# Mechanized Derivation of Linear Invariants[1]

*James A. Cavender*

Department of Mathematics, University of Colorado at Denver

Linear invariants, discovered by Lake, promise to provide a versatile way of inferring phylogenies on the basis of nucleic acid sequences (the method that he called "evolutionary parsimony"). A semigroup of Markov transition matrices embodies the assumptions underlying the method, and alternative semigroups exist. The set of all linear invariants may be derived from the semigroup by using an algorithm described here. Under assumptions no stronger than Lake's, there are >50 independent linear invariants for each of the 15 rooted trees linking four species.

## Introduction

The recent discovery by Lake (1987) of linear invariants promises to provide an extraordinarily versatile method of inferring phylogenies from nucleic acid sequences while providing naturally for statistical testing of phylogenies as hypotheses.

The method is particularly recommended by the nature of the assumptions that underlie it. These are not philosophical principles but sharply defined scientific hypotheses about the observable relative rates of replacement of particular nucleotides by others. Indeed, linear invariants themselves provide a method of testing these hypotheses. Most important, the assumptions are mild without precedent, allowing a different free choice from a large family of substitution probabilities for every branch of the tree of evolution and every position in the molecule.

In this paper, I present the theory of linear invariants in conventional mathematical language, describe a mechanical method for generating linear invariants, and expand the number of known independent linear invariants for a given tree from two to >50.

## Invariants

The problem to be treated is to distinguish among the 15 phylogenetic trees of figure 1. For data, homologous strings of RNA are provided for the four tip species, A, B, C, and D. Each of these strings is regarded as a sequence of letters from the alphabet {A, G, C, U}. Deletions, rearrangements, etc., are outside the theory; evolution proceeds by substituting letters for other letters. Substitutions at different positions on the string are independent random events. At a given position at any time in the course of evolution, the usual Markov property applies: what happens next may be influenced by the current state but is otherwise independent of the past. (While this independence assumption is probably harmless, the influence of coevolving organisms and of other changing environmental conditions could theoretically render it false.) A change from A to G, from G to A, from C to U, or from U to C is called

Key words: numerical taxonomy, nucleic acid sequence, linear invariants, evolutionary parsimony.

Address for correspondence and reprints: James A. Cavender, Department of Mathematics, University of Colorado at Denver, Campus Box 170, 1200 Larimer Street, Denver, Colorado 80204.

*Mol. Biol. Evol.* 6(3):301–316. 1989.

0737-4038/89/0603-0007$02.00

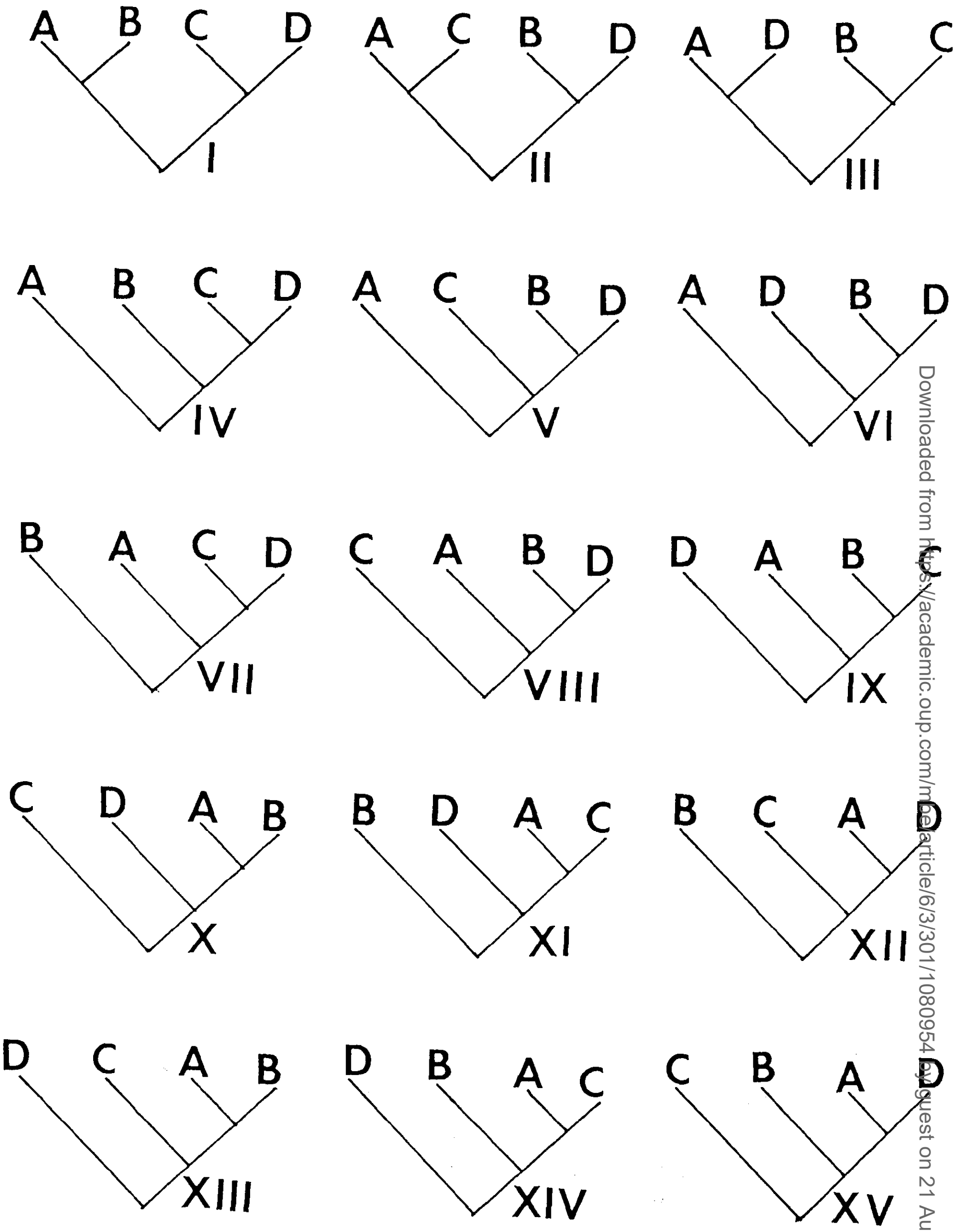

FIG. 1.—The 15 topologies on which four species can be placed. On each tree the lowest point is the root (the last common ancestor of the four species).

a *transition.* The other eight possible changes are *transversions.* (When A goes to U and then to G, it will be mathematically convenient to call the whole compound transaction a "transition from A to G.")

In this paper, I will also treat the problem of classifying five species. The way to extend the method to other numbers of species will become obvious.

To understand the results, if not the methods, presented here, one only needs the simplest ideas of Markov processes (Kemeny et al. 1974, pp. 137–144, 153–184, 203–214) and the ideas of basis and subspace, from linear algebra (Smith 1983, pp. 178–185).

For a person wishing to solve this problem, nothing could be more desirable than a function of the data whose value depends on the topology and on no other attribute of the tree (such as branch lengths and times of divergence). Since there are degenerate trees with more than one topology, one cannot hope for a single function taking 15 separate values for the 15 topologies. The best that can be expected are functions that take some constant value under some topologies but do not necessarily always take this value under the others. Such functions can reasonably be called *invariants* because their values remain unchanged over the course of evolution and because the known examples (Cavender and Felsenstein 1987) are actually invariants of algebraic varieties.

Of course, rare, chance events can always cause data from different topologies to be identical. Thus, the definition of invariant must be framed to account for randomness. There are at least three possibilities: First, one might say that a function of the expected value of the data is an invariant of a topology if for that topology there is only one value that it can take. (To be useful for discriminating topologies, it must also take some other value at least sometimes for at least one other topology. I do not add this proviso to the definition because an invariant can also be useful in other ways.) This is what I and my coauthor meant by "invariant" in a previous article (Cavender and Felsenstein 1987). Second, one could require the expected value of the function to be constant when it is applied to the random data. For linear functions of the data, this stronger definition is actually equivalent to the first one. The definition that I shall make is formally of this second type. Third, one might say a function of the data is an invariant if its distribution, rather than just the mean, is constant for all trees of the given topology (J. Felsenstein, personal communication).

At each position in the RNA sequence, there is an assignment of a letter A, G, C, or U to each species A, B, C, and D. The notation AUUG for a pattern means species A, B, C, and D have letters A, U, U, and G, respectively, at this position. There are 256 such patterns possible. The 256-dimensional vector that contains the 256 observed frequencies of the patterns is called the *spectrum* of the RNA sequence (Cavender 1978; Lake 1987). It is a complete summary of the data for the inference problem.

