxa by augmenting it with an additional branch. Such an augmentation changes the topology of the tree and so, for the reasons described above, information about two different augmentations are not directly comparable. It is the evolutionary tree connecting the original set of sequences that is of interest and this tree is common to all augmentations; the information that each augmentation gives regarding the original tree can be compared using the methods described below and can also be compared to retaining the original set of sequences but increasing the number of characters sequenced (Goldman, 1998).

Fisher Information for Augmenting a Tree

More formally, we are considering experiments where a branch of (known) fixed length δ is added to the tree, joining it in the branch indexed by b and splitting this existing branch's length in proportion ρ:1$-\rho$. Each potential augmentation is described by the triplet (b, δ, ρ), which does not depend on the values of the other branch lengths, and the original tree with branch lengths θ. If the augmented tree has branch lengths μ, it is easy to see that μ and $(\theta, b, \delta, \rho)$ are different representations of the same set of branch lengths, as shown for the example five-taxon trees in Figure 2. Because an experiment must be fully specified before performing it, the targeted branch b and the values of δ and ρ are fixed and have no information associated with them: the information the experiment gives us about the parameters of the augmented tree is equal to the information it gives us about the parameters of the original tree. By specifying the experiments in this way, all possible single-branch augmentations of the original tree can be compared on an equal basis. This form of expressing an experiment readily generalizes to augmenting a tree by adding several branches simultaneously.

In general, all branch lengths in an augmented tree can be expressed in terms of the m branches of the original tree and the three experimental design parameters (b, δ, ρ). Renumbering the branches so it is branch m that is augmented (i.e., $b = m$), the two sets of branch lengths are related by:

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{m-1} \\ \mu_m \\ \mu_{m+1} \\ \mu_{m+2} \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{m-1} \\ \rho\theta_m \\ (1-\rho)\theta_m \\ \delta \end{pmatrix} \quad (3)$$

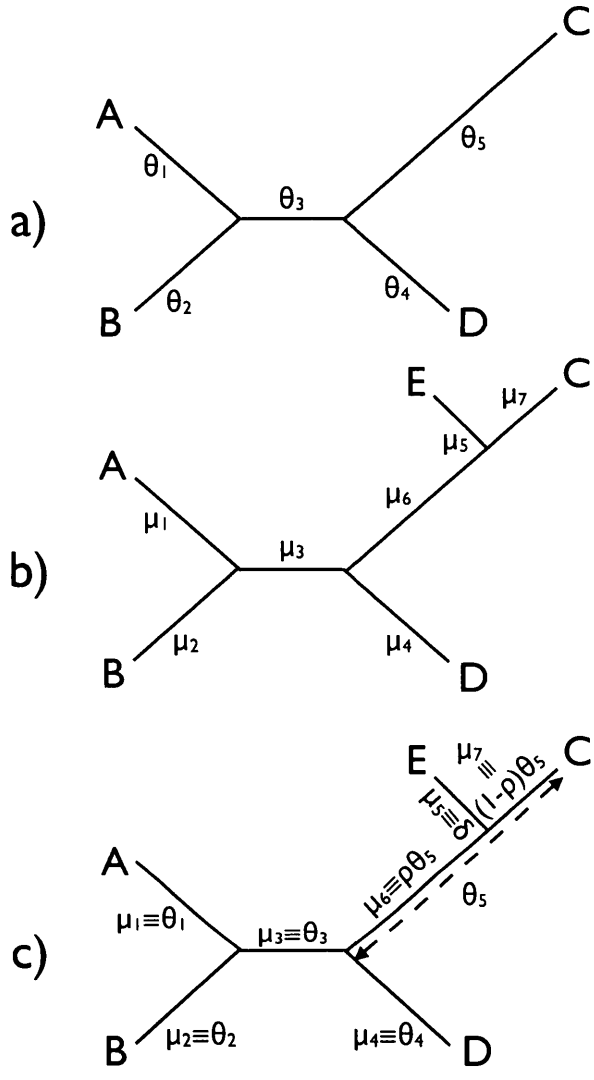

FIGURE 2. Tree augmentation and parameters. (a) The parameters $\theta_1, \ldots, \theta_5$ describe the branch lengths of the four-taxon tree shown. (b) When a branch E is added to the tree, the set $\mu_1, \ldots, \mu_7$ describes the new tree naturally. (c) Parameters $\theta_1, \ldots, \theta_5, \delta, \rho$ describe the same augmented tree in terms of the branch lengths of the original tree, the length δ of the added branch, and the proportion ρ along branch C where it is added.

and

$$M := \frac{\partial \mu}{\partial \theta} = \left(\begin{array}{c|c} \mathrm{ID}_{m-1} & \begin{matrix} Z_1 \\ \rho \end{matrix} \\ \hline \begin{matrix} Z_2 \\ \end{matrix} & \begin{matrix} 1-\rho \\ 0 \end{matrix} \end{array}\right) \tag{4}$$

where ID_{m-1} is the $(m-1)$-dimensional identity matrix and Z_1 and Z_2 are matrices of zeros of dimensions $(m-1) \times 1$ and $3 \times (m-1)$, respectively. Applying the chain-rule for differentiation, the information in terms of the original branch lengths can be expressed in terms of the new branch lengths:

$$I_{ij}(\theta) = \sum_{k,l} I_{kl}(\mu) \frac{\partial \mu_k}{\partial \theta_i} \frac{\partial \mu_l}{\theta_j} \tag{5}$$

which can be written alternatively as a product of matrices: $I(\theta) = M^T I(\mu) M$. We will use this formula for the calculations in the examples below. After transformation (reparameterization), the inverse of the Fisher information matrix is again equal to the large-sample covariance matrix of the branch lengths. The alphabetical design criteria are calculated from the transformed Fisher information matrices and optimal taxon choice can be achieved by targeting the position in the (original) tree where the criteria reach their maximum values.

Setup of Study

We used versions of the program EDIBLE (Massingham and Goldman, 2000) to calculate the information matrices for different positions of adding a branch. The transformation of the information matrices and the calculation of the A-, D-, and E-optimality criteria was performed using a combination of custom-written software and MATHEMATICA (Wolfram, 2003).

We examined the behavior of the information criteria when adding one branch at various positions of a fixed topology. Before considering some real examples, we examined three model four-taxon topologies: a symmetric tree, an asymmetric tree, and a third tree that is known to be difficult to reconstruct. First, although unrealistic as a real taxon sampling strategy, we considered the addition of an ancestral sequence (i.e., zero-length branch). Addition of such a branch splits the tree into independent parts, making the problem easier to understand intuitively. For example, adding an ancestral sequence at a node is equivalent to splitting the tree into three independent parts at that node, each part described solely in terms of the branches of the original tree. Second, for the first and second model topologies, we studied the effect on the information criteria when adding one branch, connected to the tree by a more realistic fixed length (0.1 subst./site) branch. Third, we examined the effects on the information criteria for the "difficult" tree using a range of lengths of the added branch.

