JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 1, Number 1, 1994
Mary Ann Liebert, Inc., Publishers
Pp. 77–92

Skewed Base Compositions, Asymmetric Transition Matrices, and Phylogenetic Invariants

V. FERRETTI, B.F. LANG, and D. SANKOFF

ABSTRACT

Evolutionary inference methods that assume equal DNA base compositions and symmetric nucleotide substitution matrices, where these assumptions do not hold, are likely to group species on the basis of similar base compositions rather than true phylogenetic relationships. We propose an invariants-based method for dealing with this problem. An invariant Q_T of a tree T under a k-state Markov model, where a generalized time parameter is identified with the E edges of T, allows us to recognize whether data on N observed species can be associated with the N terminal vertices of T in the sense of having been generated on T rather than on any other tree with N terminals. The form of the generalized time parameter is a positive determinant matrix in some semigroup S of stochastic matrices. The invariance is with respect to the choice of the set of E matrices in S, one associated with each of the E edges of T. We apply a general "empirical" method of finding invariants of a parametrized functional form. It involves calculating the probability f of all k^N data possibilities for each of m sets of E matrices in S to associate with the edges of T, then solving for the parameters using the m equations of form Q(f) = 0. We discuss the problems of finding asymmetric models satisfying the property of semigroup closure, of finding asymmetric models that admit invariants at all, and of the computational complexity of the method. We propose a class of semigroups S_c containing matrices of form $\begin{vmatrix} 1-a & a \\ ca & 1-ca \end{vmatrix}$ to account for A+T versus G+C asymmetries in DNA base composition. Quadratic invariants are obtained for rooted trees with three and with four terminals. In the latter case the smallest set of algebraically independent invariants is sought. These invariants are applied to data pertaining the fungal evolution and to the origin of mitochondria as bacterial endosymbionts.

Key words: phylogeny, AT-richness, semigroups, symbolic computing

INTRODUCTION

THE EVOLUTION OF NUCLEOTIDE SEQUENCES is most frequently modeled as a Markov process on the set of four bases {A,G,C,T}, with uniform initial distribution and symmetric substitution[1] matrices. Such models are inappropriate when observed base compositions deviate strongly from uniform, because evolutionary inference methods assuming uniformity and symmetry are likely to group species on the basis of similiar base compositions rather than true homology. In this paper, we investigate

Université de Montréal, CRM, Montreal, Quebec, H3T1TZ, Canada

[1] We use the term "substitution matrix" instead of "transition probability matrix" to avoid confusion with the biological term "transition," which is reserved for purine–purine (A ↔ G) or pyrimidine–pyrimidine (C ↔ T) mutations.

asymmetric substitution matrices and arbitrary initial distributions as models for evolution where the phylogenetic inference problem involves species with skewed (AT-rich or AT-poor) base compositions.

The approach we adopt is that of phylogenetic invariants and the specific methodology is that of "empirical invariants" (Ferretti and Sankoff, 1993). Our goal is to derive invariants for some meaningful asymmetric model. We also explore how these invariants may be applied to construct phylogenies based on real data. A different approach to correcting for skewed base composition, using simulation, has recently been suggested by Steel *et al.* (1993a).

THE MODEL AND THE INFERENCE PROBLEM

We denote by **T** an evolutionary tree (a rooted tree with positive lengths associated with the edges) whose branching structure is to be found. We know only that there must be N terminal vertices, each associated with an observed nucleotide sequence from one species. The N sequences are aligned and are all of length n. It is postulated that **T** contains at least one nonterminal vertex, its root, denoted ρ, such that the flow of time is directed away from ρ on all edges on the paths joining ρ to the terminal vertices. Each of the nonterminal vertices represents an idealized speciation event, and the edge-length $|XY|$ corresponds to the time elapsed between the speciation (nonterminal) or observation (terminal) events represented by vertices X and Y. We stress that the details of **T**, including the branching structure connecting its vertices as well as the edge-lengths, are unknown. What we do know, or assume, is an evolutionary model, namely a semigroup **S** of substitution matrices, each with positive determinant, on the state space $\{1, \ldots, k\}$, where some $M_{XY} \in \mathbf{S}$ is to be associated with each edge XY of **T**. (In addition, in some inference problems we assume that we know an initial distribution (generally uniform) π on the state space $\{1, \ldots, k\}$, associated with the root ρ, while in the general problem π remains unknown.)

