Methods for Inferring Phylogenies from Nucleic Acid Sequence Data by Using Maximum Likelihood and Linear Invariants

W. C. Navidi, G. A. Churchill, and A. von Haeseler
Department of Mathematics, University of Southern California

Likelihood methods and methods using invariants are procedures for inferring the evolutionary relationships among species through statistical analysis of nucleic acid sequences. A likelihood-ratio test may be used to determine the feasibility of any tree for which the maximum likelihood can be computed. The method of linear invariants described by Cavender, which includes Lake's method of evolutionary parsimony as a special case, is essentially a form of the likelihood-ratio method. In the case of a small number of species (four or five), these methods may be used to find a confidence set for the correct tree. An exact version of Lake's asymptotic \( \chi^2 \) test has been mentioned by Holmquist et al. Under very general assumptions, a one-sided exact test is appropriate, which greatly increases power.

Introduction

Likelihood methods have long been used in problems of phylogenetic inference, while methods based on linear invariants are of recent origin. Both methods are typically used to find a single tree which best fits some nucleic acid sequence data. We discuss some examples of this type of procedure. We also describe some confidence-set procedures, which are designed to yield collections of plausible trees rather than single estimates.

The use of maximum likelihood in phylogenetic inference is well established. Two papers giving excellent reviews of the subject are by Felsenstein (1983, 1988). Many models have been proposed for generating likelihood functions. Felsenstein (1981) proposed a model with one rate parameter per branch. Among the virtues of this model is that it is easy to maximize the likelihood for trees containing a fairly large number of species. We give below a numerical example using a version of this model. Cavender and Felsenstein (1987) and Barry and Hartigan (1987) describe a very general model with 12 parameters per branch. This enables maximum-likelihood methods to be used with relatively few assumptions. Most of the more restrictive models in common use can be obtained by placing constraints on this model. The general model discussed below is essentially this one, and the likelihood function we use is essentially the one described by Barry and Hartigan (1987) in their “maximum average likelihood” approach.

The first use of linear invariants was due to Lake (1987), who devised a procedure known as “evolutionary parsimony.” One appeal of linear-invariant methods is that the assumption that the evolutionary process is identical in every site is not needed. Instead, some constraints are put on the parameters of the model.

1. Key words: phylogenetic inference, invariants, likelihood, evolutionary parsimony.

Address for correspondence and reprints: William Navidi, Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113.

© 1991 by The University of Chicago. All rights reserved.
0737-4038/91/0801-0008$02.00
In the present paper, we will discuss two methods which use linear invariants. One is the method described by Cavender (1989), which relies on the asymptotic normality of the data. We will refer to this procedure as the asymptotic method of linear invariants. This method will be shown to contain Lake's (1987) method of evolutionary parsimony as a special case. We will show that Cavender's asymptotic method of linear invariants is essentially a maximum-likelihood technique. Therefore this method can be viewed as a way to base inferences on the likelihood function without assuming identical evolutionary processes in the various sites. We will also discuss an exact binomial version of evolutionary parsimony, similar to that found in the work of Holmquist et al. (1988). We will describe a one-sided test which is appropriate under very general assumptions.

The organization of the present paper is as follows: After describing the general model, we compare the features of the point-estimate type of data analysis with those of the confidence-set approach. Then we discuss in turn the likelihood-ratio method, the evolutionary-parsimony method, and the asymptotic method of linear invariants, describing in detail both point-estimate and confidence-set procedures. We then give a numerical example. Finally we summarize our results, comparing the asymptotic method of linear invariants to the likelihood-ratio method.

The General Model

We present the general model in the context of four-species unrooted trees (see fig. 1). It is easily extended to larger trees and to rooted trees. This model is also described by Navidi et al. (submitted). The species are labeled I-IV. Figure 1 shows one of the three possible trees, which we will refer to throughout as tree 1. Tree 2 is obtained from figure 1 by interchanging species II and III, and tree 3 is obtained by interchanging species II and IV.

We assume we have on hand four aligned nucleic acid sequences. Restrict attention to a single site on the molecule. Since the tree is unrooted, the root could be in any branch. Thus, as we travel along a branch, we cannot determine whether we are moving forward or backward in time. We therefore arbitrarily choose one of the interior nodes of the tree (a node labeled with an asterisk) as the initial node for the tree and specify a direction (indicated by arrows) for each branch (see figure 1). We ignore any effects due to insertions and deletions and assume that all variation in the sequences arises from transitions and transversions. We also ignore any effects due to the alignment process. We make the following assumptions:

---

**Fig. 1.—Four-species unrooted tree**
Assumption 1. The evolutionary processes in the various sites act independently of one another.

Assumption 2. All the sites evolve according to the same tree.

Assumption 3. Given the base at a site at an internal node, the two collections of bases formed by removing that site are distributed independently of each other.

Assumption 3 is a weaker form of the Markov hypothesis, which says that, given the present state of a site, the future evolution is independent of the past. It is similar to assumption 2 of Barry and Hartigan (1987, p. 200) and to assumption 2 of Felsenstein (1983, p. 248).

Let \( r = [r(A), r(G), r(C), r(U)] \), where \( r(A), r(G), r(C), \) and \( r(U) \), are the probabilities of observing A, G, C, U, respectively, at the initial node. For each branch \( i \), define the substitution probability \( m_i(A,G) \) to be the conditional probability of observing a G at the ending node of the branch, given that an A was present at the beginning node. For each of the other 15 pairs of bases A, G, C, and U, make a similar definition. The Markov matrix \( M_i \) for the \( i \)th branch is given by

\[
\begin{bmatrix}
  m_i(A,A) & m_i(A,G) & m_i(A,C) & m_i(A,U) \\
  m_i(G,A) & m_i(G,G) & m_i(G,C) & m_i(G,U) \\
  m_i(C,A) & m_i(C,G) & m_i(C,C) & m_i(C,U) \\
  m_i(U,A) & m_i(U,G) & m_i(U,C) & m_i(U,U)
\end{bmatrix}
\]

The elements of \( M_i \) are numbers between 0 and 1, and the row sums are equal to 1. This means that 12 parameters must be specified before the matrix is known. Since the components of \( r \) also sum to 1, a total of 63 parameters (5 branches \( \times \) 12 parameters/branch + 3 parameters at the initial node) are needed to determine the model. Additional constraints may be imposed on the parameters if desired. This will be necessary for linear invariants to exist. We require that any constraints be the same at all sites.

