A Rate-independent Technique for Analysis of Nucleic Acid Sequences: Evolutionary Parsimony

James A. Lake
Molecular Biology Institute and Department of Biology,
University of California, Los Angeles

The method of evolutionary parsimony—or operator invariants—is a technique of nucleic acid sequence analysis related to parsimony analysis and explicitly designed for determining evolutionary relationships among four distantly related taxa. The method is independent of substitution rates because it is derived from consideration of the group properties of substitution operators rather than from an analysis of the probabilities of substitution in branches of a tree. In both parsimony and evolutionary parsimony, three patterns of nucleotide substitution are associated one-to-one with the three topologically linked trees for four taxa. In evolutionary parsimony, the three quantities are operator invariants. These invariants are the remnants of substitutions that have occurred in the interior branch of the tree and are analogous to the substitutions assigned to the central branch by parsimony. The two invariants associated with the incorrect trees must equal zero (statistically), whereas only the correct tree can have a nonzero invariant. The $\chi^2$-test is used to ascertain the nonzero invariant and the statistically favored tree. Examples, obtained using data calculated with evolutionary rates and branchings designed to camouflage the true tree, show that the method accurately predicts the tree, even when substitution rates differ greatly in neighboring peripheral branches (conditions under which parsimony will consistently fail). As the number of substitutions in peripheral branches becomes fewer, the parsimony and the evolutionary-parsimony solutions converge. The method is robust and easy to use.

Introduction

Parsimony analysis is one of the most widely used and generally accepted methods of phylogenetic analysis (Fitch 1977). It is characterized by both intellectual and operational simplicity. Yet, under conditions of unequal rates of substitution, parsimony can select an incorrect tree. Parsimony, as a successful method of phylogenetic determination, represents a baseline against which other methods can be measured.

The best-understood instance in which parsimony can incorrectly predict an unrooted tree occurs when sequences in neighboring peripheral branches of a tree evolve at greatly different rates. Felsenstein (1978), investigating a two-state model, showed that when highly different rates occur among four sequences the most parsimonious unrooted tree places the two most highly substituted peripheral branches on one side of the tree and the two least substituted peripheral branches together on the other side. This tree will be chosen no matter what the topology of the true tree.

In this paper I propose a method of phylogenetic analysis—related to parsimony...
analysis and called evolutionary parsimony or the method of operator invariants—that can predict the correct tree even when rates of nucleotide substitution differ by an order of magnitude in adjacent branches of the unrooted tree. The method is robust, and it is easy to calculate. In it, three quantities named “operator invariants” are calculated from four aligned nucleic acid sequences. The invariants are remnants of substitutions that have occurred in the interior branch of the tree and are analogous to the substitutions assigned by parsimony to the interior branch. Both the operator invariants and the parsimony terms are derived by analysis of patterns present in the aligned sequences. These three operator invariants are then used to predict the statistically significant dendrogram. The evolutionary-parsimony method is investigated by using data calculated from a tree of known topology and shown to accurately predict the initial tree under a variety of conditions, particularly within the zone in which the Felsenstein conditions prevail.

Theory

Determining the unrooted evolutionary tree that best reconstructs the evolution of four taxa requires the discrimination of a single tree from a set of three alternative tree topologies. Hence, the problem for four taxa serves as the simplest case model for developing a method to discriminate among topologically distinct dendrograms.

Parsimony and evolutionary parsimony have related—but differing—criteria. Parsimony selects the tree that requires the minimum number of substitutions. In contrast, evolutionary parsimony selects the tree that requires the minimum number of consistent substitutions (“consistent” is used to imply consistency with evolution in the peripheral branches of the tree). In the limit that the number of substitutions in the peripheral branches of the tree becomes small relative to those in the central branch, all substitutions become consistent ones and the parsimony solution converges to the evolutionary-parsimony solution.

Two simple examples serve to illustrate the differences between substitutions and consistent substitutions. Consider the trees in figure 1. The initial tree in 1a refers to the tree used to calculate the sequences, and the most parsimonious tree is the tree inferred from analysis of the calculated sequences. In these examples, the probability of substitution is equal for all bases. Thus, for an RNA sequence (nucleotides C, U, A, or G), an A would be replaced with equal likelihood by U, C, or G. When there is a high probability of nucleotide substitution in the central branch of the initial tree and low probabilities in the other branches, one finds the pattern xxyy at most positions. This is the informative pattern for parsimony and identifies the tree that positions taxa 1 and 2 together and 3 and 4 together as being the most parsimonious. The absence of other patterns, except for xxxx, indicates that most substitutions are consistent ones. In this example, parsimony correctly predicts the initial tree.

In the second example, figure 1b, the probability of nucleotide substitutions is very large in the peripheral branches leading to taxa 1 and 3 and small in the branches leading to taxa 2 and 4 and in the central branch. Typical sequences are shown in the panel below the true tree, but the expected pattern xxyy that is diagnostic for the true tree is not present. Contrary to one’s expectations, the informative pattern for parsimony that is present is xyxy. (For this example calculations show that, in the limit of infinite substitution in branches 1 and 3, the xyxy pattern should occur at fully \( \frac{3}{16} \) of the positions). Hence, the most parsimonious—or minimum substitution—tree in figure 1b is not the initial tree but is the tree that connects taxa 1 and 3 and connects
Fig. 1.—Examples illustrating when parsimony correctly selects a tree and when it fails. Branch lengths (of either 0 or ~0.8) represent the relative probabilities of a nucleotide difference at any one position. The patterns (Cavender 1981) observed in the aligned sequences and the number of their occurrences are shown adjacent to the sequences. In 1a parsimony correctly predicts the true tree. In 1b the tree predicted by parsimony places taxa 1 and 3 and taxa 2 and 4 together in a different topological arrangement from that found in the true tree.

In this example, parsimony has picked an incorrect tree because substitutions inserted in peripheral branches of the tree have mimicked the pattern normally produced by substitutions in the central branch of an alternative tree topology. Those substitutions that mimic an incorrect pattern are described as inconsistent substitutions. The presence of a second type of pattern (xyxz) indicates, however, that xyxy might represent inconsistent substitutions. In the following sections, explicit definitions of both consistent and inconsistent substitutions are detailed and a parsimony-like procedure for determining trees using consistent substitutions is presented.

A Vector Representation

Descriptions of operator invariants and of both consistent and inconsistent substitutions are facilitated by using a vector representation of sequences. In this representation a set of four aligned sequences, each of length $n$, is represented as the vector sum of $n$ vectors. Thus, in figure 2a, each of the 256 (or $4^4$) possible combinations of nucleotides represents a direction in a 256-dimensional sequence space. For example, the vector CGGC, present at the third position along the sequence, is one of 20 subvectors that make up the sequence vector, $S$.

This 256-dimensional space can be considerably simplified if one includes information about the (molecular-biological) details of the substitution process. Because
DNA copying and repair mechanisms distinguish most readily between the larger purines and the smaller pyrimidines, exchanges that substitute one purine for another or one pyrimidine for another (transitions) occur much more frequently than those that interchange a pyrimidine and a purine (transversions). Wilson and co-workers (Brown et al. 1982), for example, have shown that, for mitochondrial DNAs, transitions occur an order of magnitude more frequently than transversions. This difference is applied to the definition of basis vectors in the following paragraph.