It is convenient to let AUUG also denote the spectrum that represents nothing but a single occurrence of the pattern AUUG, i.e., a vector with a one and 255 zeros. Using this notation, we define a spectrum $Y$ as

$$\begin{aligned} Y = {} & \mathrm{ACAC + AUAU + GCGC + GUGU + CACA + CGCG + UAUA} \\ & \mathrm{{}+ UGUG - ACAU - AUAC - GCGU - GUGC - CACG - CGCA} \\ & \mathrm{{}- UAUG - UGUA - ACGC - AUGU - GCAC - GUAU - CAUA} \\ & \mathrm{{}- CGUG - UACA - UGCG + ACGU + AUGC + GCAU + GUAC} \\ & \mathrm{{}+ CAUG + CGUA + UACG + UGCA}\,. \end{aligned}$$

Let $S$ be an observed spectrum and let a function $y$ be defined by the dot product $y(S) = Y \cdot S$. Lake (1987) showed that, under assumptions, $y$ is an invariant, with $y = 0$, of the topologies in the left and right columns of figure 1. Symmetrically, he exhibited both a vector $X$ that gives an invariant $x$ of the middle and right columns and a $Z$ that gives an invariant $z$ of the left and middle columns. Henceforth, I shall call $Y$, rather than $y$, the "invariant." I make this a definition. The notation $E[H]$ denotes the expected value of the random variable $H$.

Definition: A linear invariant of a topology is a vector $V$ such that $E[V \cdot S] = 0$ for spectra $S$ under that topology.

If $U$ and $W$ are two vectors with $E[U \cdot S]$ and $E[W \cdot S]$ both equal to nonzero constants, then it is easily shown that **U** can be obtained from **W** by adding an invariant and multiplying by a constant. Therefore, because there is essentially only one of them, vectors such as these are excluded from the definition. The archetype is $(1, 1, \ldots, 1)$, which only returns your sample size to you.

Restrict attention to a single position in the molecule. Its spectrum $S$ will comprise 255 zeros and a one. Its expected spectrum is more interesting, with 256 nonnegative numbers summing to 1. If a vector $V$ is an invariant for each position in the molecule, i.e., if $E[S_i \cdot V] = 0$, where $i$ indexes positions in the molecule, then $E[\sum S_i \cdot V] = n0 = 0$, where $n$ is the size of the sample of positions; so an invariant good for each position is good for the whole sample. This is true whether or not the spectra for different positions are statistically independent, but I retain the assumption of independence to justify statistical tests.

If $Y \cdot S$ is significantly far from zero, where $S$ is the spectrum of the whole sample, then topologies I and III can both be rejected as hypotheses. Lake splits $Y$ into two parts, $Y = Y^+ - Y^-$, where

$$\begin{aligned} Y^+ = {} & \mathrm{ACAC} + \mathrm{AUAU} + \mathrm{GCGC} + \mathrm{GUGU} + \mathrm{CACA} + \mathrm{CGCG} \\ & + \mathrm{UAUA} + \mathrm{UGUG} + \mathrm{ACGU} + \mathrm{AUGC} + \mathrm{GCAU} + \mathrm{GUAC} \\ & + \mathrm{CAUG} + \mathrm{CGUA} + \mathrm{UACG} + \mathrm{UGCA} \end{aligned}$$

and

$$\begin{aligned} Y^- = {} & \mathrm{ACAU} + \mathrm{AUAC} + \mathrm{GCGU} + \mathrm{GUGC} + \mathrm{CACG} + \mathrm{CGCA} \\ & + \mathrm{UAUG} + \mathrm{UGUA} + \mathrm{ACGC} + \mathrm{AUGU} + \mathrm{GCAC} + \mathrm{GUAU} \\ & + \mathrm{CAUA} + \mathrm{CGUG} + \mathrm{UACA} + \mathrm{UGCG}\,, \end{aligned}$$

so that the hypothesis $Y \cdot S = 0$ says $Y^+ \cdot S = Y^- \cdot S$. Both sides of this equation are merely counts of positions in the molecule. Lake tests this hypothesis with the well-known $\chi^2$ test for the equality of two events (van der Waerden 1969, pp. 40–42). However, $Y^+ \cdot S$ and $Y^- \cdot S$ are dependent random variables, so this is not a valid application of that test. The marginal distribution of $Y^+ \cdot S$, given $(Y^+ + Y^-) \cdot S$, is binomial with $n = (Y^+ + Y^-) \cdot S$, and a test of whether $p = ½$ in this binomial is a valid test of whether $E[Y \cdot S] = 0$ (Holmquist et al. 1988; David Clair, personal communication).

Lake's two invariants, $Y$ and $Z$, for topology I are basis vectors for a two-dimensional space of invariants. That is, the space consists of linear combinations of these two and of nothing more. Are all linear invariants in this space? Not at all. Under assumptions no stronger than Lake made, the subspace of invariants has 68 dimensions (for the symmetric trees I–III) or 54 dimensions (for all the others). Since many choices of assumptions are possible, it is important to have an algorithm that can generate the invariant subspace from given assumptions. These assumptions can be characterized in terms of families of Markov matrices.

## Markov Matrices

Choose a single position in the molecule and restrict attention to it. Consider two species, $\underline{\mathrm{R}}$ and $\underline{\mathrm{S}}$, anywhere on the tree. Define $p_{\mathrm{AG}}$ to be the conditional probability of G being in the chosen position at $\underline{\mathrm{R}}$, given that A is in that position at $\underline{\mathrm{S}}$. For each other pair of letters from the set { A, G, C, U }, define another $p$ similarly. The *Markov matrix* from $\underline{\mathrm{R}}$ to $\underline{\mathrm{S}}$ is

$$P_{\underline{\mathrm{RS}}} = \begin{pmatrix} p_{\mathrm{AA}} & p_{\mathrm{AG}} & p_{\mathrm{AC}} & p_{\mathrm{AU}} \\ p_{\mathrm{GA}} & p_{\mathrm{GG}} & p_{\mathrm{GC}} & p_{\mathrm{GU}} \\ p_{\mathrm{CA}} & p_{\mathrm{CG}} & p_{\mathrm{CC}} & p_{\mathrm{CU}} \\ p_{\mathrm{UA}} & p_{\mathrm{UG}} & p_{\mathrm{UC}} & p_{\mathrm{UU}} \end{pmatrix}. \qquad (1)$$

The usual term is *Markov transition matrix,* but that would be confusing here. As is usual in Markov chain theory, the probability mass function at $\underline{\mathrm{R}}$ can be written as a row vector and multiplied by $P_{\underline{\mathrm{RS}}}$, with the matrix on the right, to get a row vector of probabilities at $\underline{\mathrm{S}}$. I shall use only Markov matrices that point forward in time from an earlier species $\underline{\mathrm{R}}$ to a later species $\underline{\mathrm{S}}$. If $\underline{\mathrm{R}}$ evolves into $\underline{\mathrm{T}}$ which evolves into $\underline{\mathrm{S}}$, then $P_{\underline{\mathrm{RS}}} = P_{\underline{\mathrm{RT}}}P_{\underline{\mathrm{TS}}}$. Thus, when I admit a set of matrices to my model, I shall be forced to also admit all products of members of that set.

## A Semigroup of Markov Matrices

For each of the six branches of the tree and for each position in the molecule, there is a Markov matrix. Thus, if there are 1,000 positions, there will be 6,000 different matrices involved. This complexity can be brought under control by treating just one position and extending the invariants to the whole sample by linearity. There will be no linear invariants unless some restriction is placed on the Markov matrices. I shall specify a set of Markov matrices that satisfies a consistency property (multiplicative closure) and shall derive invariants from this set. Since every assumption about the process of evolution is potentially an error, I want to make this set as large as possible.

In particular, I shall use Markov matrices of the form

$$\begin{pmatrix} e & f & g & g \\ h & i & j & j \\ k & k & l & m \\ n & n & p & q \end{pmatrix}. \qquad (2)$$

These matrices state as generally as possible Lake's assumption that, when a transversion occurs, the two possible outcomes are equally probable. Let $\mathcal{H}$ be the set of all matrices of this form.

The product of two matrices from $\mathcal{H}$ is sometimes outside of $\mathcal{H}$ (i.e., $\mathcal{H}$ is not "multiplicatively closed"), and this would be a serious inconsistency if invariants were derived from $\mathcal{H}$. Thus, it is important to know that the union $\mathcal{L}$ of all multiplicatively closed subsets of $\mathcal{H}$ is the largest multiplicatively closed subset of $\mathcal{H}$ and is characterized by the additional constraints

$$e - f = i - h \qquad (3)$$

and

$$l - m = q - p\,. \qquad (4)$$

By multiplying two general matrices of the form of matrix (2) and insisting that $p_{AC} = p_{AU}$ in the product, one quickly sees the necessity, for multiplicative closure, of equation (4). Similarly, equation (3) follows from $p_{CA} = p_{CG}$. Sufficiency is established through a prodigy of high school algebra: multiply two general members of $\mathcal{L}$ and verify that the product is in $\mathcal{L}$.

A multiplicatively closed set of matrices such as $\mathcal{L}$ is a semigroup of matrices. Unlike some semigroups, $\mathcal{L}$ is defined almost entirely by linear constraints on the components of the matrices. (The exception is the requirement that a Markov matrix have no negative components.) This makes it only slightly different from an algebra of matrices, so something of use to taxonomy may be present in the extensive theory of algebras (Albert 1937, pp. 217–250, 1961; Deuring 1968).