Denote the base at the beginning node of branch 5 by \( b_5 \) and denote the base at the end node of branch 5 by \( b_6 \). Then the probability of observing bases \( b_1, b_2, b_3, \) and \( b_4 \) at nodes 1, 2, 3, and 4, respectively, is

\[
\sum_{b_5} \sum_{b_6} r(b_5)m_5(b_5,b_6)m_3(b_6,b_3)m_2(b_3,b_4)m_1(b_4,b_1)m_2(b_5,b_2). \tag{2}
\]

This model describes the evolutionary process at a single site. If the values of the parameters are the same in all sites, we say that the sites are identically distributed.

Let \( N \) denote the number of sites. At each site, each species I-IV contributes an A, G, C, or U. Thus, each site yields an ordered quartet of bases which falls into one of 256 categories: AAAA, \ldots, UUUU. The position in the quartet depends on the species; for example, the quartet AGCU means that species I-IV have bases A, G, C, and U, respectively, at that site. The probability that a given site will be occupied by a given quartet is determined both by a topology of the tree and by the values of the parameters in \( r \) and the \( M_i \). Number the 256 quartets from 1 to 256. Denote by \( p_k \) the 256-dimensional column vector whose \( j \)th component, \( p_{kj} \), is the probability that site \( k \) will be occupied by quartet \( j \). Equation (2) shows how to calculate \( p_k \) in terms of \( r \) and the \( M_i \). Denote by \( \bar{p} \) the average \( \frac{1}{N} \sum_{k=1}^{N} p_k \). Each component of \( \bar{p} \) is in some
sense the “average” probability that a site will be occupied by a certain quartet. If the sites are identically distributed, then $p_k$ will be the same for each site $k$, and $\tilde{p}$ will also take on this common value. Denote by $X$ the 256-dimensional vector whose $j$th component is the number of sites occupied by quartet $j$. Then $E(X) = N\tilde{p}$.

Restrict attention to a particular tree $\tau$ and to a particular site $k$. Equation (2) generates all possible values of the components of $p_k$ as the elements of $\tau$ and the $M_i$ run through their allowable values. Denote by $P^*_i$ the set of all possible values of $p_k$. Since any constraints on the parameters are the same at all sites, the set $P^*_i$ is the same at each site. Denote by $P^*_\tau$ the set of all possible values of $\tilde{p} = E(X)/N$. If the sites are identically distributed, $p_k = \tilde{p}$, so $P^*_\tau = P^*_i$. Otherwise $P^*_\tau$ is the set of all averages on $N$ vectors chosen from $P^*_i$, so $P^*_\tau$ is a larger set.

Substitution Probabilities versus Substitution Rates

In the model described above, the parameters represent substitution probabilities. Many important aspects of molecular evolution are more appropriately studied with models involving substitution rates. For example, models which are designed to estimate the expected number of substitutions occurring at a site during a given time are more naturally based on substitution rates. Substitution probabilities are less desirable for this task because they do not directly give information about the number of substitutions that may have created an observed difference. Models involving substitution rates and elapsed times have been used to estimate times between speciation events. Substitution probabilities do not offer as direct an estimation procedure.

A matrix $Q$ of substitution rates, together with a length of time $t$, determine a matrix of substitution probabilities by $e^{Qt}$. In contrast, there exist matrices of substitution probabilities which are not generated by any rate matrix. Thus models involving substitution probabilities are somewhat more general.

Models based on substitution probabilities seem to be sufficient for the task of estimating phylogenies from aligned nucleic acid sequences. The probability distribution of the vector $X$ is determined by the substitution probabilities, through equation (2). Therefore statistical methods which use $X$ as the data, such as maximum-likelihood methods and invariant methods, can be based on models whose parameters represent substitution probabilities. Models based on substitution rates and on elapsed times can also work—but only because they determine the substitution probabilities. It seems easier to work directly with the substitution probabilities themselves.

Point Estimates and Confidence Sets

Data analysis for the purpose of estimating phylogenies may pursue one of two approaches. One approach is to find the single tree which best fits the data, to serve as the estimate of the correct phylogeny. An analysis of this type may be called a point estimate approach. Such an analysis may be modified by the inclusion of a rule that, if either no tree or more than one tree fits the data well, then no estimate will be made. A disadvantage of the point-estimate approach is that its error rate depends on the true values of the parameters. Invariably, there will be some values of the parameters for which the procedure is very accurate and others for which it is not. For this reason it is often difficult to determine the accuracy of the procedure in a specific instance.

An alternative approach is to classify each tree in a collection as being plausible or implausible, rather than to single out a best tree. A procedure of this type may be described as a confidence-set approach. In this approach, for each tree $\tau$, a statistical test is made of the hypothesis that $\tau$ is the correct tree. Those trees not rejected at level $\alpha$ are deemed plausible. If the initial collection of trees is known to contain the
correct one, the set of plausible trees is a $1 - \alpha$ confidence set for the correct tree. In contrast to the point-estimate approach, the error rate for the confidence-set procedure is always known. The proportion of confidence sets which fail to contain the correct tree is $\alpha$.

In practice, point-estimate procedures and confidence-set procedures require similar calculations, so it is reasonable to do both. Both likelihood techniques and linear-invariant methods, including evolutionary parsimony, can be used to perform analyses of either type.

**The Likelihood-Ratio Method**

The use of maximum-likelihood techniques in phylogenetic inference is now common. For example, Felsenstein (1981) proposed a model in which the evolutionary process in each branch of the tree is governed by a single rate parameter. Other likelihood methods have been proposed since; among the most flexible is Barry and Hartigan's (1987) "maximum average likelihood" approach. These methods provide techniques to compute the maximum of the likelihood function for a given tree. The tree whose maximum likelihood is the largest is a point estimate of the correct tree. It has not been clear how to use the likelihood function to find a confidence set. This requires a hypothesis test to determine whether a given tree is compatible with the data. The likelihood-ratio method, described below, yields such a test. It provides a measure of the statistical significance of the maximum value of the likelihood function for any tree. Thus, given the maximum likelihood for each of several competing trees, each tree can be tested in turn and classified as being either plausible or implausible. When the number of species is small enough so that every tree can be tested, the plausible trees form a $1 - \alpha$ confidence set.

