MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 4, No. 1, March, pp. 64–71, 1995

A Frequency-Dependent Significance Test for Parsimony

MIKE STEEL,* PETER J. LOCKHART,† AND DAVID PENNY†

*Mathematics Department, University of Canterbury, Christchurch, New Zealand; and †Molecular Genetics Unit, Massey University, Palmerston North, New Zealand

Received May 19, 1994; revised September 30, 1994

We describe techniques for assessing evolutionary trees constructed by the parsimony criteria, when sequences exhibit irregular base compositions. In particular, we extend a recently described frequency-dependent significance test to handle any number of taxa and describe a modification of the Kishino–Hasegawa sites test. These modifications are useful for detecting historical signals beyond those patterns which arise purely from irregular base compositions between the compared sequences. We apply the test to extend our earlier studies on chloroplast origins using 16S rDNA sequences, where a failure to compensate for irregular base compositions between the compared sequences provides statistically significant support for unjustified phylogenetic inferences. We also describe how the techniques can be modified to determine how "tree-like" data are, given independent variation in the base frequencies.

One of the earliest, and still most widely used, methods for constructing phylogenetic trees is the maximum parsimony technique. Given a tree T, each of whose leaves correspond to an aligned sequence, and a collection C of aligned sequences, the length of T for C—denoted $L(C, T)$—is the least number of point mutations (substitutions) that needs to occur across the edges of T to account for the observed variation in the sequences.

To make this notion more precise, it is useful to regard a collection C of k parsimony sites in n aligned sequences as k functions $\chi_1, \ldots, \chi_k$, where each χ_j assigns sequence i ($i = 1, \ldots, n$) one of r possible states ($r = 4$ for DNA sequences; $r = 2$ for purine/pyrimidine sequences; $r = 20$ for amino acid sequences). For any tree T whose leaves (degree one vertices) are numbered $1, \ldots, n$, let $L(\chi_j, T)$ be the minimal number of edges of T which must have different states assigned to their ends in order to extend the function χ_j to all the vertices of T (an extension which realizes this minimization is said to be minimal). The length of T for C, written $L(C, T)$ is the sum

$$L(C, T) = \sum_{j=1}^{k} L(\chi_j, T),$$

which can be computed efficiently, indeed in $O(nk)$ steps, using Fitch's algorithm (see Hartigan, 1973). However, minimizing this function [finding the tree that minimizes $L(C, T)$] is not easy, and a fast algorithm is unlikely to exist since this problem has been shown to be NP-hard (Graham and Foulds, 1982). Nevertheless, a branch-and-bound algorithm due to Hendy and Penny (1982) works acceptably fast on "good" data for values of n up to about 20.

The *parsimony principle* regards T as a better estimate that T' of the true evolutionary tree whenever T requires fewer mutations than T'; that is, whenever $L(C, T) < L(C, T')$. Consequently, the tree (or trees) which minimizes $L(C, T)$, the *maximum parsimony tree(s)*, is taken as the best estimate of the "true" tree. Many phylogenetic studies are based on this criterion (Stewart, 1993).

There are two problems associated with this otherwise appealing and simple scheme. First, it has been known for many years (Felsenstein, 1978) that under simple stochastic models of nucleotide substitution, parsimony can be statistically inconsistent for certain parameter choices (constituting the so-called "Felsenstein Zone"). That is, the method will tend to select an incorrect tree with a probability tending to 1 as the sequence length grows (however, as pointed out by Steel *et al.* (1993b), if parsimony is applied to suitably transformed data, parsimony will be consistent for certain nucleotide substitution models). Defenders of parsimony have pointed out that the assumptions implicit in these stochastic models are overly severe and thereby unrealistic; others (notably "pattern" cladists) contend that it is better not to assume anything about the evolutionary process. While we do not agree with this second position, let us turn instead to a second problem.

A further difficulty with parsimony arises when the sequences exhibit variation in the frequency of their states, acquired independently and not due to shared ancestry. In this case parsimony will tend to group together sequences according to their base compositions. This problem has been highlighted recently (particularly for ribosomal RNA) by a number of authors [see Lockhart *et al.* (1992), Hasegawa and Hashimoto (1993), Olsen and Woese (1993), and Klenk *et al.* (1994)].

1055-7903/95 $6.00

Attempts to address this problem have included (1) Sidow and Wilson's (1990) application of compositional statistics to extend the "evolutionary parsimony" method of Lake (1987), and (2) the development of a transformation for recovering a tree when sequences evolve under nonstationary stochastic models, by Lake (1994), and Steel (1994) (see also Lockhart *et al.*, 1994). A third approach (which was confined to just four taxa) will be described shortly. A disadvantage of Method 2 is that it assumes a constant substitution rate at the variable sites, while Method 1 is based on a more restrictive stochastic mechanism than Method 2 (although it does allow rate variation between sites). Both approaches have additional drawbacks: the transformation in Method 2 is highly nonlinear and thereby biased (a potential problem for short sequences), while evolutionary parsimony has been shown in simulations (Huelsenbeck and Hills, 1993) to be considerably weaker in extracting phylogenetic information from sequences than other methods, such as parsimony (more precisely, provided one is not in the "Felsenstein zone" of inconsistency for parsimony, the probability of recovering the correct tree on four taxa using sequences of moderate length is generally higher for parsimony than for evolutionary parsimony).