Distinguishing between transitions and transversions allows one to reduce the number of basis vectors from 256 to 36. This simpler representation, which replaces each of the nucleotide letter symbols with the numbers one through four, is shown in figure 2b. Since the representation of the nucleotide in position one in a vector is arbitrary, a “1” is assigned to represent it and all others of the same type. Any nucleotide related to the nucleotide in position one by a transition is assigned a “2” to represent it. The first nucleotide (if any) that is related to the nucleotide in position one by a transversion (and all others of the same type) is represented by a “3.” Finally, any nucleotide related by a transition to the type represented by a “3” is represented by a “4.” With this notation, any combination of four nucleotides can be represented by one of 36 types.

To simplify this further, a shorthand, one-letter, notation is introduced (table 1). In the example in figure 2a, position CGGC becomes 1331 in 2b and is abbreviated as vector component G; UGGG becomes 1333 and is abbreviated as component A. Thus the set of four aligned sequences can be represented by the single line of components shown in figure 2c. Similarly, a unit vector pointing in the G direction will be represented as \( \mathbf{G} \), in the one-letter code.
This notation allows one to describe four aligned sequences either as spectral components of the aligned sequences or as a sequence vector. In the example in figure 2 the vector component $G$ (1331) occurs four times, and the value of the $G$ spectral component is listed as $G = 4$. Similarly, the sequence vector, $\mathbf{S}$—corresponding to spectral components $a, A, b, B$, etc. and unit vectors $\hat{a}, \hat{A}$, etc.—is written as

$$\mathbf{S} = a \hat{a} + A \hat{A} + b \hat{b} + B \hat{B} + \cdots.$$  \hspace{1cm} (1)

Operator-Invariant Analysis

The spectral components from the previous example are used to illustrate parsimony analysis and evolutionary parsimony analysis (fig. 3). In this example and throughout this paper, the parsimony analysis will consider only transversion substitutions in the central branch of the tree. Thus, each of the three spectral components $E$, $F$, and $G$ is most parsimoniously associated with one of three possible evolutionary trees called the $E$, the $F$, and the $G$ trees, respectively. These trees are shown in figure 4. The tree associated with the largest component is most parsimonious, i.e., requires the minimal number of transversion substitutions. Parsimony analysis of the spectral components in figure 3 identifies the $G$ tree as being most parsimonious.

The method of evolutionary parsimony is similar to parsimony but uses additional spectral components to determine consistent substitutions. As shown in figure 3 the operator invariants are linear combinations of four spectral components. As with parsimony, each invariant ($X, Y$, or $Z$) is associated with a tree (the $E$, $F$, and $G$ trees, respectively). The evolutionary interpretation of the operator invariants is that the

### ANALYSIS OF SPECTRUM

<table>
<thead>
<tr>
<th>EVOLUTIONARY PARSIMONY</th>
<th>PARSIMONY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>'E TREE'</strong></td>
<td></td>
</tr>
<tr>
<td>$X = E + u - H - J$</td>
<td>$E = 3$</td>
</tr>
<tr>
<td>$= 3 + 3 - 0 - 0$</td>
<td>SIGNIFICANT, $x^2 = 6.00, P &lt; .02$</td>
</tr>
<tr>
<td>$= 6$</td>
<td>SIGNIFICANCE UNKNOWN</td>
</tr>
<tr>
<td><strong>'F TREE'</strong></td>
<td></td>
</tr>
<tr>
<td>$Y = F + v - L - N$</td>
<td>$F = 0$</td>
</tr>
<tr>
<td>$= 0 + 0 - 1 - 0$</td>
<td>NOT SIGNIFICANT, $x^2 = 1.00$</td>
</tr>
<tr>
<td>$= -1$</td>
<td>SIGNIFICANCE UNKNOWN</td>
</tr>
<tr>
<td><strong>'G TREE'</strong></td>
<td></td>
</tr>
<tr>
<td>$Z = G + w - Q - S$</td>
<td>$G = 4$</td>
</tr>
<tr>
<td>$= 4 + 0 - 2 - 1$</td>
<td>NOT SIGNIFICANT, $x^2 = 0.14$</td>
</tr>
<tr>
<td>$= 1$</td>
<td>SIGNIFICANCE UNKNOWN</td>
</tr>
</tbody>
</table>

### CONCLUSIONS

<table>
<thead>
<tr>
<th>EVOLUTIONARY PARSIMONY</th>
<th>PARSIMONY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>'E TREE'</strong> IS STATISTICALLY FAVORED TREE ($P &lt; 0.02$)</td>
<td><strong>'G TREE' IS FAVORED BY ONE EVENT</strong></td>
</tr>
</tbody>
</table>

Fig. 3.—The operator spectral components derived in fig. 2 analyzed using both the parsimony method and the method of evolutionary parsimony. In this example, evolutionary parsimony selects the correct $E$ tree even though tree $G$ is most parsimonious.
are the remnants of transversion substitutions made in the central branch of the tree. Only the historically correct tree has contributed consistent substitutions to the sequences, and only it can have a nonzero invariant. The two incorrect trees cannot have remnants and thus will be associated with (statistically) zero invariants. In the example in figure 3, only the X invariant (the E tree) is found to be significantly greater than zero when the invariants are analyzed by the $\chi^2$-test (see Statistical Tests and Tree Selection below). The observation that the Z invariant is approximately equal to zero, even though the G tree is the most parsimonious, indicates that many of the substitutions supporting the G tree are inconsistent substitutions.

Each operator invariant (X, Y, or Z) has three types of spectral components measuring different aspects of the evolutionary process—namely, a parsimony term (E, F, or G), two peripheral branches terms (H and J, L and N, or Q and S), and a compensatory term (u, v, or w). As a guide to understanding the invariants, examples of the functioning of their components are given below.

Under conditions of low substitution rates in peripheral tree branches, the peripheral branches' terms and the compensatory term will be small and only the parsimony term be large. This is the reason that, in this limit, parsimony and evolutionary parsimony predict the same tree.

When transversion substitutions in peripheral branches of the tree are frequent, this can artifactually increase the parsimony term associated with the incorrect trees. A major function of the peripheral branches term is to compensate for this effect. This is illustrated for the Y invariant by the example in figure 1b. When frequent substitutions occur in branches 2 and 4 but not in the other branches, then both the xxyx pattern and the xxyz pattern become more frequent. In figure 1b, the components F (1313) and N (1314) occur with nearly equal frequency. Hence the contribution to Y from F is balanced by an equal but negative contribution to Y from N. If the model had introduced frequent substitutions in branches 1 and 3 but not in the other branches, then the L component (rather than N) would have increased to compensate for the increase in F. Again, since they are of opposite sign, they would have little effect on the Y invariant. Thus, both L and N contributions to the Y invariant negate increases in the F component resulting from substitutions in peripheral branches of the tree. Similar considerations apply to the two other invariants.

The compensating term performs multiple functions in the invariants. One function is to compensate for the effect of transitions. When transitions are frequent, the compensating term can increase the contribution of the parsimony term. For example, E components (1133) can be converted to u components by one transition occurring in branch 1 or 2 and one occurring in branch 3 and 4. Since the E and u components have the same sign, the X invariant is compensated for this effect. The compensating terms v and w have similar effects on the Y and Z invariants.

The above discussion is primarily to give one a feeling for the origins of these three types of terms in the invariants. Rigorous derivations are provided in the following sections.

Substitution Operators

All unrooted trees relating the evolution of four taxa contain four peripheral branches that are independent of their ancestral relations. To account for both transversion and transition substitutions on each requires specification of eight separate operators as enumerated in figure 4. The $a$, $b$, $c$, and $d$ operators introduce transitions
and the $A$, $B$, $C$, and $D$ operators introduce transversions in branches 1, 2, 3, and 4, respectively.