For two examples based on experimental data, we investigated the effect of adding a variable length branch. In a real taxon sampling problem, potential designs are constrained by the availability of sequence data and it will generally only be possible to add a contemporary sequence. Analyzing such a situation correctly requires knowledge of which sequences might be available and at what rate they diverged. For our examples based on experimental data, we used a local-clock approximation (see below for details) so that, for each position at which a branch could be added, a single, realistic branch length was taken: length zero when the new branch coincides with an existing taxon, and longer as it is placed deeper in the tree. Although this scheme is intentionally naïve, it describes the sort of sequences that might be available and so produces a useful illustration of realistic experimental design problems. The variable-length branch strategy was applied to a simplified, semi-theoretical taxon sampling problem in the balsaminoid clade, one of our study groups (Geuten et al., 2004, 2006). For this

group of taxa, we analyzed a small alignment of chloroplast *trn*M-*atp*E sequences of length 170 bp and estimated the ML topology using the Jukes-Cantor model (JC69; Jukes and Cantor, 1969).

A further example was taken from seed plants. Considerable controversy exist about the accuracy of different aspects of seed plant phylogeny: questions such as "Which is the the basal angiosperm?" and "What are the closest relative to angiosperms?" have remained debated for many years. This makes this an example of general interest to use for illustration. For simplicity of biological interpretation, we apply our methods to a single-gene phylogeny. However, as clarified above, the Fisher information is an "expected" information, and the computational burden of our method does not depend at all on the total length of the (possibly concatenated) sequences, but only on the complexity of the model used and the number of taxa in the phylogeny. The methods can equally well be applied to problems in whole-genome analyses. To define a test phylogeny and substitution model, we analyzed sequences for the RNA polymerase II gene retrieved from GenBank using the accession numbers from Nickerson and Drouin (2004). Sequences were aligned using CLUSTALX with default penalties for gap opening and gap extension; gapped positions were excluded from further analysis, leaving 2868 characters for 11 taxa. For model selection, ModelTest version 3.7 (Posada and Crandall, 1998) was used: the Akaike information criterion selected the general time-reversible model with rate heterogeneity modeled by a proportion of invariant sites and a discrete approximation of the gamma distribution (GTR+I+G). This model was used to estimate a topology with the PAUP* version 4b10 program (Swofford, 1998), applying a heuristic search with 15 random addition replicates (hold 3). Branch length estimates were optimized by setting the maximum number of branch-length smoothing passes to 200. For comparative purposes, we also used the simpler Jukes-Cantor model (Jukes and Cantor, 1969), finding a topology and estimating branch lengths using the same procedure with PAUP*. For information calculations, parameters other than branch lengths were fixed to their values found in the phylogenetic analyses. Data matrices and trees were submitted to the TreeBase database (accession number SN3345).

RESULTS

Symmetric Four-Taxon Tree

We first analyze the behavior of the optimality criteria when adding a zero-length branch (i.e., ancestral sequence) to a perfectly symmetric four-taxon tree (Fig. 3a). The three different information criteria give very similar results in this case. Addition at or close to the internal nodes of the tree is optimal. When a branch is added to one of the tips of the tree, a base level of information is attained. This necessarily represents the information of the original tree without additional branch: taxon addition at the tips would mean adding sequences identical to those already present in the analysis and therefore no information is added and these positions are least beneficial. A clear suboptimum in the center of the internal branch can be observed for the E-criterion if a zero-length branch is added. Although less clearly observable, this suboptimum is also present for the A- and D-criteria.

When adding a taxon joined to the tree by a non–zero-length branch, the length of the additional branch also does not alter the optimal positions indicated, although the relative information gain is higher when adding a shorter branch (zero-length branch) compared to a branch with a more realistic length (0.1 subst./site). The suboptima that are present for the zero-length branch are not observed for the 0.1 subst./site instance.

The effects seen for the three criteria can readily be intuitively understood. First, adding shorter length branches is more beneficial for the information level than longer ones because they add information close to the original tree and therefore give more information about the tree. Second, it is noticed that higher information results from adding a branch close to a node in a tree. Again, the addition of a branch closest to more other branches (i.e., at a node) results in higher information. Adding a branch to a tip would be no different to adding a sequence identical to one already present, resulting in no information gain at all.

We can understand the behavior of the criteria in some more detail when we compare the individual elements of the transformed information matrices. Take the instance of adding a branch to one of the tip branches; e.g., branch C, in the symmetric four-taxon tree (Fig. 3a). The diagonal elements of the transformed information matrix give the information related to the branches in the original tree. When moving the additional branch from the tip towards the internal node, all diagonal elements of the matrix increase. The branch leading to C has most information associated with it, because it has the same length as the other branches but is also connected to the additional branch. This corresponds to our intuition that the information relating to the branches in the tree increases when an additional branch is moved closer to the "interior" of the tree.

The eigensystem of the transformed information matrix not only informs us about the individual parameters but also about the interdependence between the branches. Almost all eigenvalues of the transformed information matrix increase when the additional branch is moved from tip C towards the node joining C and D. However, the largest eigenvalue, which roughly corresponds to branch C, has a more complex behavior. This eigenvalue increases but then drops when the branch approaches closer to the node, whereas all other eigenvalues keep increasing. This suggests that the information relating to branch C, taking into account the interactions between branches, is highest when the new branch is added at some point within branch C. The smallest eigenvalue or E-criterion, which represents the region of most uncertainty in the tree, here is spread over branches A, B, and the internal branch, when "sliding" an additional branch along branch C towards the node joining C and D. This is again intuitively understandable because the branch addition results in high information for branch C.