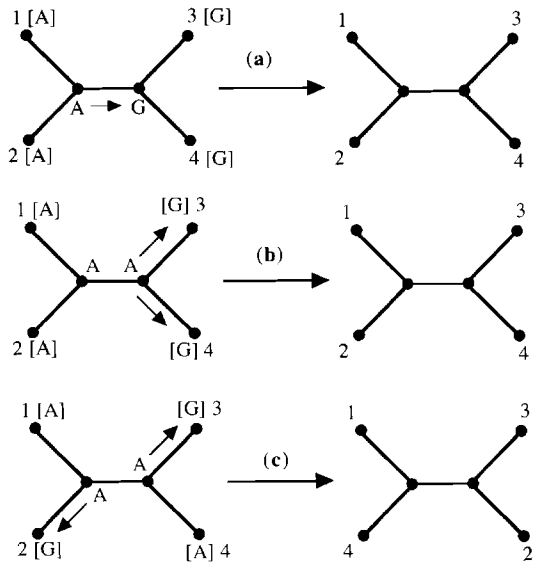

FIG. 1. (a) A most parsimonious tree for a column (site) of the data, which shows a single substitution (no homoplasy); (b) A most parsimonious tree in which the column of data has had two parallel changes and a directional substitution pressure has resulted in a pattern which is shortest on the correct tree; (c) A most parsimonious tree in which the column of data has had two parallel changes and a directional substitution pressure has resulted in a pattern which is shortest on an incorrect tree.

A NULL MODEL FOR PARSIMONY

In this paper we adopt a different approach to that outlined in the previous paragraph and consider how parsimony can be modified to handle the variation of nucleotide frequencies between sequences. It is based on a simple idea, described and developed for four sequences in Steel *et al.* (1993a). In this method parsimony scores are adjusted by subtracting the contribution to the score that one would expect if the sequences were generated randomly and if the nucleotide frequency bias was maintained from that in the observed sequences. The aim of this paper is to extend this approach to any number of sequences.

First we address a potential criticism that could be raised against such an approach. In subtracting away effects that are caused by nucleotide variation it could be argued that we are throwing away informative characteristics of the data, as that variation may itself reflect a shared history (a point raised by one of the referees). In some cases, this may be true (Fig. 1b). However, in other cases, we will avoid being falsely misled by nucleotide frequency variations (Fig. 1c); thus our approach is a conservative one. While it may lose some information, it is less likely to significantly support an incorrect phylogeny.

To explain this further, suppose the sequence sites evolve independently on a tree according to a stationary Markov model. In this case there should be little variation in the frequencies of states for sequences of realistic length (and this variation must tend toward 0 as the sequences grow in length), so that ignoring this variation is likely to be of little consequence. On the other hand suppose the sequences do exhibit significant variation in the frequencies of their states, in which case any associated Markov substitution model is unlikely to be stationary. It may be that the changing bias in the frequencies of the states implied by this nonstationarity is coherent with the evolutionary history of the sequences, in which case our correction will apparently lose phylogenetic information. However, when this is not the case (so that distantly related sequences have similar compositions) failure to correct for differences in the frequencies of states between sequences can be misleading by bringing together sequences with similar frequencies. This has been demonstrated analytically in Lockhart *et al.* (1992) and with three real data sets by Lockhart *et al.* (1994).

Thus we associate with the original data a *null model* in which sequences are generated entirely randomly, but are subject to the condition that the expected frequency of the states in each sequence is equal to its actual observed value. More precisely, if the frequency of the state α in the sequence i is π_i^α, then, under the associated null model, the states in the sequences are taken to be independent random variables, with π_i^α being the probability that any given site in sequence i has state α.

We consider two underlying models (Fig. 2) which could give rise to this null model. Both of these under-

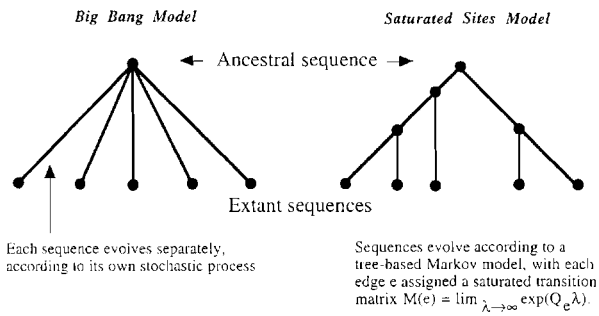

FIG. 2. Two scenarios leading to the null model: (i) the "big bang" model, wherein sequences evolve on an unresolved (star) tree according to standard stochastic models (with transition rates assigned to the edges of the tree) and (ii) the saturated sites model, in which sequences evolve on a resolved tree, but the transition rates are so high that the hierarchical information is lost.