Consider the rooted evolutionary tree shown in figure 5 and consider the evolution of a single homologous nucleotide position as substitutions occur (i.e., as substitution operators are applied). For this and all subsequent examples in this paper, calculations will be based on the E tree. In this example, the last common ancestor existed at time $t_0$ and had evolved into four separate taxa by time $t_1$. The tree topology is not significant for the definition of the eight operators and is left unspecified. At the time of their initial separation the sequences were AAUU, corresponding to unit vector $\hat{E}$ (1133). At time $t_2$ a transition in lineage 1, the $a$ operator, changes the sequence to GAUU, corresponding to unit vector $\hat{H}$ (1233). At $t_3$ a $c$ operator converts the $\hat{H}$ unit vector to the $\hat{u}$ (1235), and the final substitution, $D$, at time $t_4$ converts $\hat{u}$ to the unit vector $\hat{i}$. The usefulness of the substitution operators is that they provide a convenient mathematical description of all transitions and transversions that can occur in peripheral branches of a tree. Furthermore, these eight substitution operators may be used with all unrooted trees as well as with all rooted ones.

These operators provide a complete description of all evolution that can occur in peripheral branches of a tree. The mathematical properties of these eight operators—in effect, the multiplication table for them—are listed in table 1. This table indicates the result of applying any substitution operator to any unit vector. Thus the $a$ operator transforms unit vector $\hat{Q}$ into $\hat{G}$ and the $B$ operator transforms vector $\hat{u}$ into either $\hat{r}$
or \( \hat{m} \). The operator formalism and operator multiplication table are the central mathematical formalisms required by the theory of operator invariants.

A Vector-Space Description of Operator Invariants

The three operator invariants have a very simple vector-space representation. The invariants—\( X, Y, \) and \( Z \)—correspond to projection of the \( \vec{S} \) vector onto one of three special directions (principal axes). These three principal axes are \( \vec{X}, \vec{Y}, \) and \( \vec{Z} \) and correspond to the respective invariants. The definitions of these principal axes are

\[
\begin{align*}
\vec{X} &= \hat{E} + \hat{u} - \hat{H} - \hat{J}; \\
\vec{Y} &= \hat{F} + \hat{v} - \hat{L} - \hat{N}; \\
\vec{Z} &= \hat{G} + \hat{w} - \hat{Q} - \hat{S}.
\end{align*}
\]

Hence, when the definition of the \( \vec{S} \) vector (given in eq. 1) is used, the invariants are

\[
\begin{align*}
X &= \vec{X} \cdot \vec{S} = E + u - H - J; \\
Y &= \vec{Y} \cdot \vec{S} = F + v - L - N; \\
Z &= \vec{Z} \cdot \vec{S} = G + w - Q - S.
\end{align*}
\]

If the \( E \) tree is the historically correct tree, then the \( \vec{S} \) vector, during the course of its evolution, will evolve so that it continually remains perpendicular to the \( \vec{Y} \) and

\[
\begin{array}{c|cccc}
\text{Present Day} & 1 & 2 & 3 & 4 \\
\hline
\text{Taxa} & G & A & C & A \\
\hline
\text{Start} & t_0 \\
\text{t_1} & A & A & U & U \\
\text{t_2} & A & A & U & U \\
\text{t_3} & G & A & C & U \\
\text{t_4} & G & A & C & A \\
\hline
\end{array}
\]

\begin{align*}
\text{Tree topology} \\
\text{not relevant}
\end{align*}

Fig. 5.—A rooted evolutionary tree illustrating the evolution of sequences at a single nucleotide position. The branching relationship is not shown since the arguments based on the figure are independent of the topology. Substitutions in the peripheral branches at times \( t_2, t_3, \) and \( t_4 \) insert transversions and transitions. The operators specify both the nature of the substitution (i.e., lowercase signifies a transition and uppercase signifies a transversion) and the branch in which it occurs.
Table 1
Operator Multiplication

<table>
<thead>
<tr>
<th>OPERATORS</th>
<th>Transitions</th>
<th>Transversions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a b c d</td>
<td>A B C D</td>
</tr>
<tr>
<td>Ò (1111)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>ã (1222)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>b (1211)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>c (1121)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>d (1112)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>A (1333)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>B (1311)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>C (1311)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>D (1113)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>ã (1122)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>f (1212)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>g (1221)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>E (1133)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>F (1313)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>G (1331)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>H (1733)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>h (1322)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>i (1344)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>J (1134)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>k (1312)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>L (1323)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>m (1343)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>N (1134)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>ã (1343)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>ã (1312)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>Q (1332)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>q (1223)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>r (1334)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>s (1231)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>t (1321)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>u (1234)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>v (1324)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
<tr>
<td>w (1342)</td>
<td>ã ã ã ã</td>
<td>Æ Æ Æ Æ</td>
</tr>
</tbody>
</table>

**NOTE.**—All 36 unit vectors are listed at the left of the table, in both their single-letter and their four-digit representations. The eight peripheral branch-substitution operators are listed along the top row. The product of an operator and each unit vector is listed below each operator. In those instances in which two vector products are indicated, the operator is assumed to generate both unit vectors with equal probability. These eight operators are commutative.

̅Z vectors. The vector conditions for the E-tree evolutionary path are as follows (where \( \neq \) indicates "not necessarily equal to":)

\[
\overline{X} \cdot \overline{S} \neq 0; \\
\overline{Y} \cdot \overline{S} = 0; \\
\overline{Z} \cdot \overline{S} = 0
\]
Similarly, (1) if the $F$ tree is historically correct, then the $\mathbf{S}$ vector will evolve so that it continually remains perpendicular to the $\mathbf{X}$ and $\mathbf{Z}$ vectors and (2) if the $G$ tree is correct, then the $\mathbf{S}$ vector will evolve so that it continually remains perpendicular to the $\mathbf{X}$ and $\mathbf{V}$ vectors. A general proof of equations (8a), (8b), and (8c) is presented in the Appendix.

The time-dependent evolutionary changes of the $\mathbf{S}$ vector are illustrated in figure 6 with an $E$ tree that has been rooted in the middle of the central branch. At times preceding $t_0$, the first bifurcation of the tree, all sequence components point in the $\mathbf{O}$ (1111) direction. As the sequence evolves through the central branch of the tree, only two types of substitutions will alter the sequences; and these will convert $\mathbf{O}$ unit vectors into $\mathbf{h}$ (1122) or into $\mathbf{E}$ (1133) vectors. Thus, as one observes the projection of the 36-dimensional $\mathbf{S}$ vector onto three-dimensional $E$, $H$, and $Y$ space (fig. 6b), the component of $\mathbf{S}$ in the $E$ direction will increase until time $t_1$. At this point, all evolution will have been in the central branch of the tree and the $X$ operator invariant and the parsimony term ($E$) will have identical values. At this time the $\mathbf{S}$ vector will have no components in the $Y$ direction (or in the $Z$ direction [not shown]), since the $\mathbf{O}$, $\mathbf{h}$, and $\mathbf{E}$ unit vectors are all perpendicular to these principal axes. At times following $t_1$, evolution in peripheral branches of the tree will convert some $\mathbf{E}$ vectors into the $H$ and other directions; however, the sequence $\mathbf{S}$ will be under no constraints to evolve in a direction perpendicular to $X$. As $\mathbf{S}$ evolves further, it will (stochastically) be constrained to follow a path such that $\mathbf{V} \cdot \mathbf{S}$ continues, on average, to equal zero. The $\mathbf{S}$ vector may even evolve so that its projection onto the $F$ or $G$ axes is greater than