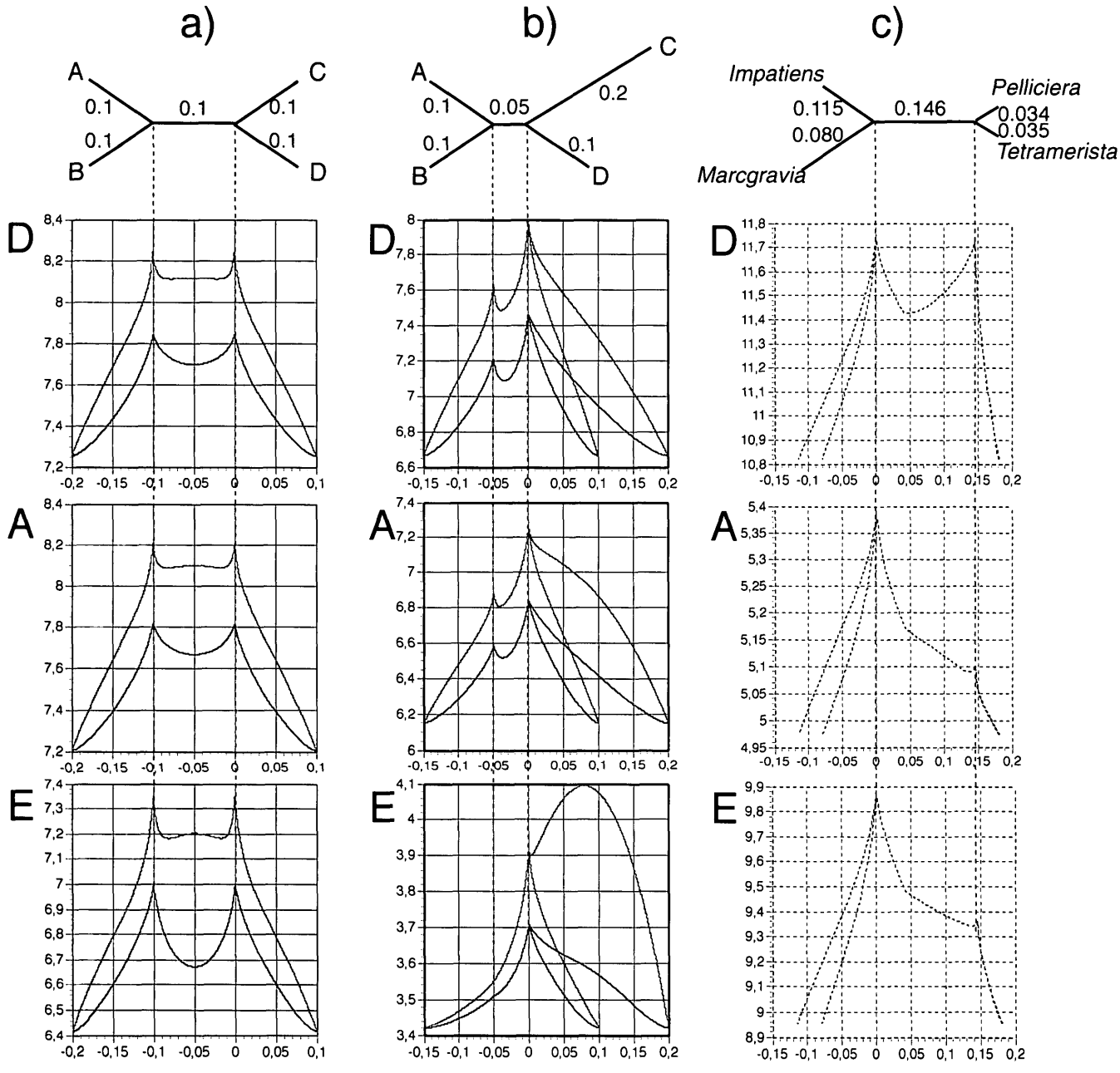

FIGURE 3. Evaluation of D-, A-, and E-optimality criteria for the addition of a branch to simple trees. The *x*-axes in the graphs correspond to the distance of the point where the branch is added from a chosen internal node, which is indicated as the value 0 on the *x*-axes. (a) D-, A-, and E-optimality criteria for the addition of a branch to a symmetric four-taxon tree. Branch lengths are equal. Length of the additional branch is either 0 subst./site (gray line) or 0.1 (black line). (b) D-, A-, and E-optimality criteria for the addition of a branch to a tree with a shorter internal branch and a long branch C. Length of the additional branch is again either 0 subst./site (upper, gray line) or 0.1 (lower, black line). Branches leading to taxa A and B overlap in the graphs. Because the branch leading to taxon C is longer (0.2 subst./site) than the branch leading to taxon D (0.1 subst./site) they can be distinguished by looking at the *x*-axis on the graphs. (c) D-, A-, and E-criteria showing the optimal position of adding a variable length branch to a balsaminoid Ericales topology (see text for details). Because of their similar length, the short branches leading to *Pelliciera* and *Tetramerista* result in overlapping lines on the graphs.

We can also try to explain the suboptimum present when an ancestral sequence (or zero-length branch) is added to the internal branch and moved from the center of the branch towards an internal node. In this case, information related to the branches towards which the additional branch is moved will increase, whereas information for the branches from which the additional branch is moved away decreases. The information related to the internal branch is highest when it is subdivided exactly in its center and decreases from its center to the node, which explains the suboptimum in the curves for the different criteria. Close to the node, however, information again increases. This is not the case for the addition of a longer branch (0.1 subst./site). For

this situation, the information associated with the internal branch only increases from its center to one of the internal nodes. Thus because a sequence is added closer to the tree in the zero-length branch case, the gain in information by bisecting the internal branch is higher.

In this symmetric four-taxon tree, the optimal strategy is to aim to subdivide the internal branch, or the tip branches close to their internal nodes. Intuitively, the most uncertain regions in the tree are the internal nodes and the deepest branch regions, which are most distant from the sequences at the tips of the tree.

Asymmetric Four-Taxon Tree

As a second theoretical example, we modified the symmetric four-taxon tree so that the internal branch is shortened to half its original length and the branch leading to taxon C is double its original length (Fig. 3b). Addition of a sequence identical to any one of the tips again necessarily results in the same base information level. The D- and A-criteria give very similar results regarding the addition of branches of different lengths. For these criteria, the node connecting taxa C and D is selected as optimal, with a suboptimum situated at the node connecting taxa A and B. This is in accord with our intuition: the long branch to node C means we know less about the region of the tree near this node and expect addition of an extra branch near here to be most beneficial.

When we add a branch of realistic length (0.1 subst./site), the E-criterion also reaches its optimum at the node connecting taxa C and D. Now, however, a suboptimum at the internal node connecting A and B is not present. Further, when we add a zero-length branch to the topology, i.e., an ancestral sequence, the optimum is reached within the longest branch C, slightly closer to the internal node than to the terminal end of the branch. This is reminiscent of long branch subdivision as a taxon sampling strategy. How can we understand this? The E-criterion seeks to minimize the largest axis of the confidence ellipsoid, which corresponds to improving the worst case in the experiment. In this case, the eigenvector corresponding to the longest axis is dominated by the branch C component, so maximizing the E-criterion is approximately equivalent to minimizing the variance of the longest branch. By placing most importance on improving the information regarding this branch, the E-criterion finds it optimal to subdivide this branch.

To investigate this effect in some more detail, we studied the behavior of the E-criterion when varying the length of the additional branch between 0 and 0.1 (results not shown). As the branch length increases from 0, the optimum within the branch to C decreases smoothly. When the branch length is approximately 0.005, the internal maximum becomes equal to the information score attained by adding the new taxon at the internal node, and for greater branch lengths the optimum position is at the internal node, as in Figure 3b for a branch length of 0.1.