lying models are special cases of the commonly used tree-based model of sequence evolution in which sites evolve independently and identically on a tree according to a Markov process (see Rodriguez *et al.*, 1990). In order to realize our null model such a tree-based model needs to be modified in one of two ways: either the underlying tree must be a "star" tree (the big bang model shown in Fig. 2) or the substitution rates on the edges of the tree must be very large (the saturated sites model in Fig. 2). We now briefly describe these two models. The star tree that underlies the big bang model was suggested by Thompson (1975) as a suitable null hypothesis. In the big bang model, sequences and their constituent sites evolve separately and independently from an ancestral sequence, according to a Markov process (see for instance Rodriguez *et al.*, 1990). This Markov process can vary between the sequences (otherwise one cannot explain the variation in nucleotide frequencies) but is constant across the sites of a given sequence. In the saturated sites model, the underlying historical tree is resolved, but this hierarchical information is effectively lost at the sequence level because the expected number of substitutions on each edge is very large. Formally, this is achieved by first writing the transition matrix $M(e)$ associated to edge e in the form $\exp(Q_e\lambda)$, where Q_e is the intensity matrix for the Markov process on edge e, and λ is an associated rate parameter, and then letting λ tend to infinity.

From now on we will suppose that the original data C have been edited and consist just of *k parsimony sites* (sites where at least two states both occur in at least two sequences). Nonparsimony sites favor all trees equally so there is no loss in discarding them from a parsimony analysis. In order to make a fair comparison between this data and the null model, we will generate under the null model a collection C^* of the same number (k) of parsimony sites and so compare C and C^*. This *modified null model* is essentially the model described for just four taxa by Steel *et al.* (1993a). Under this model we are interested in the following tree-indexed distribution, S_T, defined by

$$S_T(i) := \text{Prob}[L(C^*, T) \leq i] \qquad i = 1, 2, 3, \ldots.$$

Thus $S_T(i)$ is the probability that a collection of k parsimony sites, randomly generated under the null model, has length at most i on tree T. From the distribution S_T, one can immediately derive its mean and variance, which we denote as μ_T and σ_T^2, respectively. In the case of four taxa, S_T is described by a cumulative binomial distribution, and μ_T and σ_T^2 are easily found (see Steel *et al.*, 1993a); however, for more than four taxa, the distribution is more involved.

In the Appendix we describe an efficient method to calculate the distribution S_T. For certain problems it may be just as convenient to approximate S_T, μ_T, or σ_T^2 by simulation, as we now describe. Note that these three quantities are the expected values, in our null model, of either a random vector (in the case of S_T) or a real-valued random variable (in the other two cases). Specifically, S_T is the expected value of a random 0–1 vector, which has a 1 in the i-th coordinate whenever $L(C^*, T) \leq i$ and is 0 otherwise, while μ_T is the expected value of $L(C^*, T)$, and σ_T^2 is the expected value of

$$\sum_{i=1}^{k} \left(L(C_i^*, T) - \frac{1}{k} \sum_{j=1}^{k} L(C_j^*, T) \right)^2,$$

where C_j^* is the j-th site of C^*.

Thus, to estimate S_T, μ_T or σ_T^2 by simulation, it is sufficient to describe how to estimate the expected value of a random variable (or vector) W under the modified null model. We present two approaches, which generally give similar (although not identical) values.

(1) Simulating the Modified Null Model

A series of sites is randomly and independently generated according to the null model; thus, for each site, the states are assigned randomly and independently to all positions with the probability that state α is assigned to position i being set equal to the proportion of sites in the original sequence i that are in state α. Any nonparsimony site that arises is immediately discarded, until k parsimony sites have been generated. The value of W is then calculated. This process is repeated a large number (say, 500) of times, and then the values of W obtained are averaged to give an estimate of $E[W]$, the expected value of W.

(2) Row Randomization

Given the original data, each sequence is randomized, that is, the nucleotides are randomly reordered

(thereby preserving the nucleotide frequencies within each sequence). This is done independently for each sequence. The number of parsimony sites in the resulting (row randomized) data set is then counted (let us call it x) and the value of W is then calculated and multiplied by k/x. This process is repeated a large number of times, and the values obtained are averaged to give an estimate of $E[W]$.

Note that a further refinement of our model, which we do not pursue here, would be to preserve in randomly generated data not just the frequencies of the states, but also the (sizes of the) partition structure of the data (we do this partially, but not completely, by comparing both the original and randomized data on just parsimony sites). To explain this further, note that if the random data consists of r-state sequences, then each site in the original data partitions the sequences into at most r subsets, or "blocks" (according to each sequence's associated state at that site) and the sizes of these blocks influence the parsimony score. For instance, a site that is randomly generated under the null model will tend to have a higher parsimony score on a fixed tree T if it has approximately equal numbers of states than if there is a predominance of one state over all the others. Thus, it might be reasonable to compare the original data only with randomizations that preserve not just the frequencies of states within sequences, but also the sizes of the induced blocks within sites. In principle the simulation Eq. (1) described above can be extended to estimate the analogues of $S_T(i)$ but in practice it would be necessary to generate and discard a large number of randomly generated sites in order to match up the frequencies of the block sizes with the original data.