The easiest case to investigate is $k = 2$, while $k = 4$ is necessary to model evolution at the level of nucleotide sequences, and $k = 20$ for proteins.

It will be seen in the ensuing presentation that dealing with the matrices M_{XY} obviates any reference to the edge-lengths $|XY|$, and constitutes a somewhat more general approach. In all these models, one can identify the edge-length with $-\log \det M$.

At each of the n aligned sequence positions independently, we assume that the observed state at a terminal vertex Y is drawn from $\{1, \ldots, k\}$ according to the distribution $\pi M_{v_0v_1} M_{v_1v_2} \ldots M_{v_{r-1}v_r}$, where $\rho = v_0, v_1, \ldots, v_r = Y$ is the sequence of vertices on the path between ρ and Y. Note that $M_{v_0v_1} M_{v_1v_2} \ldots M_{v_{r-1}v_r} \in \mathbf{S}$. The paths from ρ to two different terminal vertices Y_1 and T_2 necessarily contain some of the same nonterminal vertices $\rho = v_0, v_1, \ldots, v_q = X$ (possibly with $q = 0$). Then the structure of **T** is incorporated into the model by assuming that the trajectories between ρ and Y_1, and between ρ and Y_2, are identical between ρ and X. Indeed, the sample paths of the process can be constructed by selecting a state i_0 at ρ from $\{1, \ldots, k\}$ according to π, calculating $\pi_1 = e_{i_0} M_{v_0v_1}$ for each vertex v_1 adjacent to ρ, selecting a state at each such v_1 according to the probability distribution π_1, and if v_1 is a nonterminal vertex, calculating π_2 for each v_2 adjacent to v_1 (except v_0), and so on.

The n sequence positions are assumed for present purposes to represent n independent samples of the same process.[2] For each position, the only part of the sample path we can observe is the N-tuple representing its states at the N terminal vertices of **T**. The observed frequencies of all possible N-tuples—the observed spectrum of the process—become the basic data for phylogenetic inference.

The invariants approach, introduced by Cavender and Felsenstein (1987) and Lake (1987, 1988), focuses on estimating the branching structure of **T** and not the associated edge-lengths. More precisely, it does not try to reconstruct the details of the matrices associated with each edge. This limited goal is motivated largely by the interest of the biologist primarily in the branching order of the phylogenetic tree, *e.g.*, whether X and Y are more closely related to each other than either is to Z, and only secondarily in the details of how much time has elapsed between the divergence of Z and the split of X from Y. Another major motivation of this approach is the prohibitive computational expense of a full maximum likelihood estimation of **T** and its edge-lengths (Felsenstein, 1981).

[2] Although there has been recognition that different positions evolve at different rates (*e.g.*, every third position in protein coding sequences tends to evolve more rapidly than the other two) and that positions often do not evolve independently (*e.g.*, positions that are close together in either primary or secondary structure may co-evolve).

The idea is to find a function Q_T of the data (the spectrum) and of tree topology **T** that is predicted—in terms of the process hypothesized to have generated the data—to be invariant (*e.g.*, identically equal to zero) with respect to the choice of M∈S (generalized length) associated with each edge in the correct tree, but to be sensitive to this choice (and generally to be remote from the invariant value) for all other trees **U, V**, ... Then be evaluating the functions Q_T, Q_U, Q_V, . . . on an observed spectrum, only one should take on (or, for finite *n*, be close to) the predicted invariant value, namely the function associated with the tree that generated the spectrum, so that this tree can thus be identified, and the phylogeny correctly inferred.

THE SEMIGROUP S

In the simplest semigroup $\mathbf{S_{JC}}$ for modeling biological evolution, $M_{XY} = s\mathbf{J} + (1 - ks)\mathbf{I}$, where **I** is the k×k identity matrix, **J** is the k×k matrix of 1's, and $0 < s < 1 - (k - 1)s$, so that

$$M_{XY} = \begin{bmatrix} 1-(k-1)s & s & \cdots & s \\ s & 1-(k-1)s & \cdots & s \\ \cdots & \cdots & \cdots & \cdots \\ s & s & \cdots & 1-(k-1)s \end{bmatrix} \tag{1}$$

Note that the diagonal elements are larger than the off-diagonal elements. These are essentially the "equicorrelation matrices" of multivariate statistics (*cf.* Mardia *et al.*, 1979). In biology, for the case k = 4, $\mathbf{S_{JC}}$ is known as the Jukes–Cantor (1969) model. For a fixed k, the parameter s completely determines the matrix. Setting $t = -\log(1 - ks)$, the parameter t may be identified with edge-length in the sense that Σt over any path $v_0, v_1, \ldots, v_r$ in the tree is equal to the parameter t derived from the matrix $M_{v_0v_1}M_{v_1v_2} \ldots M_{v_{r-1}v_r}$. Also $-\log \det M = t(k - 1)$.