---

**FIG. 6.—The evolution of the $\mathbf{S}$ vector in space.** The rooted evolutionary tree used to derive the sequence vector is shown in 6a. This $E$ tree is rooted in the central branch of the tree. The progress of the $\mathbf{S}$ vector through time is shown projected onto $E$, $H$, and $Y$ space in 6b. At time $t_0$ the sequence bifurcates and $\mathbf{S}$ starts to project on the $E$ axis. $\mathbf{S}$ continues to grow until time $t_1$, when taxa 1 and 2 and taxa 3 and 4 separate. At this point $\mathbf{S}$ evolves in a general direction and is subject only to the condition that it remain perpendicular to the $\mathbf{Y}$ and $\mathbf{Z}$ principal axes. There are no restrictions on the evolution of $\mathbf{S}$ in the direction of the $X$ principal axis, however.
onto the $\mathbf{E}$ axis; but, nevertheless, on average, it does not project onto the $\mathbf{Y}$ or $\mathbf{Z}$ principal axes. The difference, then, (for the $\mathbf{E}$ tree) is that while the parsimony terms—$\mathbf{F}$ and $\mathbf{G}$—may have values significantly greater than zero, the operator invariants—$\mathbf{Y}$ and $\mathbf{Z}$—will not, on average, deviate significantly from zero.

In general, the $\mathbf{X}$ component will have a maximum value at time $t_1$ and reduced values for times later than $t_1$. In this sense it is a remnant of the value for $\mathbf{E}$ found at time $t_1$. I have shown (J. A. Lake, accepted) that one can estimate the earlier value of $\mathbf{E}$ and that it is a natural and useful measure of the length of the central branch.

### Statistical Tests and Tree Selection

One expects that $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ will have two values not significantly different from zero and one positive value that may be significantly different. To test for this significance, one notes that $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are the differences between two positive terms and two negative ones. Therefore, let $P_X = E + u$; $B_X = H + J$; $P_Y = F + v$; $B_Y = L + N$; $P_Z = G + w$; and $B_Z = Q + S$. If we let $(P + B)/2$ be the expected value of $P$ and $B$ for any one set of subscripts $\mathbf{X}$, $\mathbf{Y}$, or $\mathbf{Z}$, it can be shown that $\chi^2$ for 1 df is

$$\chi^2 = (P - B)^2/(P + B).$$

A tree is considered proven when $\chi^2$ is calculated for each of the three possible sets of subscripts and only one of them shows $P$ significantly greater than $B$. In the following section I shall refer to $P$ as the parsimony-like term and to $B$ as the background term.

### Results

Evolutionary parsimony has been designed to determine topologies when substitution rates in adjacent branches of dendrograms are very different, when substitution rates vary greatly at different positions within a sequence, and when substitutions in some branches are quite frequent. Under these extreme conditions the normally robust technique of parsimony analysis can fail. Some conditions under which parsimony fails to converge to the correct tree have been analyzed in detail by Felsenstein (1978). In the following section, trees with unequal substitutions in adjacent peripheral arms, similar to those analyzed by Felsenstein, are used to test the method of evolutionary parsimony.

### When Can Parsimony Fail?

Our calculations show that parsimony analysis can fail to converge to the true tree for essentially all choices for the lengths of the peripheral branches, provided that the central branch of the tree is sufficiently small. As for all the calculations in this paper, for those in figure 7 the $\mathbf{E}$ tree has been used to calculate the sequences. In these calculations the probability of a transversion is the same in the $\mathbf{A}$ and $\mathbf{C}$ branches and the same (but generally set to a different value) in the $\mathbf{B}$ and $\mathbf{D}$ branches. The probability of a transversion occurring in the central branch is zero. The calculations have been made using two different assumptions that correspond to a single-hit (or selection) model or to a multiple-hit (or neutral) model. Details of these models are described in the legend to figure 7.

The results in figure 7 show when parsimony analysis can predict an incorrect tree. This can happen whenever the ratio of the parsimony-like term for the $\mathbf{F}$ tree, $Y_p$, exceeds the parsimony-like term from the correct $\mathbf{E}$ tree, $X_p$. Independently of
"0 0.5 1.0
a

$\begin{array}{c}
\text{NEUTRAL MODEL} \\
\text{SELECTION MODEL}
\end{array}$

\begin{align*}
\text{TRUE TREE (E TREE)}
\end{align*}

Fig. 7.—Ratios of the parsimony-like term for the "most-parsimonious" tree compared with the parsimony-like term for the true tree (i.e., $Y_p/X_p$). The E tree shown in 7c has been used to calculate the sequences, which are 1,000 nucleotides long. In these calculations the probability of a transversion, at any one sequence position, is the same in the A and C branches and the same (but generally set to a different value) in the B and the D branches. The probability of a transversion at any sequence position in the central branch is epsilon, where epsilon is < the minimum of $P_A$ and $P_B$. The calculations have been made for using two different assumptions that correspond to a single-hit (or selection) model or to a multiple-hit (or neutral) model. The ratios for the neutral model and for the selection model are shown in 7a and 7b, respectively. In the single-hit (or selection) model, it is assumed that once a transversion has been made in a segment of the tree it cannot be followed by a second transversion. In the multiple-hit (or neutral) approximation, it is assumed that transversions are completely reversible. This latter model assumes that the transversion operator is applied in multiple Bernoulli trials ($N$ trials with a probability of transversion of $P_{A}/N$). In the limit of large $N$, the probability of the A transversion is $P_{A}/(1 + P_{A})$, the probability of no change is $[1 - (P_{A}/2)]/(1 + P_{A})$, and the probability of the a transition is $(P_{A}/2)/[1 + P_{A}]$. This last term occurs because two successive transversions can create the effect of a transition. Plotting the parsimony-like terms, such as $E + u$, rather than the true parsimony terms, such as $E$, will underestimate the effect. It has the advantage that the plots are symmetrical.

whether a neutral model (fig. 7a) or a selection model (fig. 7b) is used for the calculations, one finds that, for all values of $P_A$ and $P_B$ (corresponding to the probability of substitution in branch A or B, respectively, at a single position in the aligned sequences), the $Y_p/X_p$ ratio is greater than one, except along the line $P_A = P_B$. This means that for almost any choice of transversion rates in the peripheral branches of the tree, parsimony can predict an incorrect tree (provided that $P_E$ is sufficiently small). This is true whether substitutions are reversible or irreversible. The effect is larger for the irreversible approximation. Under some conditions, an incorrect tree can be favored over the correct one by $Y_p/X_p$ ratios of 10:1 or even 10,000:1. The same curves are also obtained when transitions are included in the calculations, provided only that their relative probabilities in each of the branches are proportional (but not necessarily equal) to those used for $P_A$ and $P_B$. 