APPLICATIONS

Significance of a Tree

We discuss two ways to determine the relative support for different trees. The first method is more in line with the approach adopted in Steel *et al.* (1993a), while the latter is a generalization of the approach of Kishino and Hasegawa (1989). Both approaches are based on the modified null model and its associated distributions S_T described in the Introduction, although the second approach requires knowledge only of μ_T.

Extended FD test. There are two special values i can take in $S_T(i)$, namely, when i is the length of the original data C on T and when i is the length of a maximum parsimony tree for C. We denote these two measures $s_1(C, T)$ and $s_2(C, T)$, respectively. That is,

$$s_1(C, T) := S_T(L(C, T)),$$
$$s_2(C, T) := S_T(L(C)),$$

where

$$L(C) = \min_{T'} \{L(C, T')\}.$$

Thus,

$$s_1(C, T) = \text{Prob}[L(C^*, T) \leq L(C, T)];$$
$$s_2(C, T) = \text{Prob}[L(C^*, T) \leq L(C)].$$

A third index is $s_3(C, T) = \Phi[(L(C, T) - \mu_T)/\sigma_T]$, where $\Phi(x)$ is the cumulative standard normal distribution.

We call these closely related measures *significance indices* for T. A modification of the parsimony principle is to rank trees not by their $L(C, T)$ value but by their s_1 values. The motivation for this is as follows: if a tree T has a smaller s_1 value than another tree T', then it would be less likely, by chance, that the data would fit T as well as they do (as measured by their parsimony length on T) than to fit T' as well as they do. We leave it to the practitioner to decide, depending on the context, which of s_1, s_2, or s_3 to use. Note that s_2 cannot be used unless the length of the maximum parsimony tree is known.

Note that for two-state sequences, and with equal frequencies of the two states in each sequence, this modified parsimony principle is identical to the parsimony principle described in the Introduction (this follows from Theorem 7.1 of Steel, 1993). Perhaps surprisingly, this result does not extend exactly to r-state sequences for $r \geq 4$. That is, for four-state sequences, and a data set C with equal frequencies of the four nucleotides, the trees T which minimize $L(C, T)$ are not necessarily the trees which minimize the $s_i(C, T)$.

The significance indices can be estimated by the simulation techniques described in the Introduction, although in certain cases (particularly for larger numbers of taxa) this may be difficult, since the indices will be very small. In such cases it would be necessary to apply the algorithm described in the Appendix.

Modified Kishino–Hasegawa sites test. An alternative approach to test the relative support for two trees is to generalize the method described by Kishino and Hasegawa (1989). These authors described a simple test for determining whether a tree T_1 has a significantly lower parsimony length than a second tree T_2, as follows: Given k parsimony sites, let X_i, and Y_i denote the parsimony length of the i-th site on T_1 and T_2, respectively. Let $Z_i = Y_i - X_i$ and $Z = Z_1 + , \ldots, + Z_k$. If the sites in the sequences evolve *i.i.d.*, then Z will be distributed (approximately) normally, with variance estimated by

$$V(Z) = \frac{k}{k-1} \sum_{i=1}^{k} \left(Z_i - \frac{1}{k} \sum_{j=1}^{k} Z_j \right)^2. \qquad (1)$$

Thus, to test whether T_1 has a significantly lower parsimony length than T_2 (at a significance level α) one simply checks whether $Z/\sqrt{V(Z)} > \Psi_\alpha$, where Ψ_α is the value beyond which the standard normal density curve has area α.

This approach, however, does not take account of variation in nucleotide frequencies, and so, for random data generated under a (frequency-biased) modified null model, the most parsimonious tree will generally be significantly favored over any other tree, given sufficiently long sequences, even though there is no information in the sequences beyond the variation of nucleotide frequencies.

Consequently we suggest the following modification of the Kishino and Hasegawa sites test. Given the k parsimony sites, repeat the above calculations for Z, except replace X_i and Y_i by $X_i - \mu_T$ and $Y_i - \mu_{T_2}$, respectively, to obtain a value

$$Z^* = Z - k(\mu_{T_2} - \mu_{T_1})$$

(here μ_{T_i} is as defined in Section 1—the expected length of T_i on a single site generated under the modified null model).

Then, under the modified null model, Z^* has expectation 0, and under the same *i.i.d.* hypothesis as before Z^* will be approximately normally distributed. The unbiased estimator for the variance of Z^* is estimated by $V(Z)$, given by Eq. (1) (since we have just added a constant to Z to obtain Z^*). Thus one would reject T_2 in favor of T_1 (at a significance level α) depending on whether $Z^*/\sqrt{V(Z)} > \Psi_\alpha$ (where Ψ_α is as before).