At the other extreme, we have the semigroup $\mathbf{S_{PD}}$ of all stochastic k×k matrices with positive determinant. Here the k^2 matrix entries take on values in a k(k − 1)-dimensional subspace. No proper subset of the parameters by itself can be identified with edge-length.

In between $\mathbf{S_{JC}}$ and $\mathbf{S_{PD}}$ are a variety of strong, idealized, and weak, realistic models. One of the most frequently used is the Kimura (1980) 2-parameter model $\mathbf{S_{2K}}$ for k = 4. This has the form

$$\begin{bmatrix} 1-a-2b & a & b & b \\ a & 1-a-2b & b & b \\ b & b & 1-a-2b & a \\ b & b & a & 1-a-2b \end{bmatrix} \tag{2}$$

where $1 - a - 2b > a > b$. This distinguishes two types of state change, transitions (probability a) and transversions (each of probability b).

As an evolutionary model, the fewer the parameters necessary to specify an element in **S**, the stronger, and less realistic, the claim it implicitly makes about the mechanism of sequence evolution. Thus the one-parameter Jukes–Cantor model $\mathbf{S_{JC}}$ requires that on a given edge, the substitution probabilities between all pairs of states are the same. The weakest model, $\mathbf{S_{PD}}$, with k(k − 1) parameters, makes no constraint on the transition probabilities other than the boundaries imposed on the k^2-dimensional space by the stochasticity of the matrix and the positivity of its deteminant.

Let **f** be the probability distribution on the k^N N-tuples as determined by **T** and the matrices M associated with its edges. Suppose the matrices in **S** are determined by h parameters, where $1 \leq h \leq k(k - 1)$. Suppose further that the **T** is binary, that is ρ and all other nonterminal vertices each have two outgoing edges, and all vertices except ρ have one incoming edge, as is most often justified in these problems. Then there are 2(N − 1) edges in the tree and 2h(N − 1) + k − 1 parameters for the problem, if π is unknown. Since the k^N values of **f** are determined by only 2h(N − 1) + k − 1 parameters, there must be at least $k^N - 2h(N - 1) - k + 1$ algebraically independent relations[3] among

[3] This can be shown using the inverse function theorem. A similar calculation can be carried out for the cases where the tree has nonbinary vertices or no root, with fewer than 2(N − 1) edges, or where the distribution π is known, so that there are k − 1 fewer parameters.

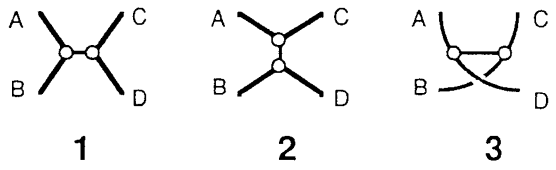

FIG. 1. The three unrooted binary trees on four terminal vertices.

the values of **f**, which hold independent of which M in **S** are associated with the edges of **T**. These are all invariants of the spectrum for a given **S** and **T**. One such relation is $\Sigma f = 1$. This relation is of course true of all **T**, as may be others. If, however, there is some relationship $Q(\mathbf{f}) = 0$, which is identically true (over all choices of M in **S**) for **T** but not for some other tree **U** on N terminal vertices, then we say that $Q_{\mathbf{T}}(\mathbf{f})$ is a phylogenetic invariant.

A SKETCH OF PREVIOUS WORK

Lake (1987) found invariant linear combinations of the 256 frequencies **f** for $\mathbf{S}_{2\mathbf{K}}$ and $\pi = (1/4,1/4,1/4,1/4)$, in the case $N = 4$ (*i.e.*, for the three unrooted trees in Fig. 1; with $\mathbf{S}_{2\mathbf{K}}$ the root may be arbitrarily placed on any edge or nonterminal vertex without changing the **f**). Similarly, Cavender and Felsenstein (1987) found two invariant quadratic combinations of the 16 frequencies **f** for $N = 4$ (the same three unrooted trees in Fig. 1), for $\mathbf{S}_{\mathbf{JC}}$ where $k = 2$ and $\pi = (1/2,1/2)$.