Downloaded from https://academic.oup.com/mbe/article-abstract/4/2/167/1184661 by guest on 27 April 2019
Evolutionary Parsimony Tested under Moderate Conditions

For the method of evolutionary parsimony to be useful, it should function under a wide range of conditions. For the initial tests I chose the tree shape used in the Felsenstein (1978) test—but used it under conditions of moderate evolutionary rates. Hence, for these first four examples, the parsimony method should normally be valid. Substitution probabilities used for these tests are listed in the legend to table 2. (In general, percent differences were ~10% for transversions and for transitions in branches A and C and ~5% for transversions and for transitions in branches B, D, and E.) In two of the four examples, a transition or a transversion was assumed to be fixed once it had occurred in any of the segments of the tree, while in the other two examples substitutions were assumed to be completely reversible. The results compiled in table 2 summarize 10 runs under each condition. If only transversions are permitted, the median $\chi^2$-value is significant at greater than the 99.99% level. If both transitions and transversions are permitted, the median significance drops to 98%-99%. This is because transitions, on average, convert $X_p$ terms into $X_b$ terms. There were no instances of incorrect trees being chosen by the invariants technique, although in two instances

<table>
<thead>
<tr>
<th>Operator Invariants</th>
<th>Parsimony-like Events</th>
<th>Background Events</th>
<th>Median Significance</th>
<th>No. of Events with &gt;95% Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Transversions Only: Selection Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$</td>
<td>39.6</td>
<td>2.7</td>
<td>&gt;99.99%</td>
<td>10/10</td>
</tr>
<tr>
<td>$Y$</td>
<td>7.5</td>
<td>5.5</td>
<td>NS</td>
<td>0/10</td>
</tr>
<tr>
<td>$Z$</td>
<td>4.6</td>
<td>5.9</td>
<td>NS</td>
<td>0/10</td>
</tr>
<tr>
<td>B. Transversions and Transitions: Selection Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$</td>
<td>29.4</td>
<td>12.9</td>
<td>&gt;98%</td>
<td>10/10</td>
</tr>
<tr>
<td>$Y$</td>
<td>7.7</td>
<td>5.3</td>
<td>NS</td>
<td>0/10</td>
</tr>
<tr>
<td>$Z$</td>
<td>5.1</td>
<td>5.4</td>
<td>NS</td>
<td>0/10</td>
</tr>
<tr>
<td>C. Transversions Only: Neutral Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$</td>
<td>36.5</td>
<td>2.5</td>
<td>&gt;99.99%</td>
<td>10/10</td>
</tr>
<tr>
<td>$Y$</td>
<td>6.1</td>
<td>4.1</td>
<td>NS</td>
<td>1/10</td>
</tr>
<tr>
<td>$Z$</td>
<td>4.0</td>
<td>6.3</td>
<td>NS</td>
<td>0/10</td>
</tr>
<tr>
<td>D. Transversions and Transitions: Neutral Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$</td>
<td>27.2</td>
<td>11.8</td>
<td>&gt;99%</td>
<td>7/10</td>
</tr>
<tr>
<td>$Y$</td>
<td>6.5</td>
<td>3.7</td>
<td>NS</td>
<td>1/10</td>
</tr>
<tr>
<td>$Z$</td>
<td>4.6</td>
<td>5.7</td>
<td>NS</td>
<td>0/10</td>
</tr>
</tbody>
</table>

**NOTE.**—In all four tests, moderate—although unequal—rates of substitution obtain. For these tests I used the following values for the transversion probabilities: $P_a = P_c = 0.1$ and $P_b = P_d = P_e = 0.05$ for all four trees. For two of the trees I also used these same values for the transitions—i.e., $P_a = P_c = 0.1$ and $P_b = P_d = P_e = 0.05$—and for the other two examples the transition probabilities were set to zero. In two of the four examples, a transition or a transversion was assumed to be fixed once it had occurred in any of the segments of the tree, while in the other two examples substitutions were assumed to be completely reversible. The sequences were 1,000 nucleotides long.
the F tree showed a significant (98%) signal. This was approximately the number of times (four) that this would have been expected on a statistical basis. In both instances the E tree was more significantly favored. The most difficult conditions for detecting the correct tree appear to be when both transitions and transversions are present and when they are reversible. Under these conditions, the \( \chi^2 \)-value associated with the E tree did not exceed the 95% level in three instances. This suggests either that a sequence somewhat longer than 1,000 nucleotides is needed for these conditions or (as discussed below) that additional, related sequences should be used. I conclude that the transition/transversion method of evolutionary parsimony accurately predicts the tree and provides a figure of merit on which to accept a tree.

Evolutionary Parsimony Tested under Conditions in Which Parsimony Fails

To test the ability of evolutionary parsimony to find the true phylogeny, I constructed three inherently difficult examples and designed them so that the parsimony technique will consistently select an incorrect tree.

The first example tests the method under conditions in which the F spectral term exceeds the E component by a factor of 10 to 1. Under these conditions the parsimony method favors both of the incorrect trees over the true tree. Because the analysis of the Felsenstein zone (fig. 7) indicated that parsimony failed most dramatically when substitutions were irreversible (the selection model), the sequences to be analyzed were calculated assuming that transversions were irreversible and using adjacent peripheral arms with transversion probabilities that differed by an order of magnitude. (Details of the calculation are given in the legend to fig. 8). The results are displayed as a significance plot in figure 8. In this plot, a running total of \( \chi^2 \)-values is presented for both the E and the F (the most parsimonious) trees. This plot is particularly useful since lines of constant significance are approximately horizontal. Each point on the curve corresponds to one additional set of sequences that have been analyzed. By the time 30 different data sets have been analyzed, the curves have separated so that the E tree is significant at greater than the 99.9% level while the significance of the most parsimonious F tree is consistent with the prediction of the null hypothesis. Summary data for this example are shown in table 3a. In spite of these extreme conditions, the technique nevertheless easily detects the correct tree.

The second test combines four elements that make it difficult to determine the true phylogeny. These are high transition and high transversion rates, large differences in rates in different branches, reversible substitutions, and a low transversion rate in the central branch. The significance plot for this example is shown in figure 9, and details of the calculation are presented in the legend. The curve corresponding to the E tree increases in significance as additional data are added. After 50 sets of sequences have been included, the tree is significant at the 99.5% level. In contrast, the most parsimonious tree—the F tree—does not violate the null hypothesis. As summarized in table 3b, the G tree also does not violate the null hypothesis. Note that under these conditions, the parsimony-like term corresponding to the correct tree, \( X_p \), is only \( \sim 25\% \) larger than the background term, indicating the considerable mixing of signal and background that has occurred.

The third example justifies using a two-sided \( \chi^2 \)-test rather than a (perhaps stronger) one-sided statistical test. As discussed in the Theory section above, the parsimony-like term corresponding to the correct tree will usually be larger than the background
Fig. 8.—Significance plot for trees with irreversible transversions. These values cause the most parsimonious tree to be preferred over the correct tree by a factor of 10. Sequence evolution was stimulated for the E tree using the following values: $P_A = P_C = 0.7; P_B = P_D = P_E = 0.07; \text{and } P_a = P_b = P_c = P_d = P_e = 0$. Transversions were irreversible. The evolving sequence for each tree was 1,000 nucleotides long. The cumulative number of trees is plotted against the square root of the cumulative sum of the $\chi^2$ values for these trees minus the square root of the number of trees less one for both the E and the F (the most parsimonious) trees. I call this a "significance plot". The X invariant, corresponding to the correct tree, is shown by black dots. The Y invariant, corresponding to the most parsimonious—but incorrect—tree is shown by open circles. The correct solution, after 50 trees, is significant at approximately the 0.03% level.