Thus we need to calculate $k(\mu_{T_2} - \mu_{T_1})$. This can be achieved analytically from S_T (given in the Appendix); however, it can also be conveniently estimated by simulation using either of the approaches in the Introduction, taking the random variable W to be the difference in the length of T_1 and T_2 for each set of k parsimony sites generated under the associated modified null model.

Application. We demonstrate an application of this modified Kishino–Hasegawa test in the context of the controversy on chloroplast origins (Lockhart *et al.*, 1992). Figure 3 shows the (unique) maximum parsimony tree T_1 (obtained by a branch-and-bound algorithm) which has parsimony length 251. Other trees tested include T_2–T_6 which group taxa with similar light harvesting complexes (unlike T_1). These trees require between 25 and 33 more mutations to fit the data than does T_1; their parsimony length is shown as the first number in the triple beside each tree. The above formulae were used to calculate the tree length differences (and its standard deviation) between T_1 and the other trees. The normalized values $Z/\sqrt{V(Z)}$ are shown as the second entry in the ordered triple beside trees T_2–T_5 in Fig. 3. In all cases the unmodified sites

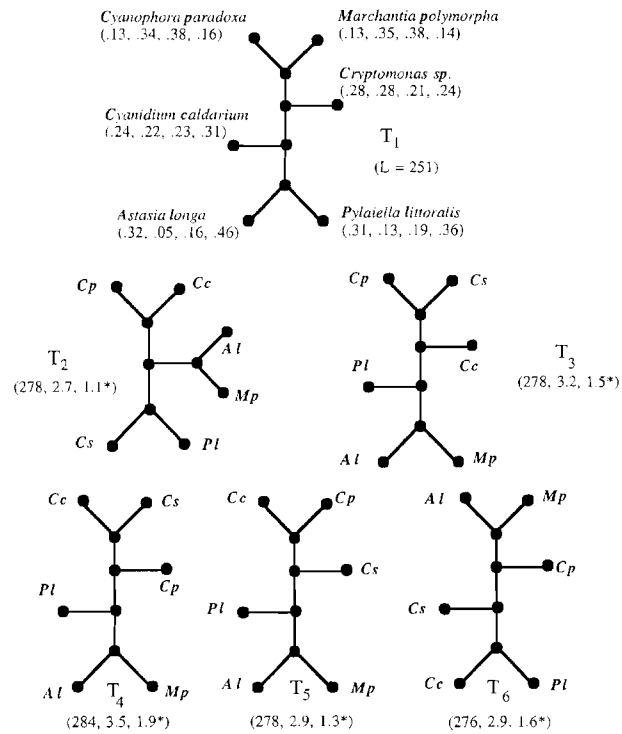

FIG. 3. Six trees used to test the evolutionary relationships between photosynthetic organelles. T_1 is the maximum parsimony tree requiring the smallest number of mutations (251) to fit the observed data. Trees T_2–T_6 group sequences from organelles having similar light harvesting complexes; their corresponding parsimony length L and values for $Z/\sqrt{V(Z)}$ are shown as the first two entries in the triple (L, z, z^*) beside each tree. For trees T_2–T_5, the third entry, z^*, is $Z^*/\sqrt{V(Z)}$, the corrected (normalized) tree length differences of T_1 versus T_2–T_5. Tree T_6 differs from T_1 only in the placement of the edge leading to *Astasia longa* and a corrected (normalized) tree length difference (the z^* value) between these two trees has been calculated by randomizing only the *A. longa* sequence (rather than all the organelle sequences as for T_2–T_5). As indicated by the z and z^* values, only when the data are not corrected for base composition effects is T_1 significantly better than the other trees at the 0.01 significance level (under the Kishino–Hasegawa sites test). The 16S rDNA sequences are from the RDP database (Olsen, 1991).

test supports T_1 at the 1% significance level. If the tree lengths are corrected for the effects of irregular a, g, c, t contents via the row randomization approach, then T_1 is still shorter than these other trees, although not at a 1% significance level—the $Z^*/\sqrt{V(Z)}$ for trees T_1–T_5 (not T_6) are shown as the third (starred) entry in the triple. Thus, much of the apparent support under parsimony for tree T_1 appears to be due simply to base frequency variations.