Every matrix in $\mathcal{L}$ is a linear combination $\lambda^1 p_1 + \lambda^2 p_2 + \cdots + \lambda^7 p_7$ (and here the superscripts are not exponents) of the seven matrices

$$p_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$p_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

$$p_3 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

$$p_4 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$p_5 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

$$p_6 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix},$$

$$p_7 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Conversely, every linear combination of these seven that is a stochastic matrix (i.e., has rows summing to 1 and lacks negative components) is in $\mathcal{L}$. Derivation of a basis such as $\{p_1, p_2, \ldots, p_7\}$ from linear constraints of the sort that define $\mathcal{L}$ is routine. One regards the elements of $\mathcal{L}$ as 16-dimensional vectors. The constraint $p_{\mathrm{AC}} = p_{\mathrm{AU}}$ says that these vectors are all orthogonal to a particular vector—$x_1$, say. Five more such vectors, $x_2, x_3, \ldots, x_6$, are apparent. To ensure that the row sums are equal, the constraints

$$p_{\mathrm{AA}} + p_{\mathrm{AG}} + p_{\mathrm{AC}} + p_{\mathrm{AU}} = p_{\mathrm{GA}} + p_{\mathrm{GG}} + p_{\mathrm{GC}} + p_{\mathrm{GU}} ,$$

$$p_{\mathrm{AA}} + p_{\mathrm{AG}} + p_{\mathrm{AC}} + p_{\mathrm{AU}} = p_{\mathrm{CA}} + p_{\mathrm{CG}} + p_{\mathrm{CC}} + p_{\mathrm{CU}} ,$$

and

$$p_{\mathrm{AA}} + p_{\mathrm{AG}} + p_{\mathrm{AC}} + p_{\mathrm{AU}} = p_{\mathrm{UA}} + p_{\mathrm{UG}} + p_{\mathrm{UC}} + p_{\mathrm{UU}}$$

must also be treated. (If the row sums equal something other than 1, there will be no harm in it.) These last three constraints give $x_7$, $x_8$, and $x_9$. To find matrices such as $p_1, p_2, \ldots, p_7$ is to find seven independent 16-dimensional vectors that are each orthogonal to all of $x_1, x_2, \ldots, x_9$. A matrix inversion algorithm, for example, will do this.

## Generating Invariants

Assume all Markov matrices are drawn from $\mathcal{L}$ and continue to restrict attention to one position in the molecule. Let $\mathscr{H}$ be the smallest subspace of 256-dimensional space that contains all expected spectra that can arise. The subspace of linear invariants is the set of all vectors that are orthogonal to $\mathscr{H}$. (A vector is orthogonal to $\mathscr{H}$ if it is orthogonal to every vector in $\mathscr{H}$, and this is true if it is orthogonal to every vector in some basis for $\mathscr{H}$.) The computation of a basis for the space of invariants from a basis for $\mathscr{H}$ is exactly like the computation of the matrices $p_1, \ldots, p_7$ from the constraining vectors $x_1, \ldots, x_n$. Thus, it suffices to compute a basis for $\mathscr{H}$. This leads into some discussion of computational shop technique.

At the heart of my computer programs is a "triangularizer" subroutine. It maintains a list of linearly independent vectors. When a new vector $h$ is presented to the triangularizer, it determines whether $h$ can be written as a linear combination of vectors already in the list. If not, the list is enlarged so that it can be. It is useful for efficiency to keep the list in row-echelon form, so this enlargement is a more complicated process than merely adding $h$ to the list. Round-off error would be disastrous to this process, so everything must be done in integer arithmetic. In particular, $h$ is a vector of integers—not an expected spectrum at all, but a multiple of one. If I throw enough vectors $h$ into this subroutine, the list becomes a basis for $\mathscr{H}$, for the span of the list is the span of the vectors thrown in. The problem is to generate enough vectors $h$ so that every expected spectrum is a combination of some of them.

For concreteness, assume topology IV. Let the Markov matrix on branch $a$ also be called $a$. Its components are of the form $a_{\mathrm{XY}}$ where X, Y $\in \{$A, G, C, U$\}$. Name the Markov matrices on the other branches $b$, $c$, $d$, $e$, and $f$. The probability of pattern AUUG, e.g., is

$$P\{\mathrm{AUUG}\} = \sum_{\mathrm{R}} \sum_{\mathrm{S}} \sum_{\mathrm{T}} r_{\mathrm{R}} a_{\mathrm{RA}} f_{\mathrm{RS}} b_{\mathrm{SU}} e_{\mathrm{ST}} c_{\mathrm{TU}} d_{\mathrm{TG}} , \qquad (5)$$

where R, S, and T, all elements of $\{A, G, C, U\}$, are possible states of interior nodes as marked in figure 2 and where $r_R$ is the probability of R in the ancestor. [I never constrain the initial distribution $r = (r_A, r_G, r_U, r_C)$.] There are analogous formulas for the other 255 patterns. Thus, you could compute the expected spectrum if you knew the matrices $a, b, c, d, e, f$ and the initial distribution $r$. You could generate all possible expected spectra if you knew all possible $a, b, c, d, e, f,$ and $r$. The formula (5) for $P\{AUUG\}$, with its 255 analogues, generalizes equation (1) of Cavender (1978) and equation (2.3) of Tavaré (1986, p. 59).

Give the name $\sigma$ to the function that assigns an expected spectrum $h = (P\{AAAA\}, P\{AAAG\}, \ldots, P\{AUUG\}, \ldots, P\{UUUU\})$ to a set of matrices $a, b, c, d, e,$ and $f$ and an initial distribution $r$. That is, $h = \sigma(a, b, c, d, e, f, r)$. Then $\sigma$ is a multilinear. That is, $\sigma(a, \lambda b_1 + \mu b_2, c, d, e, f, r) = \lambda\sigma(a, b_1, c, d, e, f, r) + \mu\sigma(a, b_2, c, d, e, f, r)$, and similarly for $a, c, d, e, f,$ and $r$. This follows from formula (5).

Let $r_1 = (1, 0, 0, 0)$, $r_2 = (0, 1, 0, 0)$, $r_3 = (0, 0, 1, 0)$, and $r_4 = (0, 0, 0, 1)$. Then every initial distribution is a linear combination of vectors from $\{r_1, r_2, r_3, r_4\}$. Let each set $\{a_1, a_2, \ldots, a_7\}, \{b_1, b_2, \ldots, b_7\}, \ldots, \{f_1, f_2, \ldots, f_7\}$ be equal to $\{p_1, p_2, \ldots, p_7\}$, the set of seven matrices that spans $\mathcal{L}$. Then the set $\{\sigma(a_i, b_j, c_k, d_l, e_m, f_n, r_q)\}$, where $q \in \{1, \ldots, 4\}$ and $i, j, k, l, m, n \in \{1, \ldots, 7\}$, of $4 \cdot 7^6$ (470,596) spectra is a spanning set for the span $\mathscr{H}$ of all possible expected spectra. In fact, every expected spectrum $h$ can be written

$$\begin{aligned} h &= \sigma(a, b, c, d, e, f, r) \\ &= \sigma\Big(\sum_i \lambda_a^i a_i, \sum_j \lambda_b^j b_i, \sum_k \lambda_c^k c_k, \ldots, \sum_q \lambda_r^q r_q\Big) \\ &= \sum_i \lambda_a^i \sigma\Big(a_i, \sum_j \lambda_b^j b_j, \sum_k \lambda_c^k c_k, \ldots, \sum_q \lambda_r^q r_q\Big) \\ &= \sum_i \sum_j \lambda_a^i \lambda_b^j \sigma\Big(a_i, b_j, \sum_k \lambda_c^k c_k, \ldots, \sum_q \lambda_r^q r_q\Big) \\ &= \sum_i \sum_j \sum_k \sum_l \sum_m \sum_n \sum_q \lambda_a^i \lambda_b^j \lambda_c^k \lambda_d^l \lambda_e^m \lambda_f^n \lambda_r^q \sigma(a_i, b_j, c_k, \ldots, r_q), \end{aligned}$$

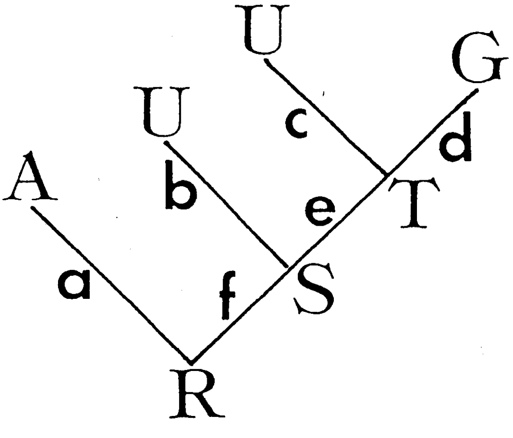

FIG. 2.—Pattern AUUG in topology IV. The edge labels may be identified with Markov matrices, which are generalized branch lengths. Capital letters are nucleotide names rather than species names.

a combination of elements from the set of 470,596. These 470,596 are the vectors $h$ that I feed to the triangularizer.

This approach is traditional. The ordered sets $(a_i, b_j, c_k, \ldots, r_q)$ are tensors that convert the multilinear map $\sigma$ into a linear map whose domain is a tensor product of six copies of the seven-dimensional space $\mathcal{L}$ and the four-dimensional space of initial distributions. $\mathscr{H}$ Is the image of this linear map (see Greub 1967, pp. 1–12, 19; Dieudonné and Carrell 1971, pp. 1–7). Lake's approach was different. He noted that $a$, $b$, $c$, and $d$ have natural actions on the space of expected spectra. That is, the 4 × 4 matrices extend to 256 × 256 matrices. Then $\mathscr{H}$ is stable (Serre 1977, pp. 1–7) under these larger matrices, which Lake calls "operators."