A Practical Test: The Endosymbiotic Theory of Mitochondrial Origins

The previous examples assumed that an essentially infinite length of sequence will be available for comparison. In practice this is not the case. I have used ribosomal RNA sequences to test the endosymbiotic theory for the origin of the fungal mitochondrion.

Wilson suggested a formal test, based solely on a comparison of the topologies of two unrooted trees, between the endosymbiotic and the autogenous theories of
<table>
<thead>
<tr>
<th>Tree</th>
<th>Parsimony-like Term</th>
<th>Background Term</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E . .</td>
<td>942</td>
<td>634</td>
<td>99.29</td>
</tr>
<tr>
<td>F . .</td>
<td>10013</td>
<td>9959</td>
<td>50.21</td>
</tr>
<tr>
<td>G . .</td>
<td>1314</td>
<td>1403</td>
<td>46.89</td>
</tr>
</tbody>
</table>

Table 3a

<table>
<thead>
<tr>
<th>Tree</th>
<th>Parsimony-like Term</th>
<th>Background Term</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E . .</td>
<td>943</td>
<td>744</td>
<td>79.72</td>
</tr>
<tr>
<td>F . .</td>
<td>1227</td>
<td>1225</td>
<td>53.55</td>
</tr>
<tr>
<td>G . .</td>
<td>393</td>
<td>362</td>
<td>57.36</td>
</tr>
</tbody>
</table>

Table 3b

<table>
<thead>
<tr>
<th>Tree</th>
<th>Parsimony-like Term</th>
<th>Background Term</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F . .</td>
<td>979</td>
<td>1775</td>
<td>74.15</td>
</tr>
<tr>
<td>F . .</td>
<td>1056</td>
<td>1030</td>
<td>45.40</td>
</tr>
<tr>
<td>G . .</td>
<td>1004</td>
<td>982</td>
<td>37.14</td>
</tr>
</tbody>
</table>

Table 3c

NOTE.—Sequence data have been calculated from the E tree and are analyzed by evolutionary parsimony. Details of the probabilities of substitution used for the calculations are provided in the legends to figs. 8-10. The results shown in 3a have been devised so that the incorrect, F, tree is favored by a factor of 10 over the correct, E, tree. In 3b the effect of both high transition and high transversion rates for branches 1 and 3 is to favor the incorrect F tree as being most parsimonious. In 3c the operator invariant corresponding to the correct tree has a negative value, and, as in 3a, both alternative trees are more parsimonious than the correct tree. In all three examples, evolutionary parsimony selects the correct tree.

* All values are for 50 df, for which values of 67.56, 71.49, 76.32, and 79.71 are significant at the 5%, 2.5%, 1%, and 0.5% levels, respectively.

mitochondrial origins (A. Wilson, personal communication). In his formalism, shown in figure 11, the endosymbiotic theory of mitochondrial origins is represented by the tree at the top of the figure, in which mitochondria and eubacteria are topologically nearest neighbors, as are eukaryotes and eocytes. On the other hand, if mitochondria had an autogenous origin—i.e., were derived from the eukaryotic nucleus—then the appropriate tree is the one at the bottom of the figure (the autogenous tree), in which mitochondria and eukaryotes are nearest neighbors, as are eubacteria and eocytes. In this test, eocytes, with their unusual sulfur-based metabolisms, represent a group clearly unrelated to the other three. With these sequences, tests of the mitochondrial-eukaryotic-eubacterial-eocyte tree shown in figure 11 were made. The significance plot indicates that these data support the endosymbiotic theory strongly. The $\chi^2$-values for the alternative trees are both within the expected range (i.e., <5%). In this instance, the same tree is predicted by the parsimony analysis (table 4), although a majority (71%) of the parsimony term has arisen from the background. This illustrates our recommendation that parsimony should be used with caution whenever either of the two alternative trees is supported by significant parsimony components.
Fig. 9.—Significance plot with reversible substitutions. Details are as in fig. 8 except that (1) the substitution probabilities are $P_a = P_A = P_c = P_C = 0.3$ and $P_b = P_B = P_d = P_D = 0.03$, (2) transversions and transitions may both occur reversibly, and (3) the correct solution, after 50 trees, is significant at approximately the 0.5% level.

Fig. 10.—Significance plot with the parsimony-like term less than the background term. $P_a = P_A = 0.9$; $P_b = P_B = P_c = P_C = P_d = P_D = 0.05$; and $P_e = P_E = 0.1$. Neither transitions nor transversions are reversible. The correct solution, after 50 trees, is significant at approximately the 2% level.
Hypothetical endosymbiotic origin of mitochondrion.

**ENDOSYMBIOTIC THEORY**

- Eubacteria
- Eukaryotic Nucleus
- Mitochondrion
- Eocytes

**AUTOGENOUS THEORY**

- Eukaryotic Nucleus
- Eubacteria
- Mitochondrion
- Eocytes

**Discussion**

The method of evolutionary parsimony has several properties that make it uniquely suited to probing deep divergences and shares many other desirable ones with parsimony.

The method is insensitive to differences in rates of substitution in different branches of the tree, since it analyzes the effects produced by operators rather than the probability that a substitution will occur in a particular branch of the tree. Even when parsimony will consistently predict the incorrect tree, evolutionary parsimony can determine the underlying phylogeny.

Evolutionary-parsimony analysis is insensitive to the varying relative rates of substitution that occur at different nucleotide positions. In the sequences of ribosomal RNAs, for example, substitution rates for nearby nucleotides can differ by one or two orders of magnitude; and similar differences can be found in mRNA coding sequences. Invariant analysis can analyze both of these sequence types well.
Evolutionary Parsimony 185

Table 4
Summary Data from a Test of the Endosymbiotic Theory of Plant Mitochondrial Origins

<table>
<thead>
<tr>
<th>Tree</th>
<th>Parsimony-like Term</th>
<th>Background Term</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endosymbiotic tree</td>
<td>979</td>
<td>702</td>
<td>38.21</td>
</tr>
<tr>
<td>Autogenous tree</td>
<td>593</td>
<td>637</td>
<td>10.78</td>
</tr>
<tr>
<td>Third alternative</td>
<td>293</td>
<td>375</td>
<td>19.08</td>
</tr>
</tbody>
</table>

NOTE.—Details of the calculation are provided in the legend to fig. 11. In this test the most parsimonious tree is also the most evolutionarily parsimonious (the endosymbiotic) tree. More than 70% of the contribution to the parsimony term, however, results from substitutions in peripheral branches.

* All values are for 16 df, for which values of 26.29, 28.81, 31.97, and 34.23 are significant at the 5%, 2.5%, 1%, and 0.5% levels, respectively.

Both evolutionary parsimony and parsimony share the property that they do not assume that relative branch lengths are necessarily conserved when adjacent sequence positions are compared. In fact, both methods only assume that all nucleotide positions share the same tree topology.

A significant difference between parsimony and evolutionary parsimony is that the latter is explicitly evolution based whereas parsimony (as emphasized by the transformed cladists; see, e.g., Patterson 1982) is not. Transformed cladists are very possibly correct in their claim that one can build trees without necessarily introducing any knowledge of the evolutionary process. While the most parsimonious tree will be correct many times, other times it will not. The method of evolutionary parsimony suggests that by explicitly introducing an evolutionary model one can obtain a more robust algorithm than parsimony. The cost is some loss of generality, but the gains are a greatly increased ability to determine evolutionary trees.

Finally, operator invariants are easy to calculate and require only slightly more computational effort than the parsimony method. The test of the endosymbiotic theory took only 45 s of real time with a time-sharing terminal accessing a VAX 11/780 computer.