Nevertheless, the randomization test described suggests that there is still some nonrandom information in the sequences which parsimony is extracting from the sequences. However, we hesitate to suggest that this residual (nonrandom) information can be simply interpreted to indicate the true historical relationships. The reason for this is suggested by a further

comparison in Fig. 3. The tree T_6 (of length 276) differs from T_1 in the placement of the edge leading to *Astasia longa*. This grouping of *A. longa* with *Marchantia polymorpha* is supported on the basis of membrane ultrastructure, pigment composition and comparative analysis of protein encoding sequences (Morden *et al.*, 1992; Lockhart *et al.*, 1994). To test this tree against tree T_1 the modified sites test can be implemented so that only one sequence is randomized (in comparing trees T_1 with T_2–T_5 all sequences were randomized). In practice this might be done for the sequence whose adjoining edge is suspected of being misplaced and when other relationships in the tree are not being disputed or investigated. However, when this is done we find that the value analogous to $Z^*/\sqrt{V(Z)}$ is 1.6 (the third starred entry in the triple under T_6 in Fig. 3). That is, the grouping of the chromophyte with the euglenophyte is still mildly favored over the grouping of the euglenophyte and the bryophyte. These results suggest that even after subtracting the effects of base composition in this way there are still misleading patterns favoring an incorrect grouping. The reason for this is not clear. However, it may be due to site saturation occurring at different base frequencies when the sites free to vary in the sequences are changing under a covarion mechanism (Fitch and Markowitz, 1970). Alternatively, if there were unequal rates of change in the different lineages then this might also contribute to the misleading patterns observed (Felsenstein, 1978).

Significance of the Data

We now describe a further index to test whether the sequences are "tree-like"—that is, whether they have nonrandom structure (as detected by parsimony) above that which is due solely to the variation in the frequencies of the states (bases). As before, let $L(C)$ denote the length of a maximum parsimony tree T_0 for k parsimony sites C and consider the following value: let $R(C)$ denote the probability that the length of the minimal length tree T for data randomly generated according to the modified null model (using frequencies based on those of C) is no more than $L(C)$. That is,

$$R(C) := \text{Prob}[L(C^*) \leq L(C)].$$

Note that the tree T in this last sentence is not a fixed tree—it is the minimal length tree for the randomized data. Then, for data C^* randomly generated by the modified null model, we have the following, almost-trivial, but useful identity:

$$\text{Prob}[R(C^*) \leq \alpha] \leq \alpha.$$

This equation gives a convenient test of the modified null model (as would be induced, for instance, by the nonhierarchical star tree in Fig. 2) against the alternative hypothesis that the data contain hierarchical tree-like information (detected by the existence of a maximum parsimony tree with a significantly small $L(C, T)$ value). For instance, for data generated by the modified null model (so putting $C = C^*$ in $R(C)$) it is highly unlikely that $R(C)$ will take a value of 10^{-5} or less, so that if the observed value for $R(C)$ is this small it then suggests that there is some significant tree information in C (of course this does not imply that a tree favored by the tests in Section 1 is necessarily the historically correct tree).

We call $R(C)$ the *tree significance* of C. It is related to, but different from, indices suggested by Archie (1989), Faith (1990), and Steel *et al.* (1992). The essential difference is that our indices randomize the data within each sequence while the other three closely related measures randomize the data within each site (a further, but unimportant difference, is that our randomizations can be carried out using a probability model, rather than a deterministic regime). Thus, Archie's index and Faith's PTP index measure the proportion of simultaneous within-site randomizations which lead to a shorter maximum parsimony tree than the data itself.

The problem with these earlier approaches is that "random" data, produced by the modified null model described above, and containing no phylogenetic information, can produce highly significant values for such indices (an example of this spurious significance was described by Steel *et al.* 1993a). This is because, when the modified null model sequences vary considerably in their nucleotide frequencies, the data will favor a tree reflecting that variation, and this tree will therefore tend to be longer under nearly all within-site randomization—such randomizations tend to average out the nucleotide variation between the sequences. Note that if $R(C)$ is not significantly small, this does not necessarily imply the absence of tree-like information in the data. For instance, there may be some phylogenetic information in the base frequency variation, but the lack of a significant $R(C)$ value suggests that the only support in the sequences for a tree arises from this variation.

When the value of $R(C)$ is very small (say, 10^{-4}) it will be difficult to determine its value accurately by simulation, since a large number of runs would be needed to find instances of shorter trees on random data. This is not a problem, however, since in these cases, it suffices simply to estimate an upperbound on $R(C)$ which is significant at the prescribed level. For example, with the data described in the legend to Fig. 3, among several runs of 100 randomizations, no tree was found which was shorter than tree T_1 was on the original data. This suggests $R(C) < 0.01$. Thus, although sequence frequencies may be misleading parsimony in favoring T_1 over T_2, it nevertheless appears that parsimony is extracting some tree information from the sequences.

Note that a weak upperbound for $R(C)$ is provided by the significance index s_2, since we have the following relationships:

$$\begin{aligned} R(C) &= \text{Prob}[L(C^*) \leq L(C)] \\ &= \text{Prob}\left[\bigcup_{T \in B(n)} \{L(C^*, T) \leq L(C)\}\right] \\ &\leq \sum_{T \in B(n)} \text{Prob}[L(C^*, T) \leq L(C)] \\ &= \sum_{T \in B(n)} s_2(C, T), \end{aligned}$$

where $B(n)$ is the set of all binary trees on n leaves. Thus, we see that $R(C)$ is always bound above by $\Sigma_{T \in B(n)}\, s_2(C, T)$. A further analytical bound for $R(C)$ is described in Steel *et al.* (1994).