Cavender (1989) subsumed Lake's results for $k = 4$ and $N = 4$ in a construction of all possible linear invariants for a semigroup much larger than $\mathbf{S}_{2\mathbf{K}}$, which we denote by $\mathbf{S}_{\mathbf{CAV}}$, where the matrices are based on six independent parameters and may be asymmetric:

$$\begin{bmatrix} 1-a-2b & a & b & b \\ c & 1-c-2d & d & d \\ p & p & 1-q-2p & q \\ r & r & s & 1-s-2r \end{bmatrix} \tag{3}$$

where $a + b = c + d$ and $s + r = p + q$. He also described a construction for higher N. These linear invariants are the only ones which have been found to date for an asymmetric model. Fu and Li (1992b) and Nguyen and Speed (1992) have also studied linear invariants for general models.

As for quadratic invariants, Drolet and Sankoff (1990) extended the Cavender–Felsenstein results for $\mathbf{S}_{\mathbf{JC}}$ from $k = 2$ to $k \geq 2$. Ferretti and Sankoff (1993) considered the same problem and found all linear and quadratic invariants for the trees in Fig. 1. Sankoff (1990) showed how to extend some of the Cavender–Felsenstein results to arbitrarily large N. Fu and Li (1992a) showed necessary and sufficient condition for the existence of a certain class of quadratic invariants for a semigroup S of symmetric Markov matrices.

Evans and Speed (1993) proposed an analytic approach for finding the set of polynomial invariants for an arbitrary tree whenever the matrices in **S** can be considered to describe random walks on an abelian group on $\{1, \ldots, k\}$, *i.e.* when **S** has an infinitesimal generator M′ whose elements $M'(i,j)$, $1 \leq i,j \leq k$, are of the form $M'(i,j) = q(|j - i|)$. They illustrated their method for $\mathbf{S}_{\mathbf{3K}}$, the 3-parameter model of Kimura (1981):

$$\begin{bmatrix} 1-a-b-c & a & b & c \\ a & 1-a-b-c & c & b \\ b & c & 1-a-b-c & a \\ c & b & a & 1-a-b-c \end{bmatrix} \tag{4}$$

where $a > b$ and $a > c$. Note that if the matrices in a semigroup **S** are asymmetric, then it cannot have an infinitesimal generator of form $M'(i,j) = q(|j - i|)$.

Székeley *et al.* (1993) also considered random walks on {1, . . . , k} and described a new class of invariants that arises in the recent extension of the Hendy–Penny 'spectral analysis' from two-state to four-state character sequences (Steel *et al.*, 1992)

THE PROBLEM OF SKEWED BASE COMPOSITION

The phylogenetic inference problems that have been attacked through the invariants approach have all been developed in the context of uniform base composition. Despite the fact that we may have invariance overall π, the symmetric nature of the M, including symmetries not only of form M(i,j) = M(j,i), but also of form M(i,j) = M(i,k) for some i,j,k, ensures that the base composition at the terminal vertices tends to be uniform. Exceptions are $\mathbf{S_{PD}}$, which has only recently been investigated (Steel *et al.* 1993b), and $\mathbf{S_{CAV}}$, which has only been shown to have linear invariants.

When base compositions are skewed, the major dimension of variability is the proportion of A and T versus the proportion of C and G. For example, in the comparative study of mitochondria (Clark-Walker, 1992; Gray, 1992), we note that the base composition of certain fungal mitochondrial DNAs can be as extreme as 82% A + T, for yeast. The $\mathbf{S_{CAV}}$ model (3) was not proposed to take into account skewed base compositions, but rather to loosen the constraints of the Kimura models (2) and (4) where there is but one transition probability and one or two transversion probabilities, respectively.

Analyzing such data with phylogenetic invariants derived from symmetric models is an invitation to systematic error. This paper investigates the possibility of deriving and using invariants from asymmetric models.