Figure 3 shows the 54 independent linear invariants of topology IV that result from this procedure. Figure 4 gives 68 invariants of topology I. If you cleverly assign and reassign the names A, B, C, and D to your four species, you can avoid treating any topologies besides I and IV.

## An Improved Algorithm

The procedure just described takes many hours on a dedicated, fast computer. Most of this time is spent in the triangularizer. If the algorithm as it stands were applied to five species, the triangularizer would be handling spectra four times as long and 49 times as many of them. Months would be needed. Fortunately, most of this computation can be avoided by computing five-species (1,024-dimensional) expected spectra $h$ from the 202 (or 188) basis vectors already identified for the four-species $\mathscr{H}$.

The idea is to graft a minimal, two-species tree onto a tree for which $\mathscr{H}$ is already known (fig. 5). Let $q^{\mathrm{U}}_{\mathrm{CA}}$ be the conditional probability of A and C in the species shown in the figure, given U at the join. Define 63 more $q^{\mathrm{R}}_{\mathrm{XY}}$ in the same way for other X, Y, and R ∈ {A, G, C, U}. Clearly, $q^{\mathrm{R}}_{\mathrm{CA}} = a_{\mathrm{RC}} b_{\mathrm{RA}}$, where $a$ and $b$ are the Markov matrices on the branches indicated in the figure. Let $s = (s_{\mathrm{AAAA}}, s_{\mathrm{AAAG}}, \ldots, s_{\mathrm{UUUU}})$ be the expected spectrum of the four-species tree. Then the probability of the pattern shown is

$$P\{\mathrm{GUCAG}\} = \sum_{\mathrm{R}} s_{\mathrm{GURG}} q^{\mathrm{R}}_{\mathrm{CA}} = \sum_{\mathrm{R}} s_{\mathrm{GURG}} a_{\mathrm{RC}} b_{\mathrm{RA}} \,. \tag{6}$$

There are 1,023 more formulas like formula (6), as there are 255 more like formula (5). With the whole set of 1,024, you generate a five-species expected spectrum. You could generate all possible five-species spectra if you knew all possible $s$, $a$, and $b$. The derivation of the algorithm continues from this point just as before. Knowing 202 basis vectors for $\mathscr{H}$ is as good as knowing all possible $s$, and knowing seven basis matrices for $\mathcal{L}$ is as good as knowing all possible $a$ or all possible $b$ in $\mathcal{L}$. The triangularizer will be called $202 \cdot 7^2$, or 9,898 times, not 23,059,204.

## Beyond Balanced Transversions

Let $\alpha$ and $\beta$ be positive numbers. The matrix

$$\begin{pmatrix} e & f & g & \alpha g \\ h & i & j & \alpha j \\ k & \beta k & l & m \\ n & \beta n & p & q \end{pmatrix} \tag{7}$$

$$\begin{aligned} V_1 = {} & CAAA - CAAG + CAGA - CAGG + 2CACA - 2CACG \\ & + CGAA - CGAG + CGGA - CGGG + 2CGCA - 2CGCG \\ & + 2CCAA - 2CCAG + 2CCGA - 2CCGG + 4CCCA - 4CCCG \\ & - UAAA + UAAG - UAGA + UAGG - 2UACA + 2UACG \\ & - UGAA + UGAG - UGGA + UGGG - 2UGCA + 2UGCG \\ & - 2UUAA + 2UUAG - 2UUGA + 2UUGG - 4UUUA + 4UUUG \end{aligned}$$

$$\begin{aligned} V_2 = {} & CAAA - CAAG + CAGA - CAGG + 2CACA - 2CACG \\ & + CGAA - CGAG + CGGA - CGGG + 2CGCA - 2CGCG \\ & + 2CCAA - 2CCAG + 2CCGA - 2CCGG + 4CCCA - 4CCCG \\ & - UAAA + UAAG - UAGA + UAGG - 2UACA + 2UACG \\ & - UGAA + UGAG - UGGA + UGGG - 2UGCA + 2UGCG \\ & - 2UUAA + 2UUAG - 2UUGA + 2UUGG - 4UUCA + 4UUCG \end{aligned}$$

$$\begin{aligned} V_3 = {} & - CAAA - CAAG - 2CAAC + CAGA + CAGG + 2CAGC \\ & - CGAA - CGAG - 2CGAC + CGGA + CGGG + 2CGGC \\ & - 2CCAA - 2CCAG - 4CCAC + 2CCGA + 2CCGG + 4CCGC \\ & + UAAA + UAAG + 2UAAC - UAGA - UAGG - 2UAGC \\ & + UGAA + UGAG + 2UGAC - UGGA - UGGG - 2UGGC \\ & + 2UUAA + 2UUAG + 4UUAU - 2UUGA - 2UUGG - 4UUGU \end{aligned}$$

$$\begin{aligned} V_4 = {} & - CAAA - CAAG - 2CAAC + CAGA + CAGG + 2CAGC \\ & - CGAA - CGAG - 2CGAC + CGGA + CGGG + 2CGGC \\ & - 2CCAA - 2CCAG - 4CCAC + 2CCGA + 2CCGG + 4CCGC \\ & + UAAA + UAAG + 2UAAC - UAGA - UAGG - 2UAGC \\ & + UGAA + UGAG + 2UGAC - UGGA - UGGG - 2UGGC \\ & + 2UUAA + 2UUAG + 4UUAC - 2UUGA - 2UUGG - 4UUGC \end{aligned}$$

$$\begin{aligned} V_5 = {} & - CAAA + CAAG - CAGA + CAGG - 2CACA + 2CACG \\ & - CGAA + CGAG - CGGA + CGGG - 2CGCA + 2CGCG \\ & - 2CCAA + 2CCAG - 2CCGA + 2CCGG - 4CCCA + 4CCCG \\ & + UAAA - UAAG + UAGA - UAGG + 2UACA - 2UACG \\ & + UGAA - UGAG + UGGA - UGGG + 2UGCA - 2UGCG \\ & + 2UCAA - 2UCAG + 2UCGA - 2UCGG + 4UCUA - 4UCUG \end{aligned}$$

$$\begin{aligned} V_6 = {} & - CAAA + CAAG - CAGA + CAGG - 2CACA + 2CACG \\ & - CGAA + CGAG - CGGA + CGGG - 2CGCA + 2CGCG \\ & - 2CCAA + 2CCAG - 2CCGA + 2CCGG - 4CCCA + 4CCCG \\ & + UAAA - UAAG + UAGA - UAGG + 2UACA - 2UACG \\ & + UGAA - UGAG + UGGA - UGGG + 2UGCA - 2UGCG \\ & + 2UCAA - 2UCAG + 2UCGA - 2UCGG + 4UCCA - 4UCCG \end{aligned}$$

$$\begin{aligned} V_7 = {} & - CAAA - CAAG - 2CAAC + CAGA + CAGG + 2CAGC \\ & - CGAA - CGAG - 2CGAC + CGGA + CGGG + 2CGGC \\ & - 2CCAA - 2CCAG - 4CCAC + 2CCGA + 2CCGG + 4CCGC \\ & + UAAA + UAAG + 2UAAC - UAGA - UAGG - 2UAGC \\ & + UGAA + UGAG + 2UGAC - UGGA - UGGG - 2UGGC \\ & + 2UCAA + 2UCAG + 4UCAU - 2UCGA - 2UCGG - 4UCGU \end{aligned}$$

$$\begin{aligned} V_8 = {} & - CAAA - CAAG - 2CAAC + CAGA + CAGG + 2CAGC \\ & - CGAA - CGAG - 2CGAC + CGGA + CGGG + 2CGGC \\ & - 2CCAA - 2CCAG - 4CCAC + 2CCGA + 2CCGG + 4CCGC \\ & + UAAA + UAAG + 2UAAC - UAGA - UAGG - 2UAGC \\ & + UGAA + UGAG + 2UGAC - UGGA - UGGG - 2UGGC \\ & + 2UCAA + 2UCAG + 4UCAC - 2UCGA - 2UCGG - 4UCGC \end{aligned}$$

$$\begin{aligned} V_9 = {} & - UAAC + UAAU - UACA - UACC + UAUA + UAUU \\ & + UGAC - UGAU + UGCA + UGCC - UGUA - UGUU \end{aligned}$$

$$\begin{aligned} V_{10} = {} & - CAAA - CAAG - 2CAAU - CAGA - CAGG - 2CAGC \\ & - 2CACA - 2CACG - 2CACC - 2CACU + CGAA + CGAG \\ & + 2CGAU + CGGA + CGGG + 2CGGC + 2CGCA + 2CGCG \\ & + 2CGCC + 2CGCU + UAAA + UAAG + 2UAAC + UAGA \\ & + UAGG + 2UAGC + 2UACG + 2UACC + 2UAUA + 2UAUC \\ & - UGAA - UGAG - 2UGAC - UGGA - UGGG - 2UGGC \\ & - 2UGCG - 2UGCC - 2UGUA - 2UGUC \end{aligned}$$