We now describe the likelihood-ratio method in terms of the general model given above. A description from a somewhat more mathematical point of view is given by Navidi et al. (submitted). Although the discussion is given in the context of four-species trees, the method can in principle be extended to trees of any size. In addition to assumptions 1–3, we impose two additional assumptions on the model.

**Assumption 4.** The sites are identically distributed.

**Assumption 5.** The values of the parameters in $\mathbf{r}$ and in the $\mathbf{M}_i$ are equal neither to 0 nor to 1.

Assumption 5 states that the true values of the parameters lie in the interior of the parameter space. It is necessary in order to ensure that the likelihood-ratio statistic has the correct asymptotic distribution. It is harmless to make this assumption, since in practice the behavior of a model when some parameters lie on the boundary is indistinguishable from its behavior when those same parameters are very near but not on the boundary.

Fix a tree, $\tau$, to be tested. As before, let $\mathbf{X}$ refer to the vector of counts of the 256 quartets, let $\mathbf{\bar{p}} = E(\mathbf{X})/N$ be the vector of expected frequencies, and denote the $i$th coordinate of $\mathbf{\bar{p}}$ by $\bar{p}_i$. Recall that $P_{\tau}$ is the set of all possible values of $\mathbf{\bar{p}}$ when $\tau$ is the correct tree. We perform a test of the null hypothesis

$$H_0: \mathbf{\bar{p}} \in P_{\tau}. \quad (3)$$

Since the sites are identically distributed, $\mathbf{X}$ has the multinomial distribution with
parameters $N$ and $p$. The likelihood function is the density of $X$ considered as a function of the parameter $p$.

$$l(\hat{p}) = \frac{N!}{\prod_{i=1}^{256} \hat{p}_i X_i}.$$  \hspace{1cm} (4)

More useful is the log likelihood, the logarithm of $l$ when the constant $N!/\prod_{i=1}^{N} X_i!$ is ignored:

$$L(\hat{p}) = \sum_{i=1}^{256} X_i \log \hat{p}_i.$$  \hspace{1cm} (5)

The maximum-likelihood estimator (MLE) of $\vec{p}$ is the vector $\hat{p} \in P$, maximizing $L$. We assume such a vector exists. Methods for computing the MLE are based on the fact that all the vectors in $P^+$ (= $P_r$) are generated by equation (2) as the parameters in $r$ and in the M, run through their allowable values. See Barry and Hartigan (1987) for an algorithm to compute $\hat{p}_r$ for the full model in which each matrix has 12 free parameters. Felsenstein (1981) describes an algorithm for the case in which each matrix is determined by a single rate parameter. The value $L(\hat{p}_r)$ is called the constrained maximum likelihood. It is the maximum value taken by the log likelihood by using values of the parameters allowable under the constraints of the model.

Previous efforts to use maximum likelihood to test the hypothesis that a tree $\tau$ is correct have involved expressing $\tau$ as a subcase of a more general tree $\tau'$, where $\tau'$ is known to be correct. The tree $\tau$ can be obtained from $\tau'$ by placing some constraints on parameters, e.g., setting their values equal to zero. See the work of Felsenstein (1988) for a discussion of this idea. The principle involved is that, if the subtree $\tau$ is in fact correct, then the statistic $2[L(\hat{p}_r) - L(\hat{p}_r)]$ has, asymptotically, a $\chi^2$ distribution with degrees of freedom equal to the number of parameter constraints.

This principle is more generally applicable. It is not necessary to express $\tau$ as a subcase of a correct tree. We need only express $\tau$ as a subcase of a model for which it is known that the correct tree is also a subcase. The most general model, of which every tree is a subcase, is the model putting no constraints at all on $\vec{p}$, except that its components sum to 1. This is known as the unconstrained or saturated model.

With no constraints on $\vec{p}$, the MLE of $\vec{p}$ is $X/N$. The unconstrained maximum likelihood is $L(X/N) = \sum_{i=1}^{256} X_i \log X_i/N$, where $X_i \log X_i/N$ is taken to be 0 if $X_i = 0$. The quantity

$$\lambda_r = 2[L(X/N) - L(\hat{p}_r)]$$  \hspace{1cm} (6)

is the generalized likelihood-ratio statistic for testing $H_0: \vec{p} \in P_r$.

Define $d$ as 255 – the number of free parameters in the model.

Under assumptions 1–5, if $\hat{p}_r$ is a local maximum of $L$, then the asymptotic null distribution of $\lambda_r$ is $\chi^2_d$ (see Rao 1973, pp. 418–419).

When more than five or six species are to be considered simultaneously, there are too many possible trees to test each one. In such cases, the likelihood-ratio statistic can be used in conjunction with various search techniques to find plausible trees for larger numbers of species. Several useful such techniques have been proposed. For example, Felsenstein (1981) describes a search strategy which begins with a two-species
tree, then adds species one by one, placing new species on the branches which yield
the greatest likelihood. His discussion is given in the context of a model in which the
matrix for each branch of the tree is determined by a single rate parameter. Barry and
Hartigan (1987) describe an application of this technique to the fully parameterized
model. Given a tree or set of trees which have been found by these or other search
procedures, the likelihood-ratio statistic can be computed to test the plausibility of
each one.

We now discuss the behavior of the procedure when the assumption of identically
distributed sites is dropped. Recall that \( p_k \) is the vector of probabilities associated with
the \( k \)th site and that the collection of possible values of \( p_k \), denoted \( P^*_k \), is the same
for each site \( k \). The vectors \( p_k \in P^*_k \) can be expressed in terms of the \( \leq 63 \) free parameters
in \( r \) and the \( M_k \), through equation (2). When the sites are not identically distributed,
\( P \) is a larger set than \( P^*_k \), and the vectors \( \hat{p} \in P \) are not expressible in this way. We
can still compute \( \lambda \), with equation (6), using the quantity \( \hat{p} \in P \) which maximizes
the likelihood. Since we will then be maximizing over too small a set, \( L(\hat{p}) \) will tend
to be less than the maximum of \( L \) taken over \( P \), so \( \lambda \) will tend to be larger under
the null hypothesis than under the \( \chi^2 \) distribution given above. Thus we will be more
likely to reject the null hypothesis than the presumed level of the test would indicate.