Difficulties with asymmetric models: closure

Modeling evolution by a semigroup **S** with less than k(k − 1) parameters has both biological significance and mathematical consequences. The fewer the parameters, the more constrained the model, so that the 1-parameter $\mathbf{S_{JC}}$, for example, is too constrained to be considered a realistic model in most contexts. $\mathbf{S_{PD}}$, on the other hand, may be considered too unconstrained and to ignore biologically accepted relationships. $\mathbf{S_{2K}}$ and $\mathbf{S_{3K}}$ incorporate biologically meaningful constraints. Note, however, the nature of this relationship. In $\mathbf{S_{2K}}$, for example, we have for any tree edge only that all transitions have the same probability and that all transversions have the same probability. There is no fixed relationship between these two probabilities that holds true from edge to edge in the tree. The key fact, however, is that the constraints embodied in the model hold not only on a single edge, but between the starting and ending points X and Y of any multiedge path in the tree. This is essential for the coherence of the model. Whether the path between X and Y contains several edges or just a single edge is only a consequence of the availability or not of data on intermediate organisms and should not affect the applicability of an evolutionary model for the mutations intervening between the organisms represented by X and Y. Mathematically, this is just the property of closure characterizing semigroups. The product of several Kimura matrices is also a Kimura matrix, including the constraint that the transition probability is greater than the transversion probability.

Thus, if we wish to model an evolutionary process by defining a representative matrix, we must ensure that closure holds. For example, suppose we wished to define an asymmetric process where A–T transversion probabilities and C–G transversion probabilities were each symmetric and equal to each other (=a). Then we might want one large transition probability c_2 and one relatively large transversion probability b_2 in the direction of A and T, and smaller ones c_1 and b_1 in the direction of C and G. The set **S** of matrices satisfying this would each have the form (N.B. Bases not in standard AGCT order.):

$$\begin{array}{c} \\ \mathrm{A} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \end{array} \\ \begin{bmatrix} - & a & b_1 & c_1 \\ a & - & c_1 & b_1 \\ b_2 & c_2 & - & a \\ c_2 & b_2 & a & - \end{bmatrix} \end{array} \tag{5}$$

Multiplying two such matrices together, however, produces a new matrix where the A $\leftrightarrow$ T transitions do not necessarily have the same probabilities as the C $\leftrightarrow$ G. The set **S**, therefore, is not a semigroup and hence (5) does not constitute a coherent model of evolution. Adding an additional parameter gives the following type of matrix, and a genuine semigroup $\mathbf{S_{AS}}$:

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} - & a_1 & b_1 & c_1 \\ a_1 & - & c_1 & b_1 \\ b_2 & c_2 & - & a_2 \\ c_2 & b_2 & a_2 & - \end{bmatrix} \end{array} \tag{6}$$

This is a mathematical convenience, however, for there is no immediate biological motivation for different transversion parameters a_1 and a_2 in (6). On the other hand, it is not unreasonable.

Another mathematical consequence of having few parameters in a model is that it may imply certain symmetries among the N-tuple probabilities in the spectrum. These are invariants (not phylogenetic invariants, however) and may be used to reduce the complexity of the search for phylogenetic invariants. For example, with $\mathbf{S_{JC}}$ where $\pi = (1/2,1/2)$, we have for any **T**, $f(1,2,2,2) - f(2,1,1,1) = 0$, $f(2,1,1,2) - f(1,2,2,1) = 0$, etc.

THE METHOD

We denote by $\mathbf{f} = (f_1, \ldots, f_kN)$ the probability distribution of N-tuples for a given rooted tree **T**, a given root distribution π, and a given set of matrices $\mathbf{M} = \{M_1, \ldots, M_E\}$ from **S** associated with the E edges of **T**. Recall that for rooted binary trees, $E = 2N - 2$. We wish to find all invariants Q having a specific form

$$Q = Q(\mathbf{f},\lambda). \tag{7}$$

where λ represents a vector of coefficients. The problem becomes that of determining all λ for which the function Q is invariant over all **M** and all π, *i.e.*, identically equal to zero, independent of the specific parameters associated with each of the edges and the root.