APPENDIX

We describe how $S_T(i)$ can be calculated by an algorithm which grows slowly (i.e., in polynomial time) with n (the number of taxa) and k (the number of parsimony sites). This in turn allows the significance indices $s_j(C, T)$ to be efficiently found. The algorithm does grow exponentially with r (the number of sequence states); however, this is not a problem for DNA/RNA sequences, where $r = 4$.

As before, denote the frequency of the state α in the i-th sequence by π_i^α and let π denote the matrix of π_i^α values. Randomly generate a second data set consisting of n sequences of length c by assigning states to sites independently and randomly, but so that for each site in sequence i, π_i^α is the probability that state α is assigned that site. Not all of the sites in the resulting randomly generated collection of sequences will be parsimony sites (sites for which at least two states appear at least twice). We edit out nonparsimony sites in our modified null model, but we first discuss the (ordinary) null model in which all sites are allowed.

If χ is randomly generated under the null model, let $P(T, \pi, j)$ be the probability that $L(\chi, T) = j$. We describe, from Steel *et al.* (1994, unpublished), a recursive formula for $P(T, \pi, j)$. Subdivide an edge of T to create a root vertex and consider, for each nonempty subset X of states, the joint probability that (i) $L(\chi, T) = j$ and (ii) X is the set of states that can be assigned to the root of T in at least one minimal extension of χ (to the vertices of T). We denote this joint probability as $P_X(T, \pi, j)$. Consider the ordinary generating function (a polynomial in x):

$$F_X(T, \pi) := \sum_j P_X(T, \pi, j) x^j.$$

Now if T has more than one leaf, let T_1 and T_2 be the two rooted subtrees of T, whose roots are adjacent to the root of T. Let $\pi^{(1)}$, $\pi^{(2)}$ be the restrictions of π to the leaves in T_1 and T_2, respectively. Applying the first pass of Fitch's algorithm [see Hartigan (1973)] we have

$$F_X(T, \pi) = \sum_{A, B} \delta(A, B, X) F_A(T_1, \pi^{(1)}) F_B(T_2, \pi^{(2)}), \quad (2)$$

where

$$\delta(A, B, X) = \begin{cases} 1 & \text{if } A \cap B = X; \\ x, & \text{if } A \cap B = \varnothing, A \cup B = X; \\ 0, & \text{otherwise.} \end{cases}$$

Note that if T has just one leaf, i, then

$$F_X(T, \pi) = \begin{cases} \pi_i^\alpha, & \text{if } X = \{\alpha\} \\ 0, & \text{if } |X| > 1. \end{cases} \quad (3)$$

From the recursion [Eq. (2)] and initial conditions [Eq. (3)] one can efficiently calculate the polynomials $\{F_X(T, \pi): X \neq \varnothing\}$ by starting from the leaves and working up to the root and storing all the intermediate polynomials generated in the construction. Then, $P(T, \pi, j)$ is simply the coefficient of x^j in $\Sigma_{X \neq \varnothing} F_X(T, \pi)$.

We wish to compute $s_T(i)$, the probability that k *parsimony* sites $\chi_1, \ldots, \chi_k$ randomly generated under the null model satisfy $\Sigma_{i=1}^k L(\chi_i, T) \leq i$. First we need to know the conditional probability that $L(\chi, T) = j$ given that χ is a parsimony site. Denote this as $P^*(T, \pi, j)$. Let $p(j)$ denote the probability of generating a nonparsimony site in which j states occur once, while a further state occurs at least once. Thus, letting

$$p = \sum_{j=0}^{r-1} p(j),$$

we see that $1 - p$ is the probability of generating a parsimony site. Thus,

$$P^*(T, \pi, j) = \frac{[P(T, \pi, j) - p(j)]}{(1 - p)}. \quad (4)$$

Next we explain how, $p(j)$ (and so p) can be calculated efficiently: Let $p_m(S, \alpha)$, $m > |S| + 1$, $\alpha \notin S$, denote the probability that, for the states assigned to sequences $1, \ldots, m$, the states in S are assigned exactly once, while α is assigned to the remaining (two or more) sequences. We have the recursions

$$p_m(S, \alpha) =$$

$$\begin{cases} p_{m-1}(S, \alpha)\pi_m^\alpha + \sum_{\beta \in S} p_{m-1}(S - \{\beta\}, \alpha)\pi_m^\beta, & \text{if } m > |S| + 2, \\ f_{m-1}(S \cup \{\alpha\})\pi_j^\alpha + \sum_{\beta \in S} p_{m-1}(S - \{\beta\}, \alpha)\pi_m^\beta, & \text{if } m = |S| + 2, \end{cases}$$

where, in the second case, $f_{m-1}(S \cup \{\alpha\})$ is the permanent of the $(|S| + 1) \times (|S| + 1)$ matrix $[\pi_i^\beta: i \leq m - 1, \beta \in S \cup \{\alpha\}]$.