**Evolutionary Parsimony**

The method of evolutionary parsimony, developed by Lake (1987), is a procedure
applicable to four-species unrooted trees (see fig. 1). Corresponding to each tree is a
pair of 256-dimensional column vectors called *linear invariants*. Each tree shares an
invariant with each other, so there are three linear invariants in all. The correct tree
has the property that both its invariants are orthogonal to \( \hat{p} \).

Lake (1987) used the term \"invariant\" to refer to the test statistics generated by
his method. Cavender (1989) used the term in a slightly different way. The following
deinition, which we will use, is equivalent to Cavender's:

**Definition 1.** A vector is a linear invariant for the tree \( \tau \) if it is orthogonal to every
vector in \( P_\tau \).

For Lake's method to work, we need assumptions 1–3. We also need some para-
meter constraints. No restrictions need be put on \( r \) or on the matrix for the middle
branch; whereas the following assumption is needed for the outer branches:

**Assumption 6.** The matrices for the outer branches are of the following form:

\[
\begin{bmatrix}
A & G & C & U \\
A & e & f & g \\
G & h & i & j \\
C & k & k & l \\
U & n & n & p & q \\
\end{bmatrix}
\]

(7)

This means that when a transversion occurs in an outer branch it is equally likely
to be either of the two possible ones. There are no other constraints. We do not need
assumption 4 or assumption 5. The method is valid regardless of whether the sites are
identically distributed.

Jin and Nei (1990) describe for the model a parameterization in which the Markov
matrices satisfy assumption 6 yet Lake’s method fails. Their matrices give substitution
probabilities over a small fraction (e.g., \( 1/50 \)) of a branch and are approximations of
Methods of Phylogenetic Inference

substitution-rate matrices. In this parameterization, additional conditions are needed, so that the collection of allowable matrices is closed under multiplication. As pointed out by Jin and Nei, sufficient conditions are given by Cavender (1989). These conditions are mentioned below. Assumption 6 is sufficient, however, when the matrices give substitution probabilities over the whole branch, as we have done here.

We now describe the three vectors which are the linear invariants: let \( v_1 \) be the 256-dimensional vector with a 1 in the positions corresponding to the quartets AACC, AAUU, GGCC, GGUU, CCAA, CCGG, UUAA, UUGG, AGCU, AGUC, GACU, GAUC, CUAG, CUGA, UCAA, UCGG, AUGC, AAUC, GGCU, GGUC, CCAG, CCGA, UUUAG, and UUGA; and 0 elsewhere.

Let \( v_2 \) be the 256-dimensional vector with a 1 in the positions corresponding to the quartets ACAC, AUUA, GCAC, UAAU, CACA, CGGC, UACG, ACGU, AUGC, GCAU, UGAC, CAGU, CUGA, UCAA, GACU, UCAG, UAGC, CGCU, GUGA, UCCG, AGCU, AUUG, GCGA, GACU, CUCG, CCAG, UUAG, and UUGA; and 0 elsewhere.

Let \( v_3 \) be the 256-dimensional vector with a 1 in the positions corresponding to the quartets ACCA, AUUA, GCCG, GUUG, CAAC, CGGC, UAAG, UGGU, ACUG, AUCA, GCGA, UAGC, CUGA, CGAU, UGAC, ACUG, CAAG, GAUG, CAGU, ACUG, UGAC, and UGAC; and 0 elsewhere.

Denote by \( v_i^T \) the transpose of \( v_i \), i.e., \( v_i \) considered as a row vector. Then \( v_i \) is orthogonal to \( \bar{p} \) if \( v_i^T \bar{p} = 0 \).

Lake (1987) showed that, if tree 1 is correct, then \( v_1^T \bar{p} = 0 \) and \( v_3^T \bar{p} = 0 \). Thus \( v_2 \) and \( v_3 \) are linear invariants for tree 1. Similarly, \( v_1 \) and \( v_3 \) are invariants for tree 2, and \( v_1 \) and \( v_2 \) are invariants for tree 3. For \( i = 1, 2, 3 \), let \( P_i \) be the sum of the components of \( X \) in those positions where \( v_i = 1 \), and let \( Q_i \) be the sum of the components of \( X \) in those positions where \( v_i = -1 \), so \( v_i^T X = P_i - Q_i \).

We now describe three point-estimate procedures and a confidence-set procedure using the setup above. The first point-estimate approach is due to Lake (1987). For \( i = 1, 2, 3 \), we test the hypothesis

\[
H_0: v_i^T \bar{p} = 0.
\]  

The hypothesis \( H_0 \) is true when tree \( i \) is incorrect and, when highly exceptional cases are disregarded, is false when tree \( i \) is correct.

The test statistic proposed by Lake is

\[
S_i = \frac{(P_i - Q_i)^2}{P_i + Q_i}.
\]  

This statistic can be justified as follows: The quantity \( P_i - Q_i = v_i^T X \), so \( E(P_i - Q_i) = E(v_i^T X) = N v_i^T \bar{p} = 0 \) under \( H_0 \). The variance of \( P_i - Q_i \) is \( N |v_i^T| \bar{p} \), where \( |v_i^T| \) denotes the vector whose components are the absolute values of the components of \( v_i^T \). We estimate \( N |v_i^T| \bar{p} \) by substituting \( X \) for \( N \bar{p} \), obtaining \( |v_i^T| X = P_i + Q_i \). When \( N \) is large, \( P_i - Q_i \) is approximately normally distributed, so, under \( H_0 \), \( S_i \) has an approximate \( \chi^2_1 \) distribution. Lake suggested computing the three quantities \( S_1, S_2, \)
and $S_3$ and declaring a tree $i$ to be correct if $S_i$ was significant at the 5% level and if $S_j$, for $j \neq i$, was not. Otherwise no tree would be preferred.

Holmquist et al. (1988) pointed out that, if tree $i$ is incorrect, then conditional on $P_i+Q_i$, $P_i$ has the binomial distribution with number of trials equal to $P_i+Q_i$, and success probability $\pi = \frac{1}{2}$. Therefore an exact binomial test can be made, avoiding the $\chi^2$ approximation. Of course, a test which has conditional level $\alpha$ for each value of $P_i+Q_i$ is an unconditional level-$\alpha$ test as well.