Since Q is to be invariant with respect to the parameters of the model, we simply choose m sets $\pi(i)$ of root distributions and m sets **M**(i) of matrices for **T** at random, $1 \leq i \leq m$, calculate explicitly the distribution **f**(i) for each set, and set up the system:

$$\begin{gathered} Q(\mathbf{f}(1),\lambda) = 0 \\ \ldots \\ Q(\mathbf{f}(m),\lambda) = 0 \end{gathered} \tag{8}$$

The set of invariants having form Q is necessarily contained in the set of nontrivial solutions of this system.

Consider for example the case of quadratic invariants. The function Q in (7) is of form

$$Q(\mathbf{f},\lambda) = \Sigma_{1 \leq i \leq j \leq k^N} \lambda_{ij} f_i f_j, \tag{9}$$

and then the equations in (8) can be written as a system of homogeneous linear equations in the unknown λ_{ij}

$$\mathbf{G}\,\lambda = \mathbf{0} \tag{10}$$

where **G** is a $m \times v$ matrix, $v = k^N(k^N + 1)/2$, with elements

$$g_{hn} = f_i(h)f_j(h), \tag{11}$$

λ_{ij} being the *n*-th component of λ. The set of solutions to (10) defines the kernel of the matrix **G**, denoted Ker(**G**). This is a vector subspace of dimension equal to $v - \text{rank}(\mathbf{G})$ for which the simplest basis is a set of vectors expressing the linear dependences existing among the columns of **G**.

The key to the choice of the m π and **M** is to assure that there are no extra solutions to (10) because of accidental dependences among its columns. This can be assured by making m as large as feasible so that

any accident must be an extremely improbable multiple coincidence, and by choosing random positive parameters, so that the set of such accidents has measure zero. In practice, of course, with pseudo-random generators this cannot be assured, but this is of no mathematical importance because spurious invariants are, as we shall see, easily detected and discarded. As we shall also see, however, it is of practical importance to keep the number of candidate invariants as small as possible.

Since **G** is a real matrix, the precise solution of (10), and of the larger problems of this sort we encounter with our method, becomes computationally cumbersome. It is easier to embed $\mathbf{G}\lambda = \mathbf{0}$ in a multiple regression problem

$$\mathbf{G}\lambda = \mathbf{0} + \epsilon \tag{12}$$

where each row of **G** is an observation of the v independent regressor variables and **0** contains the values (all zero) of the dependent variable. We can then be sure that our estimate of λ has good properties. Given that the key quantity in this problem is rank (**G**), it is not necessary to take $m > v$.

SOME ASYMMETRIC MODELS WITH NO QUADRATIC INVARIANTS

Consider the phylogenetic tree $\mathbf{T_1}$ in Fig. 2 for the simple case of N = 3 organisms A, B, C. We suppose k = 2, root distribution $\pi = (\pi_1, \pi_2)$ and semigroup $\mathbf{S_{PD}}$ consisting of generally asymmetric matrices M of form

$$\begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} \tag{13}$$

There are nine unknown parameters in the inference problem: two (a and b) for each of the four tree edges, and π_1, since $\pi_2 = 1 - \pi_1$.

We use the following abbreviated notation for the components of the spectrum $\mathbf{f} = (f_1, \ldots, f_8)$:

$$\begin{array}{ll} f_1 = f(1,1,1) & f_5 = f(2,1,1) \\ f_2 = f(1,1,2) & f_6 = f(2,1,2) \\ f_3 = f(1,2,1) & f_7 = f(2,2,1) \\ f_4 = f(1,2,2) & f_8 = f(2,2,2) \end{array} \tag{14}$$

where for all $\alpha,\beta,\gamma \in \{1,2\}$,

$$f(\alpha,\beta,\gamma) = \Sigma_{\epsilon,\phi\in\{1,2\}}\pi_\phi M_{\rho E}(\phi,\epsilon)M_{\rho C}(\phi,\gamma)M_{EA}(\epsilon,\alpha)M_{EB}(\epsilon,\beta). \tag{15}$$

Note that $\Sigma_{1\leq i\leq 8} f_i = 1$. This means that we have a system of only seven equations of form (15) relating the f_i and the model parameters, which are not sufficient to eliminate the parameters (which number nine) to produce an invariant of form (9), or indeed of any form (7). Even if the spectrum has more components than there are parameters in the model, we cannot necessarily find an invariant of specific form (9). For example, consider the case N = 3 and the semigroup $\mathbf{S_{AS}}$ in (6) for k = 4. For this problem, there are $4^3 = 64$ frequencies f_i in the spectrum, and only 27 model parameters: six (a_1, b_1, c_1, a_2, b_2, and c_2) for each of the four tree edges, and three of the four probabilities in π. Thus, there is no *a priori* reason, counting equations of form

$$f(\alpha,\beta,\gamma) = \Sigma_{\epsilon,\phi\in\{1,2,3,4\}}\pi_\phi M_{\rho E}(\phi,\epsilon)M_{\rho C}(\phi,\gamma)M_{EA}(\epsilon,\alpha)M_{EB}(\epsilon,\beta) \tag{16}$$

versus number of model parameters, to exclude the possibility of invariants.