$$V_{11} = - UGCA + UGCG + UGUA - UGUG$$

$$\begin{aligned} V_{12} = {} & CAAA + CAAG + 2CAAU + CAGA + CAGG + 2CAGC \\ & + 2CACA + 2CACG + 2CACC + 2CACU - CGAA - CGAG \\ & - 2CGAU - CGGA - CGGG - 2CGGC - 2CGCA - 2CGCG \\ & - 2CGCC - 2CGCU - UAAA - UAAG - 2UAAU - UAGA \\ & - UAGG - 2UAGC - 2UACA - 2UACG - 2UACC - 2UACU \\ & + UGAA + UGAG + 2UGAU + UGGA + UGGG + 2UGGC \\ & + 2UGCA + 2UGCG + 2UGCC + 2UGCU \end{aligned}$$

$$V_{13} = UGAC - UGAU - UGGC + UGGU$$

$$V_{14} = - UACA + UACG + UAUA - UAUG$$

$$V_{15} = UAAC - UAAU - UAGC + UAGU$$

$$\begin{aligned} V_{16} = {} & CCAA - CCAG + CCGA - CCGG + 2CCCA - 2CCCG \\ & - CUAA + CUAG - CUGA + CUGG - 2CUUA + 2CUUG \end{aligned}$$

$$\begin{aligned} V_{17} = {} & CCAA - CCAG + CCGA - CCGG + 2CCCA - 2CCCG \\ & - CUAA + CUAG - CUGA + CUGG - 2CUCA + 2CUCG \end{aligned}$$

$$\begin{aligned} V_{18} = {} & - CCAA - CCAG - 2CCAC + CCGA + CCGG + 2CCGC \\ & + CUAA + CUAG + 2CUAU - CUGA - CUGG - 2CUGU \end{aligned}$$

$$\begin{aligned} V_{19} = {} & - CCAA - CCAG - 2CCAC + CCGA + CCGG + 2CCGC \\ & + CUAA + CUAG + 2CUAC - CUGA - CUGG - 2CUGC \end{aligned}$$

$$V_{20} = - CCCA + CCCG + CCUA - CCUG$$

$$V_{21} = CCAC - CCAU - CCGC + CCGU$$

$$\begin{aligned} V_{22} = {} & CAAC - CAAU + CACA + CACC - CAUA - CAUU \\ & - CGAC + CGAU - CGCA - CGCC + CGUA + CGUU \end{aligned}$$

$$\begin{aligned} V_{23} = {} & - CAAC + CAAU + CACA + CACU - CAUA - CAUC \\ & + CGAC - CGAU - CGCA - CGCU + CGUA + CGUC \end{aligned}$$

$$V_{24} = CGCA - CGCG - CGUA + CGUG$$

$$V_{25} = CGAC - CGAU - CGGC + CGGU$$

$$V_{26} = CACA - CACG - CAUA + CAUG$$

$$V_{27} = CAAC - CAAU - CAGC + CAGU$$

$$\begin{aligned} V_{28} = {} & 2AAAC - 2AAAU + 2AACA + 2AACC - 2AAUA - 2AAUU \\ & + ACAC - ACAU + ACCA + ACCC - ACUA - ACUU \\ & + AUAC - AUAU + AUCA + AUCC - AUUA - AUUU \\ & - 2GAAC + 2GAAU - 2GACA - 2GACC + 2GAUA + 2GAUU \\ & - GCAC + GCAU - GCCA - GCCC + GCUA + GCUU \\ & - GUAC + GUAU - GUCA - GUCC + GUUA + GUUU \end{aligned}$$

$$\begin{aligned} V_{29} = {} & - 4AAAC + 4AAAU - 2AACC + 2AACU - 2AAUC + 2AAUU \\ & + 3ACAA + ACAG + 4ACAU + ACGA - ACGG + 2ACCA \\ & + 2ACCU + 2ACUA + 2ACUU - 3AUAA - AUAG - 4AUAC \\ & - AUGA + AUGG - 2AUCA - 2AUCC - 2AUUA - 2AUUC \\ & + 4GAAC - 4GAAU + 2GACC - 2GACU + 2GAUC - 2GAUU \\ & - 3GCAA - GCAG - 4GCAU - GCGA + GCGG - 2GCCA \\ & - 2GCCU - 2GCUA - 2GCUU + 3GUAA + GUAG + 4GUAC \\ & + GUGA - GUGG + 2GUCA + 2GUCC + 2GUUA + 2GUUC \end{aligned}$$

$$\begin{aligned} V_{30} = {} & GCAA - GCAG + GCGA - GCGG + 2GCCA - 2GCCG \\ & - GUAA + GUAG - GUGA + GUGG - 2GUUA + 2GUUG \end{aligned}$$

$$\begin{aligned} V_{31} = {} & - 4AACA - 2AACC - 2AACU + 4AAUA + 2AAUC + 2AAUU \\ & + 3ACAA + ACAG + 2ACAC + 2ACAU + ACGA - ACGG \\ & + 4ACUA + 2ACUC + 2ACUU - 3AUAA - AUAG - 2AUAC \\ & - 2AUAU - AUGA + AUGG - 4AUCA - 2AUCC - 2AUCU \\ & + 4GACA + 2GACC + 2GACU - 4GAUA - 2GAUC - 2GAUU \\ & - 3GCAA - GCAG - 2GCAC - 2GCAU - GCGA + GCGG \\ & - 4GCUA - 2GCUC - 2GCUU + 3GUAA + GUAG + 2GUAC \\ & + 2GUAU + GUGA - GUGG + 4GUCA + 2GUCC + 2GUCU \end{aligned}$$

$$\begin{aligned} V_{32} = {} & GCAA - GCAG + GCGA - GCGG + 2GCCA - 2GCCG \\ & - GUAA + GUAG - GUGA + GUGG - 2GUCA + 2GUCG \end{aligned}$$

$$\begin{aligned} V_{33} = {} & - GCAA - GCAG - 2GCAC + GCGA + GCGG + 2GCGC \\ & + GUAA + GUAG + 2GUAU - GUGA - GUGG - 2GUGU \end{aligned}$$

$$\begin{aligned} V_{34} = {} & - GCAA - GCAG - 2GCAC + GCGA + GCGG + 2GCGC \\ & + GUAA + GUAG + 2GUAC - GUGA - GUGG - 2GUGC \end{aligned}$$

$$V_{35} = GCCA - GCCG - GCUA + GCUG$$

$$V_{36} = GCAC - GCAU - GCGC + GCGU$$

$$\begin{aligned} V_{37} = {} & GAAC - GAAU + GACA + GACC - GAUA - GAUU \\ & - GGAC + GGAU - GGCA - GGCC + GGUA + GGUU \end{aligned}$$

$$\begin{aligned} V_{38} = {} & - GAAC + GAAU + GACA + GACU - GAUA - GAUC \\ & + GGAC - GGAU - GGCA - GGCU + GGUA + GGUC \end{aligned}$$

$$V_{39} = GGCA - GGCG - GGUA + GGUG$$

$$V_{40} = - GGAC + GGAU + GGGC - GGGU$$

$$V_{41} = - GACA + GACG + GAUA - GAUG$$

$$V_{42} = - GAAC + GAAU + GAGC - GAGU$$

$$\begin{aligned} V_{43} = {} & ACAA - ACAG + ACGA - ACGG + 2ACCA - 2ACCG \\ & - AUAA + AUAG - AUGA + AUGG - 2AUUA + 2AUUG \end{aligned}$$

$$\begin{aligned} V_{44} = {} & ACAA - ACAG + ACGA - ACGG + 2ACCA - 2ACCG \\ & - AUAA + AUAG - AUGA + AUGG - 2AUCA + 2AUCG \end{aligned}$$

$$\begin{aligned} V_{45} = {} & - ACAA - ACAG - 2ACAC + ACGA + ACGG + 2ACGC \\ & + AUAA + AUAG + 2AUAU - AUGA - AUGG - 2AUGU \end{aligned}$$

$$\begin{aligned} V_{46} = {} & - ACAA - ACAG - 2ACAC + ACGA + ACGG + 2ACGC \\ & + AUAA + AUAG + 2AUAC - AUGA - AUGG - 2AUGC \end{aligned}$$

$$V_{47} = ACCA - ACCG - ACUA + ACUG$$

$$V_{48} = - ACAC + ACAU + ACGC - ACGU$$

$$\begin{aligned} V_{49} = {} & AAAC - AAAU + AACA + AACC - AAUA - AAUU \\ & - AGAC + AGAU - AGCA - AGCC + AGUA + AGUU \end{aligned}$$

$$\begin{aligned} V_{50} = {} & - AAAC + AAAU + AACA + AACU - AAUA - AAUC \\ & + AGAC - AGAU - AGCA - AGCU + AGUA + AGUC \end{aligned}$$

$$V_{51} = AGCA - AGCG - AGUA + AGUG$$

$$V_{52} = AGAC - AGAU - AGGC + AGGU$$

$$V_{53} = - AACA + AACG + AAUA - AAUG$$

$$V_{54} = AAAC - AAAU - AAGC + AAGU$$

FIG. 3.—A basis for the subspace of all linear invariants under topology IV. This is probably neither the simplest nor most symmetric basis that could be found.