Overall, the results in Figure 3b agree with our intuition that least information is associated with the longest branch of the tree, and so taxon addition in this region is most beneficial. The different information criteria each allow discovery of the optimal positions for taxon addition, placing different emphasis on different regions according to what they seek to optimize. The A- and D-criteria consider "overall information," whereas the E-criterion concentrates attention on the least-well-estimated region of the tree.

Difficult Four-Taxon Tree

As a third theoretical example, we were interested in investigating the behaviour of the criteria for a tree that is known to be difficult to reconstruct using most methods of analysis, including ML (Gaut and Lewis, 1995). The tree, as shown in the bottom right panel of Figure 4, has a short internal branch (0.01 subst./site), and in addition two short external branches (0.01 subst./site), each joined to a long branch (0.4 subst./site).

Addition of a taxon to the tips of this tree again cannot result in any increase relative to the base level of information. Because of the symmetry about the middle branch of the tree, the information attained for each of the criteria is the same for each of the internal nodes and along equivalent branches. In the previous examples, the A- and D-criteria behaved similarly. In this example, all three criteria behave differently along the branches and nodes. The behavior of the D-criterion is very similar to what can be seen in both our previous examples, with optima at the internal nodes, and can be interpreted in the same way.

In contrast, the A-criterion indicates an optimum position within either one of the long branches in the tree. This is again reminiscent of long-branch subdivision, a rule of thumb in molecular systematics when trying to improve the accuracy of tree estimation (e.g., Poe, 2003). The exact location of the optimum is close to the middle of these branches, but slightly shifted towards the internal nodes. The longer the added branch, the more the location of this optimum is shifted towards the nodes. This effect is somewhat similar to what could be seen for the E-criterion, when adding a branch of infinitesimal length to the asymmetric tree in the previous example. However, here the effect is also strong for addition of longer, realistic branch lengths.

Different from both the D- and A-criteria, the E-criterion advises addition of a branch bisecting the internal branch. The E-optimality decreases slowly along the longest branches and tapers off towards the end of this branch. It seems that for this extreme case, the suboptimum observed for the E-criterion as in Figure 3a is now transformed into an absolute optimum. As the E-criteria is sensitive to the worst estimate in the tree (the longest axis of the confidence ellipsoid), least information is present furthest away from the tips of the tree. The different behavior of the criteria illustrates their complementarity in finding optimal designs. The E-criterion identifies the longest axis of the confidence ellipse. This is situated in the middle of the internal branch. The D-criterion is strongly influenced by this and selects the area around the internal branch as most optimal. Because the

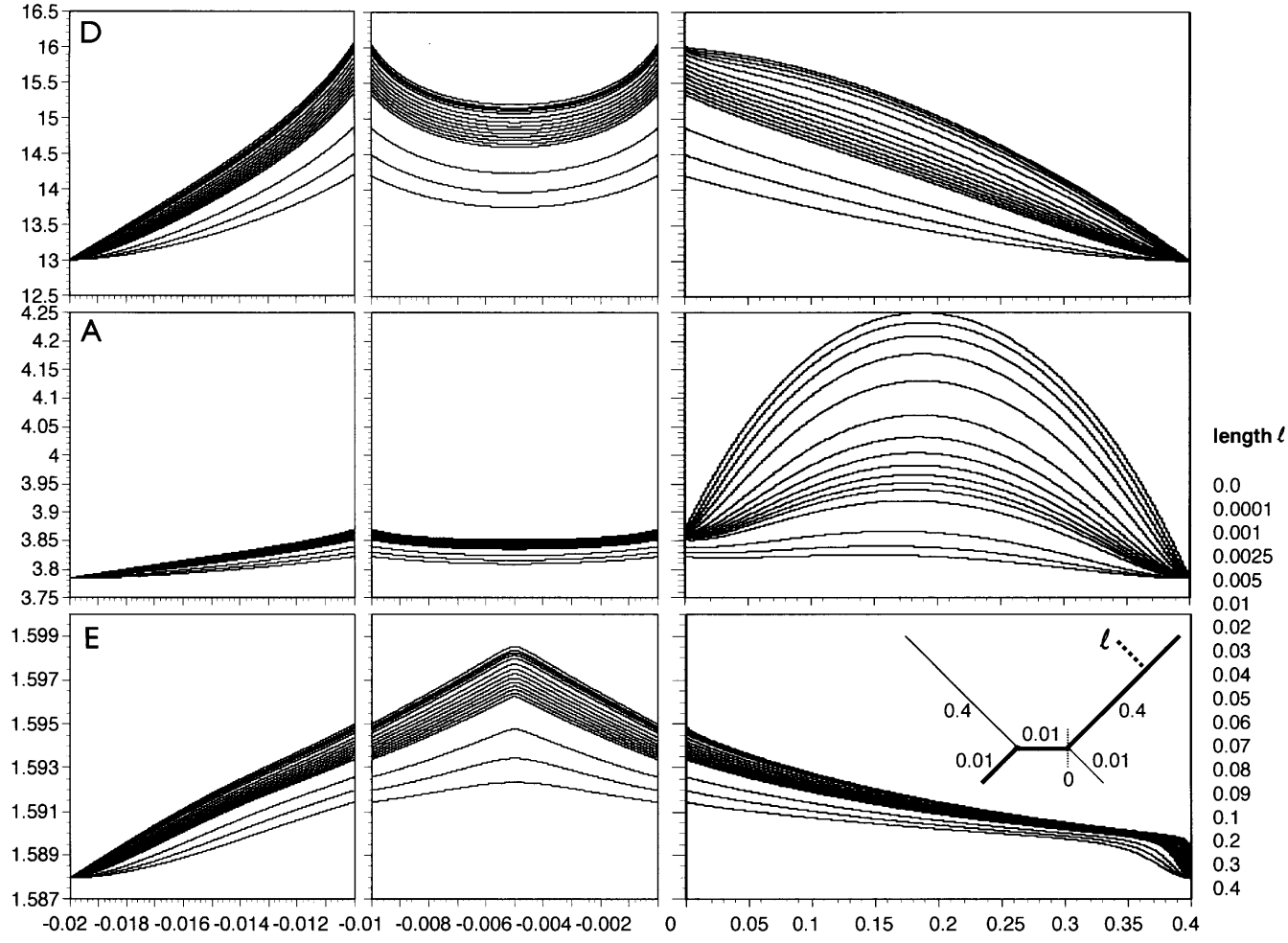

FIGURE 4. Evaluation of D-, A-, and E-optimality criteria for the addition of a branch to a tree that is notoriously difficult to reconstruct. The original tree, with branch lengths, is shown in the bottom right panel. The y-axis measures how optimal it is to add a taxon to this original tree, for each criterion. The x-axis shows where the branch is added along the original tree: this path of attachment points for the added branch corresponds to the thicker lines in the tree shown. The left-hand panels show the information criteria as the added branch ℓ moves from the tip of the short branch (indicated as $x = -0.02$) towards its ancestral node ($x = -0.01$); center panels show the information criteria as the added branch proceeds towards the other internal node ($x = 0$); and right-hand panels show the information criteria as the added branch moves along the long branch of the original tree ($x = 0$ to $x = 0.4$; note different x-axis scale). In all panels the length of the added branch varies, with the different lengths (ordered top to bottom corresponding to the order of the curves shown in all plots) shown in the list to the bottom right of the figure.