We now make an assumption which will ensure that $\chi > \frac{1}{2}$ when tree $i$ is correct.

Assumption 7. In each branch of the tree, the probability of no difference between the nodes at either end is greater than the probability of a transition difference.

Theorem 1. Under assumptions 1–3, 6, and 7, if tree $i$ is correct, then conditional on $P_i+Q_i$, $P_i$ has the binomial distribution with number of trials equal to $P_i+Q_i$, and success probability $\pi > \frac{1}{2}$. (The proof is in the Appendix.)

We now define our second point-estimate procedure for choosing a tree using Lake's invariants.

Procedure A:

A1. Choose a critical value $\alpha$ and compute $P_1$, $P_2$, and $P_3$.

A2. Compute the one-sided upper-tail significance level of $P_i$ by using the binomial distribution with parameters $P_i+Q_i$ and $\frac{1}{2}$.

A3. Declare a tree to be correct if its significance level is $<\alpha$ while the significance levels of the other two trees are $>\alpha$. Make no decision otherwise.

This is a one-sided exact version of Lake's (1987) procedure, which we have described above. Alternatively, we could simply choose the tree with the lowest significance level.

It has been suggested (Li 1989) that when Lake's method is used with significance level $\alpha$ the true significance level is $3\alpha$, since three tests are being made. The significance level measures the frequency with which a true null hypothesis is rejected. In the context of tree-selection procedures, such as evolutionary parsimony, another error rate seems to be more important. This is the frequency with which the procedure chooses an incorrect tree. This frequency depends on the true values of the parameters and is difficult to determine. In general, it need not be well estimated by $3\alpha$.

A procedure which yields confidence sets for the correct tree can be constructed by considering sums of pairs of the $P_i$. Define

$$Y_1 = P_2 + P_3; \quad Y_2 = P_1 + P_3; \quad Y_3 = P_1 + P_2$$

and

$$Z_1 = Q_2 + Q_3; \quad Z_2 = Q_1 + Q_3; \quad Z_3 = Q_1 + Q_2.$$  (11)

If tree $i$ is correct, then conditional on $Y_i+Z_i$, $Y_i$ has the binomial distribution with parameters $Y_i+Z_i$ and $\frac{1}{2}$. If tree $i$ is incorrect, then $Y_i$ is conditionally binomial with success probability $>\frac{1}{2}$. For each tree $i$ we may compute the one-sided upper-tail significance level of $Y_i$ by using the binomial distribution with parameters $Y_i+Z_i$ and
The collection of trees for which this significance level is < 0.05 is an exact level 1 - 0.05 confidence set for the correct tree.

A reviewer has pointed out that, if we define \( Y'_1 = P_2 + Q_3 \); \( Y'_2 = P_1 + Q_3 \); \( Y'_3 = P_1 + Q_2 \) and \( Z'_1 = Q_2 + P_3 \); \( Z'_2 = Q_1 + P_3 \); \( Z'_3 = Q_1 + P_2 \), then, if tree \( i \) is correct, it is the case that, conditional on \( Y'_1 + Z'_3 \), \( Y'_1 \) has the binomial distribution with parameters \( Y'_1 + Z'_3 \) and \( 0.05 \). If tree \( i \) is incorrect, then \( Y'_1 \) is conditionally binomial with success probability \( \neq 0.05 \). For each tree \( i \) we may compute the two-sided significance level of \( Y'_1 \) by using the binomial distribution with parameters \( Y'_1 + Z'_3 \) and \( 0.05 \).

We may combine the procedure above with the procedure based on equations (10) and (11) by performing each and then rejecting tree \( i \) if either procedure rejects it at level 0.025. Some preliminary calculations indicate that, in some situations, combining the two procedures may increase power by \( \sim 10\% \) over that of the procedure based on equations (10) and (11) alone. In other situations, the power of the combined procedure seems to be about the same.

The statistics \( Y_i, Z_i \) can be used to formulate another point-estimate procedure for choosing the correct tree, as described below:

Procedure B:

B1. Choose a critical value \( \alpha \) and compute \( Y_1, Y_2, \) and \( Y_3 \).

B2. Compute the one-sided upper-tail significance level of \( Y_i \) by using the binomial distribution with parameters \( Y_i + Z_i \) and \( \alpha \).

B3. Declare tree \( i \) to be correct if \( Y_i \) is not significant at level \( \alpha \) whereas \( Y_j \) for \( j \neq i \) is significant. Make no decision otherwise.

Again, an obvious modification is to choose the tree with the largest significance level. This procedure can also be combined with the two-sided procedure based on \( Y'_1 \) and \( Z'_1 \). Preliminary simulation results suggest that procedures A and B are about equally effective in selecting the correct tree.

We now describe another confidence-set procedure, valid when the number of sites is large, which can be extended to a more general setting. Fix in mind a tree \( i \) and let \( V_i \) be the 256 \( \times \) 2 matrix whose columns are the invariants for the tree \( i \). Thus \( V_i \) is the matrix whose columns are \( v_j \) and \( v_k \), where \( i \neq j \) and \( i \neq k \). Denote by \( V_i^T \) the transpose of \( V_i \), the 2 \( \times \) 256 matrix whose rows are the columns of \( V_i \). Then, if tree \( i \) is correct, \( V_i^T \tilde{p} = 0 \). We seek a test statistic for the hypothesis

\[
H_0: V_i^T \tilde{p} = 0 .
\] (12)
$T_i = X^T V_i (V_i^T \hat{\Sigma} V_i)^{-1} V_i^T X.$  

Under $H_0$, $T_i$ has an approximate $\chi^2$ distribution. In fact, $T_i = S_j + S_k$, the sum of two of Lake’s statistics. See equation (9). The level $1-\alpha$ confidence set is determined by including those trees $i$ for which $T_i$ is not significant at level $\alpha$.

This confidence procedure is itself not very useful, since the exact procedure given above is more accurate and easier to compute. What is of value is its method of construction, which will be used below to develop a very general procedure.