Thus, $p_m(S, \alpha)$ can be calculated efficiently (in n) for all nonempty subsets S of states and all $m = 1, \ldots, n$ (we assume, for convenience, that $n > r$). Now, $p(j)$ is the sum of $p_n(S, \alpha)$ over all sets S of states of size j and all $\alpha \notin S$, and so $p(j)$ and hence p can be found in polynomial (in n) time. In this way, $P^*(T, \pi, j)$ can be calculated efficiently for all j by Eq. (4). Thus, letting $P^*(T, x)$ denote the associated ordinary generating function,

$$P^*(T, x) = \sum_j P^*(T, \pi, j)x^j,$$

we note that $S_T(i)$ is the sum, over $s = 0, 1, \ldots, i$, of the coefficients of x^s in $P^*(T, x)^k$, and this can be calculated in time which is polynomial in n and k, as claimed.

REFERENCES

Archie, J. (1989). A randomization test for phylogenetic information in systematic data. *Syst. Zool.* **38**(3): 239–252.

Carter, M., Hendy, M., Penny, D., Székely, L. A., and Wormald, N. C. (1990). On the distribution of lengths of evolutionary trees. *SIAM J. Disc. Math.* **3**(1): 38–47.

Charleston, M., and Steel, M. A. (1995). Five surprising properties of parsimoniously colored trees, *Bull. Math. Biol.* **57**(2): 367–375.

Faith, D. (1990). Chance marsupial relationships. *Nature* **345:** 393–394.

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

Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20:** 406–416.

Fitch, W. M., and Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixations of mutations in evolution. *Biochem. Genet.* **4:** 579–593.

Graham, R. L., and Foulds, L. R. (1982). Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosci.* **60:** 133–142.

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

Hasegawa, M., and Hashimoto, T. (1993). Ribosomal RNA trees misleading? *Nature* **361:** 23.

Hendy, M. D., and Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* **59:** 277–290.

Huelsenbeck, J. P., and Hillis, D. M. (1993). Success of phylogenetic methods in the four taxon case. *Syst. Biol.* **42:** 247–264.

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29:** 170–179.

Klenk, H. P., Palm, P., and Zillig, W. (1994). DNA-dependent RNA polymerases as phylogenetic marker molecules. *Syst. Appl. Microbiol.* **16:** 638–647.

Lake, J. A. (1987). A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Biol. Evol.* **4:** 167–191.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA* **91:** 1455–1459.

Lockhart, P. J., Penny, D., Hendy, M. D., Howe, C. J., Beanland, T. J., and Larkham, A. W. D. (1992). Controversy on chloroplast origins. *FEBS Lett.* **301:** 127–131.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11(4): 605–612.

Moon, J. W., and Steel, M. A. (1993). A limiting distribution for parsimoniously bicolored trees. *Appl. Math. Lett.* **6**(4): 5–8.

Morden, C. W., Delwiche, C. F., Kuhsel, M., and Palmer, J. D. (1992). Gene phylogenies and endosymbiotic origin of plastids. *J. Mol. Evol.* **28:** 75–90.

Olsen, G. J., Larsen, N., and Woese, C. R. (1991). The ribosomal RNA database project. *Nucleic Acids Res.* **19:** 2017–2018.

Olsen, G. J., and Woese, C. R. (1993). Ribosomal RNA: A key to phylogeny. *FASEB J.* **7:** 113–123.

Rodriguez, F., Oliver, J. L., Marin, A., and Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142:** 485–501.

Sidow, A., and Wilson, A. C. (1990). Compositional statistics: An improvement of evolutionary parsimony and its application to deep branches in the tree of life. *J. Mol. Evol.* **31:** 51–68.

Steel, M. A. (1993). Distributions on bicoloured binary trees arising from the principle of parsimony. *Discr. Appl. Math.* 41: 245–261.

Steel, M. A. (1994). Recovering a tree from the leaf colorations it generates under a Markov model. *Appl. Math. Lett.* **7:** 19–23.

Steel, M. A., Hendy, M. D., and Penny, D. (1992). Significance of the length of the shortest tree. *J. Classification* **9:** 71–90.

Steel, M. A., Lockhart, P. J., and Penny, D. (1993a). Confidence in evolutionary trees from biological sequence data. *Nature* **364:** 440–442.

Steel, M. A., Penny, D., and Hendy, M. (1993b). Parsimony can be consistent! *Syst. Biol.* **42:** 581–587.

Steel, M. A., Waterman, M., and Goldstein, L. (1995). A central limit theorem for parsimony length of trees. *Adv. Appl. Prob.*, submitted for publication.

Stewart, C. B. (1993). The power and pitfalls of parsimony. *Nature* **361:** 603–607.

Thompson, E. A. (1975). "Human Evolutionary Trees," Cambridge Univ. Press, Cambridge.