D-criterion takes the other axes of the confidence ellipsoid also into account, the optimum is shifted towards the nodes. The A-criterion aims at minimizing the average variance and is less sensitive to the parameter with largest variance.

Balsaminoid Ericales

The balsaminoid Ericales clade has been studied by Geuten et al. (2004, 2006) and here a possible topology is examined. In this tree, the internal branch is longest (Fig. 5). In real taxon sampling problems, the range of feasible experiments is restricted because only contemporary taxa can be added to the tree; a somewhat arbitrary scheme that adjusts the length of additional branches depending where they are added in the tree was created to capture this effect. (Of course, our example is illustrative and for other problems users of our methods are free to develop whatever taxon-sampling schemes they consider appropriate.) In this scheme, closer to the tips of the tree, very short branches are added. Towards the internal nodes and branches, longer branches are added (Fig. 5). In detail: at the node a joining *Impatiens* and *Marcgravia*, a branch of the average branch length to *Impatiens* and *Marcgravia* is added. Similarly, at the node b joining taxa *Pelliciera* and *Tetramerista*, a branch of the average length of the terminal branches leading to these taxa is added. When a taxon is added to a terminal branch, its length varies linearly from the internal node (length as above) to the tip (where the additional branch will have length zero). For the internal branch, the midpoint along this branch (m) and the length of the added branch on

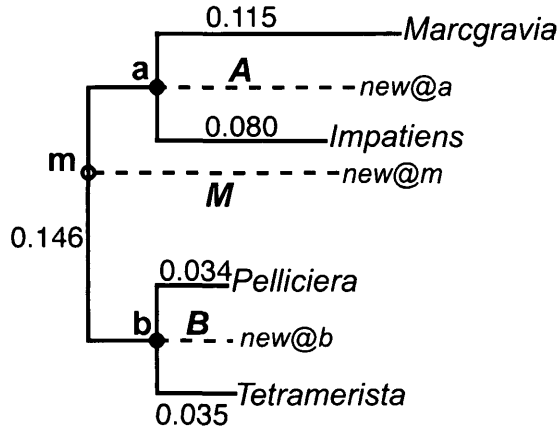

FIGURE 5. Variable additional branch length scheme for the balsaminoid Ericales example. The length of the newly added branch varies with the position in the tree where the branch is added. At positions between the possible additional branches labeled "new," the length of the branch leading to the additional taxon varies linearly between the values shown (or between 0 and the single value shown, for external branches). At node a, the length of the new branch is $A = 0.0975$, which is the average of the branches leading to *Impatiens* and *Marcgravia*, and analogously for node b ($B = 0.0345$). Along the internal branch, between nodes a and m, the length increases linearly between A and M, and similarly the length decreases linearly between nodes m and b. The length M at position m equals the sum of A and the distance between a and m; similarly, M also equals the sum of B and the distance from m to b.

this midpoint (M) were taken so that the average of the external branches at one side (A or B) summed with the distance to the midpoint along the internal branch was equal on both sides of the midpoint and equal to the length of M. When a taxon is added elsewhere in the internal branch, its length varies linearly between the internal nodes and the midpoint of the tree.

The results of this analysis are shown in Figure 5. In this more realistic design scheme, all criteria indicate internal nodes as the optimal positions to add a branch. The A- and E-criteria select the node joining *Impatiens* and *Marcgravia*, which seems reasonable as this is between the longest (and therefore most uncertain) branches of the tree. The D-criterion has no real preference as to which internal node is chosen. This is surprising, but becomes clearer after inspecting the elements of the information matrix and is related to a strong correlation between the *Pelliciera* and *Tetramerista* branches when the new taxon is added at the *Impatiens-Marcgravia* internal node.

Seed Plant Phylogeny from RNA Polymerase II Sequences

We analyzed a set of seedplant RNA polymerase II sequences using the GTR + I + G substitution model. This gave the phylogeny shown in Figure 6a. The tree depicts plausible relationships between seed plants, congruent with relationships found in recent analyses (Chaw et al., 2000; Hajibabaei et al., 2006). We used a variable branch length addition scheme, as with the balsaminoid Ericales example, to suggest realistic branch lengths depending on where the tree was to be augmented. Calculations of the A-, D-, and E-optimality criteria all indicate node b, connecting *Arabidopsis* to the tree, and its near neighborhood as optimal for the addition of another taxon. The second node chosen by all the criteria is node f, connecting *Psilotum* to the interior of the tree. However, the general order of preference for different locations in the tree is different for the three criteria. The D-criterion finds the f, g, and h nodes to be better than the a, c, d, and e nodes. Besides choosing nodes b and subsequently f as best, the A-criterion differentiates less between other regions in the tree. The E-criterion, again choosing nodes b and f as best, shows a clear preference for this part of the tree, with the next suboptima at the intervening nodes c, d, and e.

To address in more detail what happens when transformation is made back to the original branch lengths, we compared the changes in edge length variances resulting from adding a branch at the different internal nodes. Results are shown in Figure 6c. The greatest reduction in variance is achieved for branch 5 when a branch is added at node b. Branch 3 also has a significantly reduced variance in this case; note that branches 5 and 3 are the internal branches connected to node b. Similarly, a large reduction in variance is obtained for both branches 11 and 13 when adding a branch to node f, and a large reduction in variance is obtained for the terminal branches leading to *Zea* and *Oryza* when adding a branch to their common ancestor node a. In general, there is a strong correspondence between the position in the tree at which a taxon is added and the branches that show the greatest reduction in variance.

As in all experimental design, the methods presented here are derived from theory applicable to the true tree topology and branch lengths. However, it is relevant to know what happens when applying the information criteria to an incorrect tree, a situation that might occur in practice in phylogenetics. Possible reasons for incorrect trees include model misspecification (usually oversimplicity), a bias that could be present in genome-scale studies, or insufficient sequence length, a possibility in single-gene/many-taxa studies. In general, it is not known what improvement can be expected from taxon addition in remedying different kinds of biases in tree reconstruction. We illustrate here that it is possible for taxon addition, guided by the use of experimental design criteria, to remedy such a bias present in a phylogeny estimation. The tree in Figure 7a, derived from the RNA polymerase II data using ML under the Jukes-Cantor model (1969), shows very improbable relationships. It places a core eudicot, *Arabidopsis*, and *Psilotum* together and finds no sister-group relationship for *Welwitschia* and *Pinus*, contradicting most recent phylogenies for seed plants (Chaw et al., 2000; Hajibabaei et al., 2006). We chose this example because a wide consensus exists about this part of the phylogeny being wrong. For real-life taxon sampling problems, such knowledge will not be available; such an example would not allow us to illustrate that an improvement can result from judicious taxon addition using information calculations.