The Asymptotic Method of Linear Invariants

When appropriate constraints are placed on the matrix parameters, it is in principle possible to compute invariants for trees of any size. Cavender (1989) describes a method for calculating all linear invariants for four-species-rooted trees with six branches. In principle, his method can be generalized to larger trees, although the amount of computation becomes too great for trees with more than about five species. For linear invariants to exist, some constraints must be put on the parameters. For example, Cavender (1989) considers matrices of the form of matrix (7) with two additional constraints:

$$e + h = f + i, \text{ and } l + p = m = q.$$  

The set of all matrices of this form is closed under matrix multiplication (and is thus a semigroup). No strictly larger semigroup than this is useful, since then there are no linear invariants (Cavender, submitted). Models in which the probabilities of transversions are not balanced are also feasible (Cavender 1989).

Fix in mind a tree $\tau$ to be tested. We presume that appropriate parameter constraints are in place and that a number of linear invariants for $\tau$ have been calculated. Let $m$ represent this number. It is not necessary to calculate all of the linear invariants; any subset will do. However, each invariant adds 1 degree of freedom to the distribution of the test statistic, so more invariants should generally result in a more powerful test.

Let $V_i$ be the $256 \times m$ matrix whose columns are linear invariants for $\tau$. We construct a test statistic for testing the null hypothesis

$$H_0: V_i^T \hat{p} = 0.$$  

The hypothesis $H_0$ is true if $\tau$ is the correct tree. The test statistic will be of the form of equation (13). Its construction is analogous to the one carried out at the end of the previous section. When the number of sites is large, the quantity $V_i^T X$ is distributed approximately $m$-variate normal, with mean $E(V_i^T X) = NV_i^T \hat{p} = 0$ under $H_0$. As before, let $\Sigma$ denote the covariance matrix of $X$. We describe a procedure to estimate $V_i^T \Sigma V_i$, the covariance matrix of $V_i^T X$. In this case, we will estimate $V_i^T \Sigma V_i$ by estimating $\Sigma$. Consider the case in which the sites are identically distributed, although the procedure works equally well when they are not. Denote the $ij$th element of $\Sigma$ by $\sigma_{ij}$. Then, if the sites are identically distributed,

$$\sigma_{ij} = -N \hat{p}_i \hat{p}_j \text{ if } i \neq j, \text{ and } \sigma_{ii} = N \hat{p}_i (1 - \hat{p}_i).$$  

The estimator $\hat{\Sigma}$ will be based on an estimator $\hat{p}$ of $\tilde{p}$. Any consistent estimator of $\hat{p}$
will work asymptotically. One possible choice is \( \hat{\rho} = X/N \), but we will use \( \hat{\rho} = X/N - V_1(V_1^T V_1)^{-1} V_1^T X/N \), subtracting from \( X/N \) its projection onto the space spanned by the columns of \( V_1 \). This should be more accurate, and it seems to produce less complicated expressions for the test statistic. This procedure will give Lake's test statistic (9) under the assumptions of evolutionary parsimony. Denote the \( ij \)th element of \( \Sigma \) by \( \hat{\sigma}_{ij} \). Now define \( \Sigma \) as follows:

\[
\hat{\sigma}_{ij} = -N \hat{\mu}_i \hat{\mu}_j \text{ if } i \neq j, \quad \text{and } \hat{\sigma}_{ii} = N \hat{\mu}_i (1 - \hat{\mu}_i).
\]

The expression on the right hand side of equation (13) now yields a test statistic for the null hypothesis (15). We denote this test statistic by \( T \). It has been shown by Navidi et al. (submitted) that under assumptions 1-3 the asymptotic null distribution of \( T \) is \( \chi^2_n \), regardless of whether the sites are identically distributed. Now, given a collection of trees, we can compute \( T \) for each one. The trees for which \( T \) is not significant at level \( \alpha \) are plausible candidates for the correct tree.

In practice, with samples of moderate size, some problems may be encountered with the above procedure. Cavender (1989) reports getting negative components in \( \hat{\rho} \) with a sample size of 1,095 sites. He replaced the negative values with zeroes, which is probably the best one can do under the circumstances. This indicates that unless the number of sites is quite large the asymptotic approximation should not be trusted, and so the null distribution should be estimated with a simulation.

It turns out that there is a great similarity between the asymptotic method of linear invariants, which is described above, and the likelihood-ratio method. To see this, again fix in mind a tree \( \tau \), and let \( V_\tau \) be a matrix whose columns are linear invariants. Let \( V \) be the collection of all vectors \( w \) for which \( V^T w = 0 \). The null hypothesis (15) can now be written as

\[
H_0: \hat{\rho} \in V_\tau.
\]

It follows from definition 1 that \( P_\tau \subset V_\tau \). Comparing hypothesis (18) with hypothesis (3) shows that the asymptotic method of linear invariants differs from the likelihood-ratio method in that it tests whether \( \hat{\rho} \) belongs to the larger set \( V_\tau \) rather than to \( P_\tau \).

The next result states that this is essentially the only difference between the two methods.

**Theorem 2.** Let \( \tau \) denote the correct tree. Let \( \hat{\rho}_\tau \) be the vector in \( V_\tau \) maximizing \( L \). Let \( T \) be the test statistic for the asymptotic method of linear invariants. Then, under assumptions 1-3 and 5, as the number of sites grows large, the difference \( T - 2[L(X/N) - L(\hat{\rho}_\tau)] \to 0 \).

See Navidi et al. (submitted) for a proof. Theorem 2 shows that this method of linear invariants is asymptotically equivalent to the likelihood-ratio method in which the likelihood is maximized over \( V_\tau \) instead of over \( P_\tau \). Thus the asymptotic method of linear invariants is essentially a maximum-likelihood technique.

**A Numerical Example**

To provide an example of the methods discussed above, we extracted four aligned sequences from the collection of small-subunit RNAs (Dams et al. 1988). The sequences are I, *Sulfolobus solfataricus* (archaebacteria); II, *Halobacterium salinarium* (archaebacteria); III, *Escherichia coli* (eubacteria); and IV, *Homo sapiens* (eukaryota). This set of four aligned sequences contains 1,352 positions without insertions or dele-
We analyze the data first with evolutionary parsimony, then with the likelihood-ratio method.