$$\begin{aligned} V_{1} = {} & ACAA-ACAG+ACGA-ACGG+2ACCA-2ACCG \\ & -AUAA+AUAG-AUGA+AUGG-2AUCA+2AUCG \\ & +CAAA-CAAG+CAGA-CAGG+2CACA-2CACG \\ & +CCAA-CCAG+CCGA-CCGG+2CCCA-2CCCG \\ & -UAAA+UAAG-UAGA+UAGG-2UACA+2UACG \\ & -UUAA+UUAG-UUGA+UUGG-2UUUA+2UUUG \end{aligned}$$

$$\begin{aligned} V_{2} = {} & ACAA-ACAG+ACGA-ACGG+2ACCA-2ACCG \\ & -AUAA+AUAG-AUGA+AUGG-2AUCA+2AUCG \\ & +CAAA-CAAG+CAGA-CAGG+2CACA-2CACG \\ & +CCAA-CCAG+CCGA-CCGG+2CCCA-2CCCG \\ & -UAAA+UAAG-UAGA+UAGG-2UACA+2UACG \\ & -UUAA+UUAG-UUGA+UUGG-2UUCA+2UUCG \end{aligned}$$

$$\begin{aligned} V_{3} = {} & ACAA+ACAG+2ACAC-ACGA-ACGG-2ACGC \\ & -AUAA-AUAG-2AUAC+AUGA+AUGG+2AUGC \\ & +CAAA+CAAG+2CAAC-CAGA-CAGG-2CAGC \\ & +CCAA+CCAG+2CCAC-CCGA-CCGG-2CCGC \\ & -UAAA-UAAG-2UAAC+UAGA+UAGG+2UAGC \\ & -UUAA-UUAG-2UUAU+UUGA+UUGG+2UUGU \end{aligned}$$

$$\begin{aligned} V_{4} = {} & ACAA+ACAG+2ACAC-ACGA-ACGG-2ACGC \\ & -AUAA-AUAG-2AUAC+AUGA+AUGG+2AUGC \\ & +CAAA+CAAG+2CAAC-CAGA-CAGG-2CAGC \\ & +CCAA+CCAG+2CCAC-CCGA-CCGG-2CCGC \\ & -UAAA-UAAG-2UAAC+UAGA+UAGG+2UAGC \\ & -UUAA-UUAG-2UUAC+UUGA+UUGG+2UUGC \end{aligned}$$

$$\begin{aligned} V_{5} = {} & ACAA-ACAG+ACGA-ACGG+2ACCA-2ACCG \\ & -AUAA+AUAG-AUGA+AUGG-2AUCA+2AUCG \\ & -CAAA+CAAG-CAGA+CAGG-2CACA+2CACG \\ & -CUAA+CUAG-CUGA+CUGG-2CUCA+2CUCG \\ & +UAAA-UAAC+UAGA-UAGC+2UACA-2UACG \\ & +UCAA-UCAG+UCGA-UCGG+2UCUA-2UCUG \end{aligned}$$

$$\begin{aligned} V_{6} = {} & ACAA-ACAG+ACGA-ACGG+2ACCA-2ACCG \\ & -AUAA+AUAG-AUGA+AUGG-2AUCA+2AUCG \\ & -CAAA+CAAG-CAGA+CAGG-2CACA+2CACG \\ & -CUAA+CUAG-CUGA+CUGG-2CUCA+2CUCG \\ & +UAAA-UAAG+UAGA-UAGG+2UACA-2UACG \\ & +UCAA-UCAG+UCGA-UCGG+2UCCA-2UCCG \end{aligned}$$

$$\begin{aligned} V_{7} = {} & ACAA+ACAG+2ACAC-ACGA-ACGG-2ACGC \\ & -AUAA-AUAG-2AUAC+AUGA+AUGG+2AUGC \\ & -CAAA-CAAG-2CAAC+CAGA+CAGG+2CAGC \\ & -CUAA-CUAG-2CUAC+CUGA+CUGG+2CUGC \\ & +UAAA+UAAG+2UAAC-UAGA-UAGG-2UAGC \\ & +UCAA+UCAG+2UCAU-UCGA-UCGG-2UCGU \end{aligned}$$

$$\begin{aligned} V_{8} = {} & ACAA+ACAG+2ACAC-ACGA-ACGG-2ACGC \\ & -AUAA-AUAG-2AUAC+AUGA+AUGG+2AUGC \\ & -CAAA-CAAG-2CAAC+CAGA+CAGG+2CAGC \\ & -CUAA-CUAG-2CUAC+CUGA+CUGG+2CUGC \\ & +UAAA+UAAG+2UAAC-UAGA-UAGG-2UAGC \\ & +UCAA+UCAG+2UCAC-UCGA-UCGG-2UCGC \end{aligned}$$

$$\begin{aligned} V_{9} = {} & AAAC-AAAU+AACA+AACC-AAUA-AAUU \\ & -AGAC+AGAU-AGCA-AGCC+AGUA+AGUU \\ & +GAAC-GAAU+GACA+GACC-GAUA-GAUU \\ & -GGAC+GGAU-GGCA-GGCC+GGUA+GGUU \\ & +2CAAC-2CAAU+2CACA+2CACC-2CAUA-2CGAC \\ & +2CGAU-2CGCA-2CGCC+2CGUA-2UAUU+2UGUU \end{aligned}$$

$$\begin{aligned} V_{10} = {} & -AAAC+AAAU+AACA+AACU-AAUA-AAUC \\ & +AGAC-AGAU-AGCA-AGCU+AGUA+AGUC \\ & -GAAC+GAAU+GACA+GACU-GAUA-GAUC \\ & +GGAC-GGAU-GGCA-GGCU+GGUA+GGUC \\ & -2CAAC+2CAAU+2CACA+2CACU-2CAUA+2CGAC \\ & -2CGAU-2CGCA-2CGCU+2CGUA-2UAUC+2UGUC \end{aligned}$$

$$\begin{aligned} V_{11} = {} & -CACA+CACG+CAUA+CGCA-CGCG-CGUA \\ & +UACA-UACG-UAUA+UGUG \end{aligned}$$

$$V_{12} = CAUA-CGUA-UAUA+UGUA$$

$$V_{13} = -CACU+CGCU+UACU-UGCU$$

$$V_{14} = CACC-CGCC-UACC+UGCC$$

$$V_{15} = CACG-CGCG-UACG+UGCG$$

$$V_{16} = CACA-CGCA-UACA+UGCA$$

$$\begin{aligned} V_{17} = {} & -CAAC+CAAU+CACC+CGAC-CGAU-CGCC \\ & +UAAC-UAAU-UACC+UGGU \end{aligned}$$

$$V_{18} = CAGC-CGGC-UAGC+UGGC$$

$$V_{19} = CAGG-CGGG-UAGG+UGGG$$

$$V_{20} = CAGA-CGGA-UAGA+UGGA$$

$$V_{21} = -CAAU+CGAU+UAAU-UGAU$$

$$V_{22} = CAAC-CGAC-UAAC+UGAC$$

$$V_{23} = CAAG-CGAG-UAAG+UGAG$$

$$V_{24} = -CAAA+CGAA+UAAA-UGAA$$

$$V_{25} = UACA-UACG-UAUA+UAUG$$

$$V_{26} = UAAC-UAAU-UAGC+UAGU$$

$$V_{27} = -CUCA+CUCG+CUUA-CUUG$$

$$V_{28} = CUAC-CUAU-CUGC+CUGU$$

$$V_{29} = -CCCA+CCCG+CCUA-CCUG$$

$$V_{30} = CCAC-CCAU-CCGC+CCGU$$

$$\begin{aligned} V_{31} = {} & AAAC-AAAU+AACA+AACC-AAUA-AAUU \\ & -AGAC+AGAU-AGCA-AGCC+AGUA+AGUU \\ & +GAAC-GAAU+GACA+GACC-GAUA-GAUU \\ & -GGAC+GGAU-GGCA-GGCC+GGUA+GGUU \\ & +2CAAC-2CAAU+2CACA+2CACC-2CAUA-2CAUU \\ & -2CGAC+2CGAU-2CGCA-2CGCC+2CGUA+2CGUU \end{aligned}$$

$$\begin{aligned} V_{32} = {} & -AAAC+AAAU+AACA+AACU-AAUA-AAUC \\ & +AGAC-AGAU-AGCA-AGCU+AGUA+AGUC \\ & -GAAC+GAAU+GACA+GACU-GAUA-GAUC \\ & +GGAC-GGAU-GGCA-GGCU+GGUA+GGUC \\ & -2CAAC+2CAAU+2CACA+2CACU-2CAUA-2CAUC \\ & +2CGAC-2CGAU-2CGCA-2CGCU+2CGUA+2CGUC \end{aligned}$$

$$V_{33} = CGCA-CGCG-CGUA+CGUG$$

$$V_{34} = -CGAC+CGAU+CGGC-CGGU$$

$$V_{35} = -CACA+CACG+CAUA-CAUG$$

$$V_{36} = CAAC-CAAU-CAGC+CAGU$$

$$\begin{aligned} V_{37} = {} & AAAC-AAAU+AACA+AACC-AAUA-AAUU \\ & +AGAC-AGAU+AGCA+AGCC-AGUA-AGUU \\ & +2ACAC-2ACAU+2ACCA+2ACCC-2ACUA-2AUUU \\ & -GAAC+GAAU-GACA-GACC+GAUA+GAUU \\ & -GGAC+GGAU-GGCA-GGCC+GGUA+GGUU \\ & -2GCAC+2GCAU-2GCCA-2GCCC+2GCUA+2GUUU \end{aligned}$$