The evolutionary assumptions of the theory of evolutionary parsimony seem to be in accord with current knowledge about the molecular-biological mechanisms of substitution. The relevant assumption is that transition and transversion operators are a reasonable approximation to biological reality. Clearly, the molecular details of transversions and transitions are extremely different, since, for some sequences, transitions can occur an order of magnitude more frequently than transversions (Brown et al. 1982). For this reason, a realistic model of evolution should consider these processes separately. Although the current knowledge of transition mechanisms does not provide any information concerning the frequency of the four transitions (C to T, T to C, A to G, and G to A), this does not affect the theory, since the equations themselves are rate invariant.

Even less is known about the molecular mechanisms that result in transversions. Hence, the mathematical treatment of transversions is couched in approximate terms. I have assumed that the transversion operator converts a vector to either one of the
two alternatives with equal probability. Thus, for example, the $A$ operator can convert the $E$ vector to either a $B$ or an $h$ and both alternatives are assumed to be equally likely. Results recently reported by Goodman and co-workers (Koop et al. 1986) indicate that this basic assumption is met. Specifically, they have measured the relative frequencies of the four transitions and the eight transversions in studies of eukaryotic protein-coding genes. They also provide figures for relative nucleotide usage. From these data I calculate that the $A$ operator acting on the $E$ unit vector has a 49.9% probability of converting it to a $B$ vector over a 50.1% probability of converting it to an $h$ vector. Similar probabilities are calculated for the other transversions as well. Thus, the invariants will have slight errors associated with them that would probably become significant only for extremely long sequences. The effect of this error in extreme cases would be to cause the two invariants associated with incorrect trees to have (statistically) nonzero values. For this reason, an essential part of the theory of evolutionary parsimony is to test whether the null hypothesis is obeyed for the two less favored trees. By performing this comparison one is independently testing the evolutionary assumptions on which the theory is based. (The comparable test when using parsimony would be to test the two components supporting the less parsimonious trees to see whether they are statistically zero; if not, one should consider the results of the parsimony analysis suspect.) Since evolutionary parsimony can be extended to include other choices for substitution operators, it is possible that other invariants might be found for those cases when the null hypothesis is not valid for two or more of the equations. However, the transition/transversion operators used in this paper are the most useful set that I have yet found and seem to be quite generally valid.

Finally, one needs a few practical rules for using the method. The usual rules that apply to use of the $\chi^2$-test also are needed for this method. There are two types of failure. In the first type the wrong tree can be indicated if the $\chi^2$-value for an incorrect tree is too large. If a 95% significance level is used, this can happen, on average, one time in 20 for each tree considered. In the second type, the correct tree is not chosen because its $\chi^2$-value is too small. The first type of error can be reduced by setting the rejection levels higher (I recommend 2% or even 1% as the one-sided upper cutoff and 0.5% or perhaps an even smaller one-sided lower cutoff). Errors of the second type can be reduced by including more data or by considering data that have the largest possible signal:noise ratio.

In general, signal:noise ratios can be increased by considering two factors. The first factor is that the most reliable predictions are obtained when the central branch is as long as possible. This is because the central branch is the source of the signal. For example, in the test of mitochondrial origins proposed in this paper, sequences from three groups (the halobacteria, the methanogens, or the eocytes) could have been used to define the fourth branch of the tree. Eocytes were selected because they appear to be the closest neighbors to the eukaryotes and hence provide a longer central branch than do the two other groups. (When halobacteria are the fourth branch of the tree, no tree topology is favored because the central branch is too short. When methanogens are used, the endosymbiotic theory is supported—but not as strongly as when eocytes are used.) The second factor is that low substitution rates in the peripheral branches of a tree produce a low background term and provide the best conditions for detecting a weak signal. For example, if one has available sequences from two sister taxa, one rapidly evolving and one slowly evolving, the sequence from the more slowly evolving taxon should be used for phylogenetic analysis when trees having short central branches are to be detected. Thus, the sensitivity and accuracy of the method of evolutionary
parsimony can, to some extent, be maximized by carefully choosing organisms to obtain the longest central branch and shortest peripheral branches.

The technique of evolutionary parsimony is a robust algorithm for probing deep evolutionary branchings. It should extend the range of phylogenetic analysis and make possible investigations of distant evolutionary branchings that were formerly impossible.

Acknowledgments

I thank Roderick Little for suggestions on statistical tests of the basic equations and Michael Ghil for suggesting the investigation of Markov techniques. These ideas have been greatly influenced by many colleagues. Walter Fitch influenced this paper in profound ways during our many conversations on parsimony methods, and the ideas on statistical tests of trees are patterned after those first described to me by Allan Wilson. I also thank Joe Felsenstein for suggestions on a draft of this paper and for sharing his results in advance of their publication (Cavender and Felsenstein, in press). I thank James Cavender for many helpful suggestions—and, in particular, for pointing out the usefulness of the dot product. This work was supported by grants from the National Science Foundation (PCM 76-14718) and the National Institute of General Medical Sciences (GM24034) to the author.

APPENDIX

Given the rules of operator algebra listed in table 1, the proof of the central equations of evolutionary parsimony proceeds iteratively. First note that the sequence vector $\mathbf{S}$ is the sum of $N$ vectors, each of length 1, that correspond to single positions in a sequence of length $N$. If one can show that equations (8a), (8b), and (8c) are valid for each of the subcomponents of $\mathbf{S}$, then they must be valid for $\mathbf{S}$, their vector sum. Hence, the proof is for a single subcomponent vector of length 1, corresponding to a single position in sequence S.

For purposes of calculation I assume the rooted tree in figure 6. Although this tree is rooted (at the center of the interior branch), results calculated from it will be valid for all rootings, since the substitution operators are reversible (e.g., see Felsenstein 1983). For the following calculations, I also rely on the operator relationships defined in table 1.

Transition operators have the following property: they can only convert $\tilde{E}$ into one of four vectors contained in the set \{E, $\hat{u}$, H, or J\}. Let $\mathbf{V}_{X}$ represent any vector in this set and let $\mathbf{V}_{NX}$ represent any of the 32 remaining basis vectors not in that set. Then, the effect of transitions on the $\mathbf{V}_{X}$ is to shuffle them as shown in equation A1:

$$
\begin{bmatrix}
E \\
H \\
\hat{u} \\
J \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
E \\
H \\
\hat{u} \\
J \\
\end{bmatrix}
$$

The effect of transitions on the $\mathbf{V}_{NX}$ is also to shuffle them within their own sets. Hence no transition can convert any $\mathbf{V}_{NX}$ into any $\mathbf{V}_{X}$, and vice versa. Since all of the $\mathbf{V}_{NX}$ are perpendicular to $\mathbf{X}$ and none of the $\mathbf{V}_{X}$ are perpendicular to $\mathbf{X}$, the following relationships are obeyed:

$$
\begin{align*}
\mathbf{X} \cdot \mathbf{V}_{X} &= 1; \\
\mathbf{X} \cdot \mathbf{V}_{NX} &= 0.
\end{align*}
$$

Transition operators shuffle the vectors so that (where $o$ is any transition operator)
\[
\mathbf{X} \cdot o(\mathbf{V}_X) = -\mathbf{X} \cdot \mathbf{V}_X; \tag{A4}
\]
\[
\mathbf{X} \cdot o(\mathbf{V}_{NX}) = 0 \tag{A5}
\]