To perform an evolutionary parsimony analysis, the following values were calculated:

\[ P_1 = 38, \quad P_2 = 21, \quad P_3 = 25, \quad Q_1 = 30, \quad Q_2 = 12, \quad Q_3 = 15 \]

and

\[ Y_1 = 46, \quad Y_2 = 63, \quad Y_3 = 59, \quad Z_1 = 27, \quad Z_2 = 45, \quad Z_3 = 42 \].

We compute a \( p \) value for each tree. First, under evolutionary parsimony procedure A, the \( p \) value for tree 1 is found by computing the probability that a quantity distributed binomially with number of trials equal to \( P_1 + Q_1 = 68 \) and with success probability \( \frac{1}{2} \) is greater than or equal to \( P_1 = 38 \). This \( p \) value is 0.1981. Similar calculations for trees 2 and 3 yield \( p \) values of 0.0814 and 0.0769, respectively. With these data, the sample size is large enough so that the normal approximation to the binomial would have been appropriate, although we give the exact \( p \) values. In procedure A small \( p \) values are characteristic of the correct tree. Thus evolutionary-parsimony procedure A slightly favors tree 3, which clusters I with IV and II with III. This is the tree proposed by Lake (1988). No tree is significant at the 5% level. The results for procedure B are similar. Here the \( p \) value for tree 1 is found by computing the probability that a quantity distributed binomially with number of trials equal to \( Y_1 + Z_1 = 73 \) and with success probability \( \frac{1}{2} \) is greater than or equal to \( Y_1 = 46 \). This \( p \) value is 0.0172. Similar calculations for trees 2 and 3 yield \( p \) values of 0.0507 and 0.0555, respectively. In this procedure small \( p \) values are characteristic of incorrect trees, so tree 3 is slightly favored. Tree 1 is rejected at the 5% level, and trees 2 and 3 form a 95% confidence set for the correct tree.

To implement the likelihood-ratio method, we use a model of Felsenstein (1981). There are no constraints on \( r \), while the evolutionary process in each branch is described with a single rate parameter. We can describe this model in terms of the general model of section 2 as follows: Denote the rate parameters \( t_1, t_2, t_3, t_4, \) and \( t_5 \). Then the elements of the matrices \( M_i \) are given by

\[ m_{ij}(k,k) = e^{-r_i} \delta_{jk} + (1 - e^{-r_i}) r(k). \] (20)

Here \( \delta_{jk} = 1 \) if \( j = k \) and is equal to 0 otherwise. There are a total of eight free parameters, so the asymptotic distribution of the likelihood-ratio statistic has \( 255 - 8 = 247 \) degrees of freedom. The unconstrained maximum likelihood is \( L(X/N) = -5591.057 \). The values of the parameters maximizing the likelihood for tree 1 are

\[ r_1 = 0.229, \quad r_2 = 0.309, \quad r_3 = 0.259, \quad r_4 = 0.203 \]

and

\[ t_1 = 0.211, \quad t_2 = 0.244, \quad t_3 = 0.407, \quad t_4 = 0.652, \quad t_5 = 0.085. \] (21)

We substitute these quantities into (20) to compute the \( M_i \), then use equation (2) to compute \( \hat{\mu}_t \). The constrained maximum likelihood under tree 1 turns out to be \( L(\hat{\mu}_t) \)
\[ -5863.775 \]. Now the likelihood-ratio statistic (6) for testing tree 1 is \( \lambda_r = 2 [ -5591.057 - ( -5863.775 ) ] = 545.436 \). The \( p \) value is \( \sim 2.6 \times 10^{-24} \). The likelihood-ratio method rejects the other two trees just as soundly. Similar calculations for trees 2 and 3 yield likelihood-ratio statistics \( \lambda_r = 550.531 \) for tree 2 and \( \lambda_r = 553.191 \) for tree 3, so the \( p \) values for these trees are slightly smaller than those for tree 1. Tree 1 is thus favored, although every tree is rejected. Tree 1 is the tree proposed by Woese (1987).

The analysis above is intended to illustrate the methods involved and is obviously inadequate for drawing firm conclusions about the correct tree. The small \( p \) values in the evolutionary-parsimony analysis indicate that some deviations from equal transition probabilities are being picked up in all of the trees, particularly in trees 2 and 3. The very small \( p \) values for the likelihood-ratio test indicate that the model is not describing the data at all well. The problem may be that the parameter constraints are unrealistic, or that the sites are not independent and identically distributed, or both. The likelihood-ratio test is probably quite powerful, in which case it will detect deviations from the model even when they are small.

Discussion

We can now compare the asymptotic method of linear invariants with the likelihood-ratio approach. Felsenstein (1988) also provides a very good discussion of some of the ideas mentioned below. The likelihood-ratio approach allows full generality in the model, i.e., 63 free parameters in an unrooted four-species tree, and involves a \( \chi^2 \) test with \( 255 - 63 = 192 \) degrees of freedom. In contrast, linear invariants require parameter constraints. For example, in Cavender's application of the method of linear invariants to a rooted tree with six branches, he allowed 39 free parameters, and the resulting test statistics had either 54 or 68 degrees of freedom, depending on the tree being tested. The likelihood-ratio test in this case would allow 75 free parameters (6 branches \( \times \) 12 parameters/branch + 3 parameters at the initial node) and would have 180 degrees of freedom. In practice, more degrees of freedom usually means more power, so under the assumption of identically distributed sites the likelihood-ratio test is likely to be more powerful than the asymptotic method of linear invariants. This leads to more accurate point estimates and smaller confidence sets.

The advantage of linear invariants is the lack of a need for the assumption of identically distributed sites. Whether this is worth the price of extra constraints on the parameters and of a loss of many degrees of freedom depends on how nonidentically the sites are distributed. If in fact the evolutionary process differs greatly from site to site, the likelihood-ratio method may be unreliable, and the asymptotic method of linear invariants may be preferred. On the other hand, if most of the sites evolve somewhat similarly, the extra degrees of freedom in the likelihood-ratio test probably outweigh the falseness of the assumption of identical distributions.