$$\begin{aligned} V_{38} = {} & -AAAC+AAAU+AACA+AACU-AAUA-AAUC \\ & -AGAC+AGAU+AGCA+AGCU-AGUA-AGUC \\ & -2ACAC+2ACAU+2ACCA+2ACCU-2ACUA-2AUUC \\ & +GAAC-GAAU-GACA-GACU+GAUA+GAUC \\ & +GGAC-GGAU-GGCA-GGCU+GGUA+GGUC \\ & +2GCAC-2GCAU-2GCCA-2GCCU+2GCUA+2GUUC \end{aligned}$$

$$\begin{aligned} V_{39} = {} & -ACCA+ACCG+ACUA+AUCA-AUCG-AUUA \\ & +GCCA-GCCG-GCUA+GUUG \end{aligned}$$

$$V_{40} = ACUA-AUUA-GCUA+GUUA$$

$$V_{41} = -ACCU+AUCU+GCCU-GUCU$$

$$V_{42} = ACCC-AUCC-GCCC+GUCC$$

$$V_{43} = -ACCG+AUCG+GCCG-GUCG$$

$$V_{44} = ACCA-AUCA-GCCA+GUCA$$

$$\begin{aligned} V_{45} = {} & ACAC-ACAU-ACGC-AUAC+AUAU+AUGC \\ & -GCAC+GCAU+GCGC-GUGU \end{aligned}$$

$$V_{46} = -ACGC+AUGC+GCGC-GUGC$$

$$V_{47} = ACGG-AUGG-GCGG+GUGG$$

$$V_{48} = ACGA-AUGA-GCGA+GUGA$$

$$V_{49} = -ACAU+AUAU+GCAU-GUAU$$

$$V_{50} = ACAC-AUAC-GCAC+GUAC$$

$$V_{51} = -ACAG+AUAG+GCAG-GUAG$$

$$V_{52} = -ACAA+AUAA+GCAA-GUAA$$

$$\begin{aligned} V_{53} = {} & AAAC-AAAU+AACA+AACC-AAUA-AAUU \\ & +AGAC-AGAU+AGCA+AGCC-AGUA-AGUU \\ & +2ACAC-2ACAU+2ACCA+2ACCC-2ACUA-2ACUU \\ & -GAAC+GAAU-GACA-GACC+GAUA+GAUU \\ & -GGAC+GGAU-GGCA-GGCC+GGUA+GGUU \\ & -2GCAC+2GCAU-2GCCA-2GCCC+2GCUA+2GCUU \end{aligned}$$

$$\begin{aligned} V_{54} = {} & -AAAC+AAAU+AACA+AACU-AAUA-AAUC \\ & -AGAC+AGAU+AGCA+AGCU-AGUA-AGUC \\ & -2ACAC+2ACAU+2ACCA+2ACCU-2ACUA-2ACUC \\ & +GAAC-GAAU-GACA-GACU+GAUA+GAUC \\ & +GGAC-GGAU-GGCA-GGCU+GGUA+GGUC \\ & +2GCAC-2GCAU-2GCCA-2GCCU+2GCUA+2GCUC \end{aligned}$$

$$V_{55} = GCCA-GCCG-GCUA+GCUG$$

$$V_{56} = -GCAC+GCAU+GCGC-GCGU$$

$$V_{57} = -GGCA+GGCG+GGUA-GGUG$$

$$V_{58} = GGAC-GGAU-GGGC+GGGU$$

$$V_{59} = GACA-GACG-GAUA+GAUG$$

$$V_{60} = -GAAC+GAAU+GAGC-GAGU$$

$$V_{61} = -AUCA+AUCG+AUUA-AUUG$$

$$V_{62} = AUAC-AUAU-AUGC+AUGU$$

$$V_{63} = -ACCA+ACCG+ACUA-ACUG$$

$$V_{64} = -ACAC+ACAU+ACGC-ACGU$$

$$V_{65} = AGCA-AGCG-AGUA+AGUG$$

$$V_{66} = -AGAC+AGAU+AGGC-AGGU$$

$$V_{67} = -AACA+AACG+AAUA-AAUG$$

$$V_{68} = -AAAC+AAAU+AAGC-AAGU$$

FIG. 4.—A basis for the subspace of all linear invariants of topology I

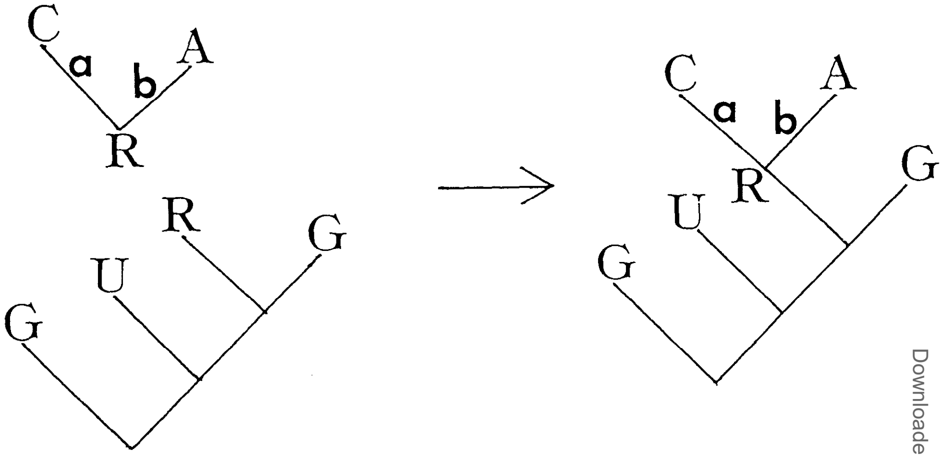

FIG. 5.—A five-species tree built by joining a two-species tree to a four-species tree

generalizes matrix (2). If the additional constraints

$$\alpha l - m = \frac{q}{\alpha} - p \tag{8}$$

and

$$\beta e - f = \frac{i}{\beta} - h \tag{9}$$

are imposed, these matrices form a multiplicatively closed set. That is, for each $\alpha$ and $\beta$ there is an alternative semigroup. I should emphasize that equations (8) and (9) do not follow from matrix (7). They are required to keep matrix (7) from breaking down over the millions of years. Each of these semigroups leads to invariants, as I shall show next.

## An Assay for Semigroups

The following theorem provides a quick way of screening semigroups to see whether they give invariants.

Theorem: *If the two-species tree has no invariants under a given semigroup, then larger trees also have no invariants.*

I shall sketch the proof for three-species trees. Let $\mathscr{H}_2$ and $\mathscr{H}_3$ be the two-species and three-species versions, respectively, of $\mathscr{H}$. Say $\mathscr{H}_2$ contains every pattern AA, AG, . . . , UU. I show that $\mathscr{H}_3$ contains every pattern AAA, AAG, . . . , UUU.

For example, here is how the pattern AGU arises. Let $\sigma(R, b_l, c_m)$ be the two-species expected spectrum from initial state $R$ and Markov matrices $b_l$ and $c_m$ on the branches indicated in figure 6. Since $\mathscr{H}_2$ contains every pattern, lambdas exist for which

$$\mathrm{GU} = \sum_l \sum_m \lambda^{\mathrm{A}lm}\sigma(\mathrm{A}, b_l, c_m) + \sum_l \sum_m \lambda^{\mathrm{G}lm}\sigma(\mathrm{G}, b_l, c_m)$$
$$+ \sum_l \sum_m \lambda^{\mathrm{C}lm}\sigma(\mathrm{C}, b_l, c_m) + \sum_l \sum_m \lambda^{\mathrm{U}lm}\sigma(\mathrm{U}, b_l, c_m) .$$

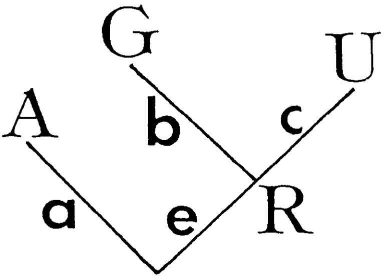

FIG. 6.—Pattern AGU of a linear combination of possible spectra

Similarly, for each R ∈ {A, G, C, U}, thetas and rhos exist, with ρ ∈ {A, G, C, U}, for which

$$\mathrm{AR} = \sum_k \sum_i \sum_j \theta_{\mathrm{R}}^{kij} \sigma(\rho_k, a_i, e_j) .$$

Then

$$\mathrm{AGU} = \sum_k \sum_i \sum_j \sum_l \sum_m \Big(\sum_{\mathrm{R}} \theta_{\mathrm{R}}^{kij} \lambda^{\mathrm{R}lm}\Big) \sigma(\rho_k, a_i, b_l, c_m, e_j) .$$

Notice that I am taking the point of view here that the probability mass function of the root is always unconstrained so that invariants do not depend on it, only on the semigroup.