As shown in Figure 7a, the same region in the tree is identified by all three criteria as optimal for taxon addition. Although the precise limits of this optimal

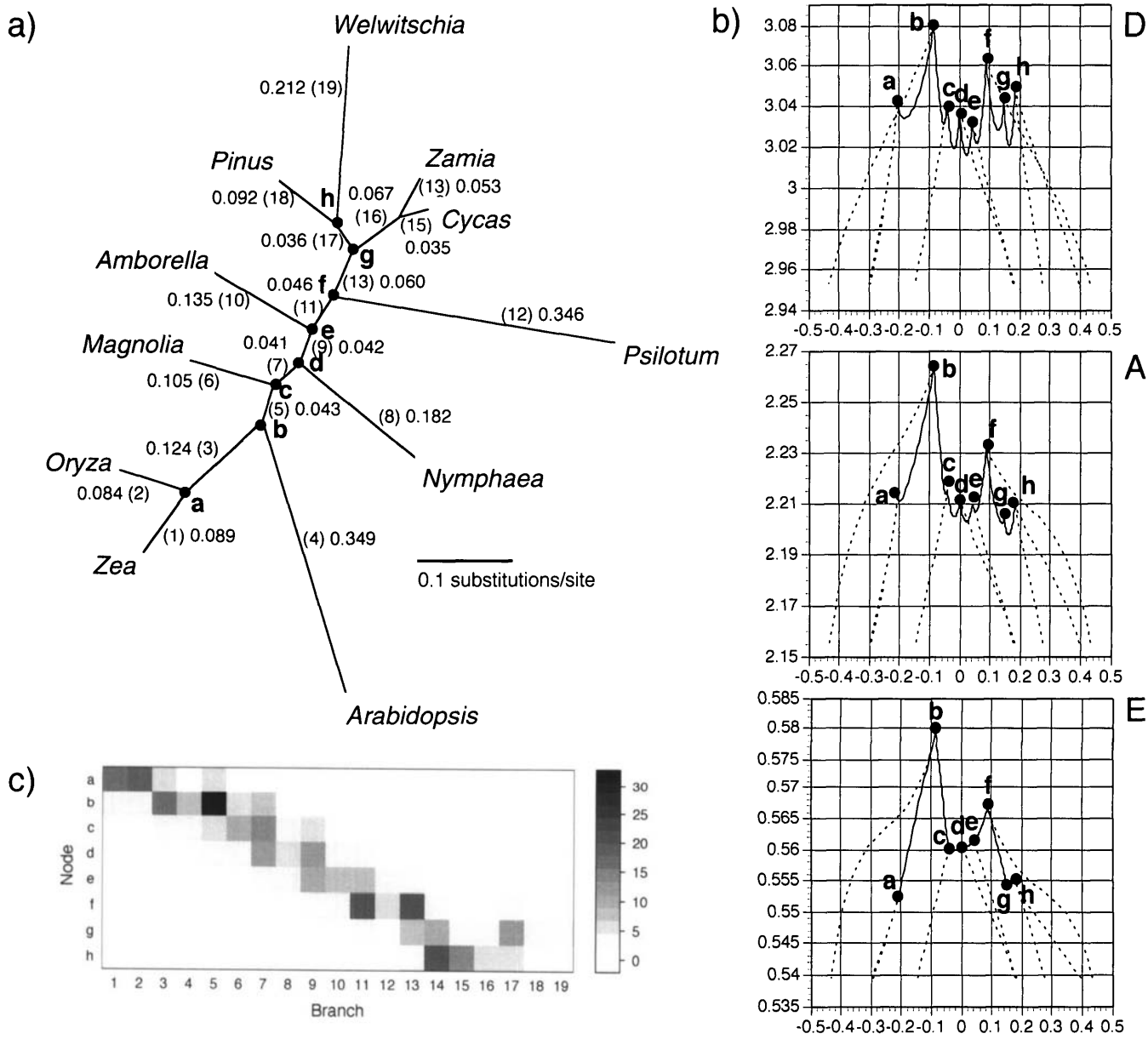

FIGURE 6. Evaluation of D-, A-, and E-criteria for adding a branch to a seed plant phylogeny. The scheme for determining the length of the additional branch is described in the text. The *y*-axes measure the different optimality criteria. The *x*-axes show where in the tree a branch is added, which can be more easily followed using the node labeling on the graphs and the tree. (a) The ML phylogeny inferred using the GTR + I + G model. The internal nodes are labeled alphabetically, and these labels referred to in (b) and (c). Branches are numbered, and the same numbering used in (c). (b) Graphs showing the behavior of the three information criteria as an extra branch (taxon) is sampled at different positions of the tree in (a). (c) The percentage reduction in variance for each of the branches (1–19), depending on which node (a–h) the new sequence is added at. The highest reduction of variance (31%) is in branch 5 (internal branch between *Arabidopsis* and *Magnolia*) and is attained when a new sequence is placed at node (b).

region are smaller or broader depending on the criterion used, in general a relatively broad area around the node connecting *Arabidopsis* and *Psilotum* is optimal for targeting. Two strategies seem plausible: either adding a known relative of *Arabidopsis* or adding a known relative of *Psilotum*. No appropriate sequence closely related to *Psilotum* could be found in GenBank so we chose a sequence from the core eudicot *Antirrhinum majus*. *Antirrhinum* is an asterid genus, reasonably distant from the rosid genus *Arabidopsis*, but can still be expected to group with *Arabidopsis*. Based on this extended alignment, the ML phylogeny was reestimated using the Jukes-Cantor model. The resulting topology, depicted in Figure 7b, restores plausible relationships of the seed plants congruent with relationships found in recent analyses (Chaw et al., 2000; Hajibabaei et al., 2006).