The discussion in the last two paragraphs is applicable when the sample sizes are large enough for the asymptotics to hold. Without further study, it is difficult to give a rule of thumb as to how large such a sample must be. It is likely that the asymptotics do not hold well for small or moderate sample sizes. In practice it is probably best to estimate significance levels by simulations rather than by asymptotic approximations. This does not apply, of course, in the case of evolutionary parsimony, since exact tests are available, the accuracy of which is known for small as well as large sample sizes.
Acknowledgments

This research was supported by the Deutsches Forschungsgemeinschaft, the National Science Foundation, the National Institutes of Health, and the University of Southern California Faculty Research and Innovation Fund. We would like to express our appreciation to Simon Tavare, whose frequent consultations were invaluable. Thanks also to Walter Fitch, Terry Speed, and Michael Waterman for many helpful suggestions and to an anonymous reviewer for suggesting the statistics \( Y_i \) and \( Z_i \).

APPENDIX

Proof of theorem 1: Direct calculation using equation (2) shows that, under assumption 6, if tree \( i \) is correct, then

\[
\frac{E(P_i - Q_i)}{N} = r(A)m_5(A,C)\left( [m_1(A,A) - m_1(A,G)][m_2(A,A) - m_2(A,G)]
\right.
\]

\[
\times \left. \left[ [m_3(C,C) - m_3(C,U)][m_4(C,C) - m_4(C,U)] \right] \right)
\]

\[
+ r(A)m_5(A,U)\left( [m_1(A,A) - m_1(A,G)][m_2(A,A) - m_2(A,G)]
\right.
\]

\[
\times \left. \left[ [m_3(U,C) - m_3(U,U)][m_4(U,C) - m_4(U,U)] \right] \right)
\]

\[
+ r(G)m_5(G,C)\left( [m_1(G,A) - m_1(G,G)][m_2(G,A) - m_2(G,G)]
\right.
\]

\[
\times \left. \left[ [m_3(C,C) - m_3(C,U)][m_4(C,C) - m_4(C,U)] \right] \right)
\]

\[
+ r(G)m_5(G,U)\left( [m_1(G,A) - m_1(G,G)][m_2(G,A) - m_2(G,G)]
\right.
\]

\[
\times \left. \left[ [m_3(U,C) - m_3(U,U)][m_4(U,C) - m_4(U,U)] \right] \right)
\]

\[
+ r(C)m_5(C,A)\left( [m_1(C,C) - m_1(C,U)][m_2(C,C) - m_2(C,U)]
\right.
\]

\[
\times \left. \left[ [m_3(A,A) - m_3(A,G)][m_4(A,A) - m_4(A,G)] \right] \right)
\]

\[
+ r(C)m_5(C,G)\left( [m_1(C,C) - m_1(C,U)][m_2(C,C) - m_2(C,U)]
\right.
\]

\[
\times \left. \left[ [m_3(G,A) - m_3(G,G)][m_4(G,A) - m_4(G,G)] \right] \right)
\]

\[
+ r(U)m_5(U,A)\left( [m_1(U,C) - m_1(U,U)][m_2(U,C) - m_2(U,U)]
\right.
\]

\[
\times \left. \left[ [m_3(A,A) - m_3(A,G)][m_4(A,A) - m_4(A,G)] \right] \right)
\]

\[
+ r(U)m_5(U,G)\left( [m_1(U,C) - m_1(U,U)][m_2(U,C) - m_2(U,U)]
\right.
\]

\[
\times \left. \left[ [m_3(G,A) - m_3(G,G)][m_4(G,A) - m_4(G,G)] \right] \right)
\].

(A1)

Since \( P_i \) and \( Q_i \) are disjoint sums of multinomial counts, then, conditional on \( P_i + Q_i \), \( P_i \) has the binomial distribution with parameters \( P_i + Q_i \) and \( \pi \), where \( \pi = E(P_i)/E(P_i + Q_i) = 1/2 + E(P_i - Q_i)/2E(P_i + Q_i) \). Equation (22) shows that under assumption 7, when tree \( i \) is correct, \( E(P_i - Q_i) > 0 \), so \( \pi > \frac{1}{2} \).

Proof that \( P_j - Q_j \) and \( P_k - Q_k \) are uncorrelated under the null hypothesis (12) is as follows: By definition, \( P_j - Q_j \) and \( P_k - Q_k \) are uncorrelated if \( E[(P_j - Q_j)(P_k - Q_k)] = E[(P_j - Q_j)]E[(P_k - Q_k)] \). Since, under \( H_0 \), \( E[(P_j - Q_j)] = E[(P_k - Q_k)] = 0 \), it suffices to show that \( E[(P_j - Q_j)(P_k - Q_k)] = 0 \).

Focus on the \( i \)th site where \( 1 \leq i \leq N \). Define \( P_{ij} = 1 \) if the quartet at the \( i \)th site corresponds to one of the quartets for which \( v_j = 1 \). Otherwise \( P_{ij} = 0 \). Define \( Q_{ij} = 1 \) if the quartet at the \( i \)th site corresponds to one of the quartets for which \( v_j = -1 \).
Otherwise $Q_{ij} = 0$. Define $P_{ik}, Q_{ik}$ similarly. Then at most one of $P_{ij}, Q_{ij}, P_{ik}, Q_{ik}$ is nonzero, so $(P_{ij} - Q_{ij})(P_{ik} - Q_{ik}) = 0$. Now $P_{j} - Q_{j} = \sum_{i=1}^{N} (P_{ij} - Q_{ij})$, and $P_{k} - Q_{k} = \sum_{m=1}^{N} (P_{mk} - Q_{mk})$. Therefore $E[ (P_{j} - Q_{j})(P_{k} - Q_{k})] = \sum_{i=1}^{N} \sum_{m=1}^{N} E[ (P_{ij} - Q_{ij})(P_{mk} - Q_{mk})]$. Because the sites are independent, when $i \neq m$, $E[ (P_{ij} - Q_{ij})(P_{mk} - Q_{mk})] = E[ (P_{ij} - Q_{ij})]E[ (P_{mk} - Q_{mk})] = 0$. Therefore $E[ (P_{j} - Q_{j})(P_{k} - Q_{k})] = \sum_{i=1}^{N} E[ (P_{ij} - Q_{ij})] = 0$.

**LITERATURE CITED**


———. Necessary conditions for the method of inferring phylogeny by linear invariants. (submitted).


BARRY G. HALL, reviewing editor

Received May 29, 1990; revision received July 8, 1990

Accepted August 20, 1990