For an example, assume that $b$ and $c$ are Markov matrices of the form of matrix (7) so that $\alpha b_{\mathrm{AC}} - b_{\mathrm{AU}} = 0$, etc. Let $V = \alpha\beta\mathrm{AC} - \beta\mathrm{AU} - \alpha\mathrm{GC} + \mathrm{GU}$. Let $s = (s_{\mathrm{AA}}, s_{\mathrm{AG}}, \ldots, s_{\mathrm{UU}})$ be any expected spectrum. Then $V$ is an invariant, since

$$\begin{aligned} \mathrm{V} \cdot s &= \alpha\beta s_{\mathrm{AC}} - \beta s_{\mathrm{AU}} - \alpha s_{\mathrm{GC}} + s_{\mathrm{GU}} \\ &= \sum_{\mathrm{R}} r_{\mathrm{R}}(\alpha\beta b_{\mathrm{RA}} c_{\mathrm{RC}} - \beta b_{\mathrm{RA}} c_{\mathrm{RU}} - \alpha b_{\mathrm{RG}} c_{\mathrm{RC}} + b_{\mathrm{RG}} c_{\mathrm{RU}}) \\ &= \sum_{\mathrm{R}} r_{\mathrm{R}}(\beta b_{\mathrm{RA}} - b_{\mathrm{RG}})(\alpha c_{\mathrm{RC}} - c_{\mathrm{RU}}) \\ &= 0 . \end{aligned}$$

This is because the binomial $(\beta b_{\mathrm{RA}} - b_{\mathrm{RG}})$ is zero when R ∈ {C, U} and the other binomial is zero when R ∈ {A, G}.

One can see trivially that the converse of the theorem is true by considering the four-species invariant for topology IV:

$$W = \sum_{\mathrm{X,Y} \in \{\mathrm{A,G,C,U}\}} \sum (\alpha\beta\mathrm{XYAC} - \beta\mathrm{XYAU} - \alpha\mathrm{XYGC} + \mathrm{XYGU}) ,$$

which, since it can be verified or rejected without even looking at species A and B, is merely a restatement of the invariant $V$ just proved.

## General Semigroups

Let $\mathcal{S}$ be any semigroup of Markov matrices. Let $\mathcal{T}$ be the algebra spanned by $\mathcal{S}$, i.e., the set of all matrices $t$ of the form

$$t = \sum_{k=1}^{n} \lambda^k s_k \, ,$$

where each $s_k$ is in $\mathcal{S}$. Since $\mathcal{T}$ has at most 16 dimensions, there is a set of at most 16 matrices $t_k$ that span $\mathcal{T}$. Each of these is a linear combination of finitely many elements from $\mathcal{S}$, so there is a finite subset of $\mathcal{S}$ that can serve in place of $\{p_1, p_2, \ldots, p_7\}$ in the algorithm to generate all linear invariants resulting from $\mathcal{T}$ (or, more exactly, from the set of all stochastic matrices in $\mathcal{T}$, which is a semigroup contained in $\mathcal{T}$ and containing $\mathcal{S}$). This shows that, despite appearances, the algorithm does not depend on having a semigroup that is defined by linear constraints.

In practice one would ordinarily use the larger semigroup rather than $\mathcal{S}$, since one's assumptions are then weaker. There is no loss in doing this; the two semigroups have the same invariants, as one can easily prove using by-now familiar methods.

The most familiar semigroup of Markov matrices is the one-parameter semigroup, a set of matrices of the form $e^{Mt}$ with $M$ a fixed matrix and $t$ a varying, real, "time" or "evolutionary-distance" parameter. Such a semigroup will have a basis of at most four matrices $s_k$ (Brogan 1985, pp. 202–205, 208–209).

## An Application

The expected spectrum for a string of RNA is the sum of the expected spectra for many individual positions. Each summand is in $\mathcal{H}$, and therefore the sum is also. For a given tree and semigroup, an algorithm is at hand to compute $\mathcal{H}$. If the actual spectrum is significantly far from $\mathcal{H}$, either the tree or the semigroup must be rejected. The most fortunate user of the method would have some grounds for total faith in the semigroup and would find that a good statistical test rejects 14 of the 15 trees.

I have not located a test of the hypothesis that the parameters of a multinomial satisfy given linear equations. Here I treat the problem with a multivariate normal approximation. I hope someone can show me a better way.

Lake provided me with excellent data: aligned sequences of the 16S rRNA from the four species *Homo sapiens, Desulfurococcus mobilis, Halobacterium volcanii,* and *Escherichia coli.* I call these $\underline{\text{A}}$, $\underline{\text{B}}$, $\underline{\text{C}}$, and $\underline{\text{D}}$, respectively. The data are part of those treated by Lake (1988). There are 1,095 aligned patterns in the set. To create an observed spectrum $S$, I merely count the number of AAAA, AAAG, AAAC, . . . , UUUU. [They happen to be 121, 5, 4, . . . , 74. Thus $S = (121, 5, 4, \ldots, 74)$.]

I decompose $S$ into a sum of two vectors $S = h + x$, where $h$ satisfies the invariants (i.e., $h \in \mathcal{H}$) and where the "error" $x$ is orthogonal to $\mathcal{H}$ (which requires that $x$ be a linear combination of the invariants).

I assume that the distribution of $S$ is multinomial, with parameters given by the frequencies in $h$. The reader should be aware of three objections to this commonly accepted practice. First, $S$ is a mixture of multinomials, not a multinomial. The result is that my estimated variances will be too large. This does not invalidate the test, for the power rather than the size is what suffers damage. The nonlinear invariants—there are >140 of them—might be used here to improve the power. Second, $h$ is merely the element of $\mathcal{H}$ that lies closest to $S$ in *Euclidean distance.* It is a rough approximation for the element of $\mathcal{H}$ that is best able to survive the subsequent $\chi^2$ test. Third, I always got absurd negative components in $h$. I arbitrarily replaced these with zeros.

I next replace the multinomial with a multivariate normal having the same variances and covariances. This is common practice, but, despite the reassurances of authorities (van der Waerden 1969, pp. 226–228), one should wonder about the effects of small sample size.

The rest is routine. The normal distribution is projected (Graybill 1961, theorem 3.22, p. 68) into the subspace spanned by the invariants (because $S$ was projected into this subspace to create $x$); $x$ is recoordinatized in a frame where the normal is standard; and the length of $x$ is tested with a one-tailed $\chi^2$ with 54 or 68 degrees of freedom.

And what resulted? Almost nothing. I reject only trees III and IX at the 1% (or 5%) level. There are three possible explanations. First, the test I used, with its many approximations, could be at fault. Second, the semigroup $\mathcal{L}$ may be wrong. Third, maybe information about the ancestral connections of these four species is just not in their 16S ribosomal RNA.

Are the conclusions of Lake (1988) unjustified, then? I think not. One valid test can reject while another accepts, and it is a grave sin in statistics to try one test after another *on the same data,* looking for the conclusion you prefer. I do not wish to be guilty of this (although I really have no preference among the 15 trees for these four species), so I urge the reader to regard my reanalysis as merely a demonstration of method. Clearly, a better-founded statistical technique is needed—and fresh data.

## Conclusion: A Mathematician's Perspective

If people must infer phylogenies from nucleic acid sequences, the method of linear invariants is, in my opinion, the best available today. But before I place too much confidence in it, I would like to know whether transversions are truly balanced. This will require statistical studies (designed first, conducted afterward!) involving large samples—and not just globin genes and not just vertebrates. I would also like to know how badly the method is affected by small deviations from this key assumption. Finally, and what is probably most difficult, I would like to know whether the assumptions of statistical independence can be justified.

## Acknowledgments

I thank Glen E. Campbell, David Clair, Joseph Felsenstein, James A. Lake, Joe V. Petty, and John A. Williamson. Part of this study was funded by National Science Foundation grant BSR 88-05729.

## LITERATURE CITED

ALBERT, A. A. 1937. Modern higher algebra. University of Chicago Press, Chicago.

———. 1961. Structure of algebras. American Mathematical Society, Providence.

BROGAN, W. L. 1985. Modern control theory. Prentice-Hall, Englewood Cliffs, N.J.

CAVENDER, J. A. 1978. Taxonomy with confidence. Math. Biosci. **40**:271–281 [erratum, **44**: 308 (1979)].

CAVENDER, J. A., and J. FELSENSTEIN. 1987. Invariants of phylogenies in a simple case with discrete states. J. Classification **4**:51–71.

DEURING, M. 1968. Algebren. Springer, Berlin.

DIEUDONNÉ, J. A., and J. B. CARRELL. 1971. Invariant theory, old and new. Academic Press, New York.

GRAYBILL, F. A. 1961. An introduction to linear statistical models. Vol. **1**. McGraw-Hill, New York.

GREUB, W. H. 1967. Multilinear algebra. Springer, New York.

HOLMQUIST, R., M. M. MIYAMOTO, and M. GOODMAN. 1988. Analysis of higher-primate phylogeny from transversion differences in nuclear and mitochondrial DNA by Lake's methods of evolutionary parsimony and operator metrics. Mol. Biol. Evol. **5**:217–236.

KEMENY, J. G., J. L. SNELL, and G. L. THOMPSON. 1974. Introduction to finite mathematics. 3d ed. Prentice-Hall, Englewood Cliffs, N.J.

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

———. 1988. Origin of the eukaryotic nucleus as determined by rate-invariant analysis of rRNA sequences. Nature **331**:184–186.

SERRE, J.-P. 1977. Linear representations of finite groups. Springer, New York.

SMITH, K. T. 1983. Primer of modern analysis. Springer, New York.

TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Pp. 57–86 *in* R. M. MIURA, ed. Lectures on mathematics in the life sciences, vol. **17**: Some mathematical questions in biology: DNA sequence analysis. American Mathematical Society, Providence.

VAN DER WAERDEN, B. L. 1969. Mathematical statistics. Springer, New York.

WALTER M. FITCH, reviewing editor

Received March 23, 1988; revision received November 15, 1988