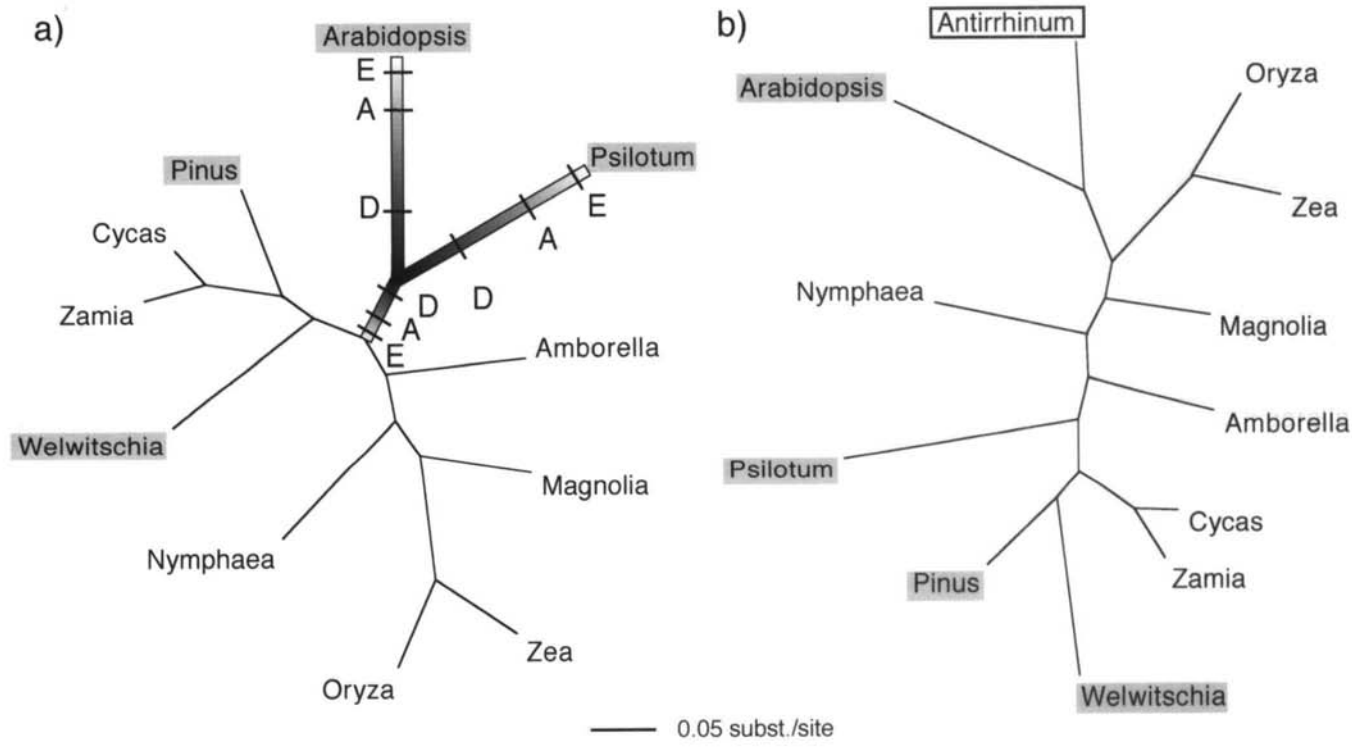

FIGURE 7. (a) Seed plant phylogeny estimated using the Jukes-Cantor model. All three information criteria, when applying the variable branch length scheme, select the same region of the tree for optimal taxon addition (grey-scale coloring in the region of *Arabidopsis* and *Psilotum*). Different criteria select slightly larger or smaller regions (delimited by bars labeled D, A, or E according to which criterion is used). (b) Altered relationships in seed plant phylogeny when adding the sequence of *Antirrhinum majus* with the aim of adding a branch in the region as suggested by (a). Note the positions of *Pinus*, *Welwitschia*, *Arabidopsis*, and *Psilotum* (highlighted) have changed.

DISCUSSION

Currently there is no method other than simulation studies to assess where to optimally add taxa to a given topology. We have investigated several information-based experimental design criteria to gain insight into where a branch should optimally be added to improve inference. Our examples illustrate that the information calculations, which are a direct extension of standard likelihood theory, give plausible advice on where to add a taxon.

Because the branch lengths are readily observable for a tree, it may be tempting to think that the different locations of the optima for the D-, A-, and E-criteria can be identified without calculation of the Fisher information, so replacing calculations like ours with a few simple rules. This is not the case in practice, because there is no direct correspondence between the eigenvectors of the information matrix (axes of the confidence ellipsoid) and the branches of the tree. By inspecting the transformed information matrices and their corresponding eigensystem, it is possible to identify the balancing factors in the tree in retrospect, but in general, calculations will be necessary to identify the optimal position to add a branch in the tree. At this stage in our research, we are not able to give firm recommendations about which criteria should be used for a particular problem. Nevertheless, the ability of the information-based approach to give interpretable results means we have a technique for future investigation and comparison of criteria, as well as a theoretical basis that affords a greater understanding of the problem of experimental design in phylogenetics than does any other approach we know.

We cannot currently assess the exact relation of information to topological accuracy and further exploration of the connections may inspire new, more closely associated, experimental, design criteria. Nevertheless, the methods applied here result in strikingly similar advice for taxon addition to that reported in simulation studies. Graybeal (1998) and Poe (2003) found that trees reconstructed least accurately were those where taxa were added close to the tips of long branches and trees reconstructed most accurately were those where a taxon was added close to the base of long branches. This is exactly what we found by using information calculations.

All of the three criteria show a general preference for augmenting the tree at deep internal nodes connected to long branches, increasing information about the more uncertain regions of the tree. For more extreme phylogenetic trees, with combinations of branch lengths that make them difficult to reconstruct accurately (our third example), the information criteria do not necessarily choose nodes as their optimal location in the tree for targeted taxon addition. In these cases, targeting a long branch for subdivision can be the most optimal strategy. It is reassuring that the examples given are in agreement

with rules of thumb derived from experiments and simulation studies.

The total number of possible site patterns grows exponentially as more taxa are added, making exact computation of information infeasible when the number of taxa is large. In this case, however, it is still possible to estimate the Fisher information using Monte Carlo methods (Ripley, 1987; Massingham and Goldman, 2000) and so explore alternative experimental designs. The computational work can be further reduced by noting that most experimental designs are not biologically feasible: information only needs to be calculated for the addition of those taxa available to the experimenter.

Because the topology is unknown, it must be estimated from the data, probably using one of the many heuristic methods available. Atteson (1999) showed that the popular neighbor-joining heuristic correctly reconstructs the topology if the error in all pairwise distances is less than half the smallest branch length (see also Huson et al., 1999; Mihaescu et al., 2006), so to reconstruct the correct topology, it is sufficient to accurately estimate all pairwise distances. Because the pairwise distances are linear combinations of the branch lengths of the true tree, and this relationship may remain for some pairwise distances in a slightly incorrect tree, their respective errors are closely connected and the Fisher information (via the Cramer-Rao lower-bound) may be considered to approximate a limit on how well these pairwise distances can be estimated for any estimator. The E-criterion is an especially good fit to this problem, because improving this information measure by augmenting a data set ensures that the error of the worst possible linear combination of branch length parameters must decrease.