In contrast, transversion operators, \(O\), do not shuffle components within sets but instead transform all subcomponents of \(\mathbf{S}\) into vectors perpendicular to \(\mathbf{X}\). The following examples illustrate this point: Inspection of table 1 shows that 28 of the 36 basis vectors are converted into vectors that have no components in the set of \(\mathbf{V}_X\). The remaining basis vectors—\(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{h}, \mathbf{i}, \mathbf{j}, \text{ and } \mathbf{k}\)—merit an explicit calculation. Consider the operation \(A(\mathbf{B})\). Reference to table 1 shows that \(A(\mathbf{B}) = 0.5\mathbf{E} + 0.5\mathbf{H}\), where the stochastic relationship (i.e., that both of the outcomes of the transversion operators are equally likely) has been explicitly assumed. Hence \(\mathbf{X} \cdot A(\mathbf{B}) = 0.5 - 0.5 = 0\). Similar relationships hold for the seven other related basis vectors. Thus, for all 36 basis vectors one has
\[
\mathbf{X} \cdot O(\mathbf{V}_X) = 0; \tag{A6}
\]
\[
\mathbf{X} \cdot O(\mathbf{V}_{NX}) = 0. \tag{A7}
\]

Thus, in general, operators transform vectors perpendicular to the \(\mathbf{X}\) direction. Only those vectors that initially at time \(t_1\) projected onto \(\mathbf{X}\) can continue to point in the direction of \(\mathbf{X}\) at later times. Even then, some of those special subcomponents will be transformed into vectors perpendicular to \(\mathbf{X}\) by transversions. Only those that have been altered exclusively by transitions (or are unaltered) will not be perpendicular to \(\mathbf{X}\). Hence, the only way in which a component can project onto \(\mathbf{X}\) at the present time is to have projected in the \(\mathbf{X}\) direction at time \(t_1\) (see fig. 6). (The same arguments can be derived for \(\mathbf{Y}\) and \(\mathbf{Z}\) by permuting sequences 2 and 3 or 2 and 4, respectively.) The proof is completed by noting that only components \(\mathbf{O}, \mathbf{e}, \text{ and } \mathbf{E}\) existed at time \(t_1\). Hence, for any vector subcomponent of \(\mathbf{S}, \mathbf{V}\), one has for all times after \(t_1\) (where \(\neq\) indicates “not necessarily equal to”)
\[
\mathbf{X} \cdot \mathbf{V} \neq 0; \tag{A8}
\]
\[
\mathbf{V} \cdot \mathbf{V} = 0; \tag{A9}
\]
\[
\mathbf{Z} \cdot \mathbf{V} = 0. \tag{A10}
\]

Since these relationships are true for all the subcomponents of \(\mathbf{S}\), the three equations (A8, A9, and A10) apply for the entire sequence vector, \(\mathbf{S}\).

**Note Added in Proof** (in collaboration with Roderick J. Little [Department of Biomathematics, University of California, Davis]): In the event that each tree does not represent an independent measurement, the significance may be calculated by estimating the correlations from the data. We first consider \(n\) variants sampling a single branch and then generalize to variants in all four branches.

\(\chi^2\)-Test from Correlated Data on \(n\) Variants

**Notation:**

\[
X_{PE}^{(j)} = \text{count for parsimony-like term, topology } E, \text{ variant } j (j = 1 \ldots n);
\]
\[
\mu_{PE} = \text{mean of } X_{PE}^{(j)} \text{ (assumed same for all } j); 
\]
\(X_{BE}^{(j)}\) = count for background term, topology E, variant \(j\);

\(\mu_{BE}\) = mean of \(X_{PE}^{(j)}\) (assumed same for all \(j\)).

Also define:

\[d_j = X_{PE}^{(j)} - X_{BE}^{(j)}\]

\[s_j = X_{PE}^{(j)} + X_{BE}^{(j)}\]

\[\bar{d} = \frac{\sum_{j=1}^{n} d_j}{n}; \quad \bar{s} = \frac{\sum_{j=1}^{n} s_j}{n};\]

\[s_d^2 = \frac{\sum_{j=1}^{n} (d_j - \bar{d})^2}{(n - 1)} = \text{sample variance of } d_j's \text{ over variants}.\]

Testing Hypothesis

To test the hypothesis \(H_0: \mu_{PE} = \mu_{BE}\), compute

\[\chi^2 = \frac{nd^2}{\bar{s}(1 + [n - 1]\hat{\rho})},\]  \hfill (1)

where

\[\hat{\rho} = 1 - \frac{s_d^2}{\bar{s}} = \text{estimated correlation between } d_i \text{ and } d_k.\]  \hfill (2)

Derivation

Assume

\[E(d_j) = \mu_{PE} - \mu_{BE};\]

\[\text{Var}(d_j) = \mu_{PE} + \mu_{BE} (\text{Poisson assumption for counts});\]

\[\text{Cov}(d_j, d_k) = \rho(\mu_{PE} + \mu_{BE}) (\text{constant correlation } \rho \text{ between variants});\]

then

\[E(\bar{d}) = \mu_{PE} - \mu_{BE} = 0 \text{ under } H_0, E(\bar{s}) = \mu_{PE} + \mu_{BE};\]

\[\text{Var}(\bar{d}) = \frac{(\mu_{PE} + \mu_{BE})}{n} (1 + [n - 1]\rho).\]

The estimate of the variance is then given by

\[\text{Var}(\bar{d}) = \frac{\bar{s}}{n} (1 + [n - 1]\hat{\rho}).\]
190 Lake

Estimate of $\hat{\rho}$

It can be easily shown that

$$E(s^2_\theta) = (1 - \rho)(\mu_{PE} + \mu_{BE}).$$

So

$$\rho = 1 - \frac{E(s^2_\theta)}{E(s^2)}$$

This suggests the simple estimate

$$\hat{\rho} = 1 - \frac{s^2_\theta}{\tilde{s}}.$$

More efficient estimates could be derived, based on maximum likelihood. They seem to be much more complex, however.

Extension to Simultaneous Variation in All Branches

In the event that variants are used in all four branches, the hypothesis can be tested by the following $\chi^2$-test for 1 df: To test the hypothesis $H_0$, compute

$$\chi^2_1 = nd^2/\tilde{s},$$

where

$$c = 1 + \sum_j (n_j - 1)p_j + \sum_{j<k} (n_j - 1)(n_k - 1)p_{jk} + \sum_{j<k<l} z(n_j - 1)(n_k - 1)(n_l - 1)p_{jkl},$$

where $n_j$ is the number of variants used for branch $j$, for $j = 1, \ldots, 4$; $p_j$ is the correlation between variants of a branch $j$, holding branches $k$, $l$, and $m$ fixed; $p_{jk}$ is the correlation between variants of branch $j$ and $k$, holding branches $l$ and $m$ fixed; and so on. If we make the plausible assumptions that $p_{jk} = p_j p_k$ for all $j < k$; $p_{jkl} = p_j p_k p_l$ for all $j < k < l$; and $p_{1234} = p_1 p_2 p_3 p_4$, then eq. (N4) reduces to the simpler form

$$c = (1 + [n_1 - 1]p_1) \ldots (1 + [n_4 - 1]p_4).$$

If either eq. (N4) or eq. (N5) is used, the correlations can be calculated directly from the pairwise comparison of variants by means of eq. (N2).

LITERATURE CITED


W. M. FITCH, reviewing editor

Received October 2, 1986; revision accepted November 6, 1986.