It has been suggested that it is not advisable to use a phylogeny-based method to improve phylogenies (Lyons-Weiler and Hoelzer, 1997), the argument being that the topology and conditional branch lengths could be wrong. The information calculations we have presented rely on the accuracy of both the assumed branch lengths and topology, and errors in either will affect the estimated optimality of the overall design. However, the seed plant example in Figure 7 shows that our methods can give meaningful advice on how to improve even a wrong topology. In addition, this potential weakness of the method may in future be remedied by averaging over parameters and plausible topologies, perhaps using ideas from the field of Bayesian experimental design (see Chaloner and Verdinelli, 1995, for a review).

The lack of general guidelines on how to design a phylogenetic inference experiment has led to a noticeable gap in many research articles: details and discussion of how taxa were chosen are often absent from the Materials and Methods section, the choice perhaps being guided more by intuition than any concrete criterion. The trend towards sequencing large regions of a few genomes (ENCODE 2004), and interest in producing phylogeny from entire genomes (e.g., Rokas et al., 2003; Goremykin et al., 2003) necessarily means that fewer taxa are sequenced and it becomes all the more important to make a judicious choice; having some tool to guide this choice should be of interest. Better experimental design means increased statistical power for an equal cost, or that equivalent accuracy can be achieved more cheaply: exploring alternative designs is not just about justifying the taxa chosen but should be part of every scientist's due diligence in minimizing error and ensuring that research funding is spent efficiently.

ACKNOWLEDGEMENTS

K.G. was financially supported by research grants of the K. U. Leuven OT/01/25 and the Fund for Scientific Research—Flanders (Belgium) (G.0104.01;1.5.069.02;1.5.061.03) awarded to E.S. and would like to acknowledge a fellowship from the D. Collen foundation and the Belgian American Educational Foundation. N.G. received support from the Wellcome Trust. T.M. was funded by BBSRC grant 721/BEP17055. K.G. would like to thank P. Schols for enthusiastic feedback on several versions of the manuscript. The authors would like to thank O. Gascuel, R. Page, and three anonymous reviewers for valuable feedback.

REFERENCES

Atkinson, G., and A. Donev. 1992. Optimum experimental designs. Oxford University Press, Oxford, UK.

Atteson, K. 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25:251–278.

Chaloner, K., and I. Verdinelli. 1995. Bayesian experimental design: A review. Stat. Sci. 10:273–304.

Chaw, S.-M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc. Natl. Acad. Sci. USA 97:4086–4091.

Edwards, A. 1972. Likelihood. Cambridge University Press, Cambridge, UK.

ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fisher, R. A. 1926. The arrangement of field experiments. J. Ministry Agriculture Great Britain 33:503–513.

Fisher, R. A. 1935. The design of experiments. Oliver and Boyd, Edinburgh.

Gaut, B. S., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

Geuten, K., A. Becker, K. Kaufmann, P. Caris, S. Janssens, T. Viaene, G. Theissen, and E. Smets. 2006. Petaloidy and petal identity MADS-box genes in the balsaminoid genera *Impatiens* and *Marcgravia*. Plant J. 47:501–518.

Geuten, K., E. Smets, P. Schols, Y.-M. Yuan, S. Janssens, P. Küpfer, and N. Pyck. 2004. Conflicting phylogenies of balsaminoid families and the polytomy in Ericales: Combining data in a Bayesian framework. Mol. Phyl. Evol. 31:711–729.

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

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and H. F. Hellwig. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol. Biol. Evol. 20:1499–1505.

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

Hajibabaei, M., J. Xia, and G. Drouin. 2006. Seed plant phylogeny: Gnetophytes are derived conifers and a sister group to Pinaceae. Mol. Phyl. Evol. 40:208–217.

Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522–529.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Huson, D. H., S. M. Nettles, and T. J. Warnow. 1999. Disk-Covering, a fast-converging method for phylogenetic tree reconstruction. J. Comput. Biol. 6:369–386.

Jukes, T. H., and C. R. Cantor. 1969. Mammalian protein metabolism. Pages 21–132 *in* Evolution of protein molecules (H. N. Munro and J. B. Allison, eds). Academic Press, New York.

Kiefer, J. 1959. Optimal experimental design. J. Royal Stat. Soc. 21:272–319.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

Lyons-Weiler, J., and G. A. Hoelzer. 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. Mol. Phyl. Evol. 8:375–384.

Martin, W., O. Deusch, N. Stawski, N. Grünheit, N., and V. Goremykin. 2005. Chloroplast genome phylogenetics: Why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:203–209.

Massingham, T., and N. Goldman. 2000. EDIBLE: Experimental design and information calculations in phylogenetics. Bioinformatics 16:294–295.

Mihaescu, R., D. Levy, and L. Pachter. 2006. Why neighbor-joining works. arXiv:cs.DS/0602041 v2 (http://arxiv.org/abs/cs.DS/0602041).

Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. 47:61–76.

Nickerson, J., and G. Drouin. 2004. The sequence of the largest subunit of RNA polymerase II is a useful marker for inferring seed plant phylogeny. Mol. Phyl. Evol. 31:403–415.

Nylander, J. A. A. 2001. Taxon sampling in phylogenetic analysis: Problems and strategies reviewed. Introductory research essay no. 1, Department of Systematic Zoology, Uppsala University.

Pawitan, Y. 2001. In all likelihood: Statistical modelling and inference using likelihood. Oxford University Press, Oxford, UK.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Poe, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47:18–21.

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. 17:1854–1858.

Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

Ranwez, V., and O. Gascuel. 2002. Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. Mol. Biol. Evol. 19:1952–1963.

Ripley, B. D. 1987. Stochastic simulation. John Wiley and Sons, New York.

Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rosenberg, M. S., and S. Kumar. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl Acad. Sci. USA 98:10751–10756.

Rosenberg, M. S., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan and S. G. Brady. 2000 Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17:782–797.

Schadt, E., and K. Lange. 2002. Codon and rate variation models in molecular phylogeny. Mol. Biol. Evol. 19:1534–1549.

Sempel, C., and M. Steel. 2003. Phylogenetics. Oxford Lecture Series in Mathematics and Its Applications, 24. Oxford University Press, Oxford, UK.

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y. L. Qiu, M. W. Chase, J. S. Farris, S. Stefanovic, D. W. Rice, J. D. Palmer, and P. S. Soltis. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": A cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.

Swofford, D. 1998. PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0 beta. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl Acad. Sci. USA 101:11030–11035.

Wolfram, S. 2003. Mathematica, 5th Edition. Addison-Wesley, Reading, Massachusetts.

Yang, Z. 1996a. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42:294–307.

Yang, Z. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556.

Yang, Z., N. Goldman, and A. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

Zaretskii, K. A. 1965. Constructing trees from the set of distances between pendant vertices. Uspehi Mathematiceskih Nauk 20:90–92 (in Russian).

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

First submitted 12 October 2006; reviews returned 29 January 2007; final acceptance 8 April 2007
Associate Editor: Olivier Gascuel