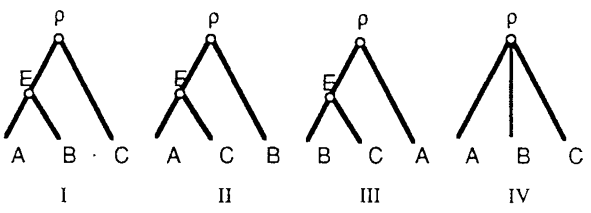

FIG. 2. The four rooted trees on three terminal vertices.

We first construct the matrix **G**. To do this we randomly choose $m = v = 4^3(4^3 + 1)/2 = 2{,}080$ points $\mathbf{p}(1), \ldots, \mathbf{p}(2{,}080)$ in 27-dimensional space to be the parameters in the various M and in π. We can then accurately calculate the 64 $f(\alpha,\beta,\gamma)$ for each of the 2,080 spectra $\mathbf{f}(1), \ldots, \mathbf{f}(2{,}080)$. Using all of these values we can construct the $2{,}080 \times 2{,}080$ matrix **G**. The set of quadratic invariants must be in Ker(**G**), but we find (using routines from the IMSL library) that rank (**G**) = 2,080, so that Ker(**G**) = {0}. Therefore, no quadratic invariant can exist for this model. This implies, as well, that no linear invariants exist, *i.e.*, no invariants of form $Q(\mathbf{f},\lambda) = \Sigma_{1 \leq i \leq k} N\, \lambda_i f_i$; for if one did exist, denote it by L. Then $\{f_i \times L : 1 \leq i \leq 64\}$CKer(**G**) would all be quadratic invariants.

A TRACTABLE SEMIGROUP OF ASYMMETRIC MATRICES

The previous two sections lead to three conclusions: first, that it is not always trivial to define a constrained semigroup modeling a given biological concept; second, that even if such a semigroup is found, there may be no low-degree polynomial phylogenetic invariants; and third, that our "empirical" methodology quickly becomes computationally difficult as the number of parameters increases.

In the search for a tractable model for the evolution of skewed base composition, we postulated the following scenario. A group of organisms evolves in similar environments, possibly different from that of their common ancestor, and this environment puts a constant asymmetric pressure on mutation tendencies, so that the changes of changing state in one direction is a constant times that of the other direction, this constant being universal across the group of organisms.

Thus, given a constant $c > 1$, we consider matrices of form:

$$M = \begin{bmatrix} 1 - a & a \\ c\,a & 1 - c\,a \end{bmatrix} \tag{17}$$

where a varies between 0 and 1/2c. It is easy to prove that these matrices form a semigroup, which we denote $\mathbf{S_c}$. This has been noted independently by Steel *et al.* (1993b). In a biological context, we might imagine that c might range as high as 2 or 3.

Consider first the case of N = 3 species and the tree $\mathbf{T_I}$ of Fig. 2. The parameter space now is 5-dimensional (the probability π_1 and the 4 matrix parameters a) and the matrix **G**, as computed from $v = 2^3(2^3 + 1)/2$, is 36×36.

We find rank (**G**) = 35 for c = 1,2,3, . . . so that a basis of the subspace Ker(**G**) contains only one element. Simply by inspecting this invariant for different values of c, we find that they can all be expressed as

$$f_2f_3 - f_1f_4 - f_2f_5 + (c - 1)\, f_4f_5 + f_1f_6 - (c - 1)\, f_3f_6 + cf_4f_7 - cf_6f_7 - cf_3f_8 + cf_5f_8. \tag{18}$$

The facts that our λ are, strictly speaking, only estimated by the computer program on the basis of a (pseudo)-random sample, and that the form of the hypothesized invariant is derived by inspection, might seem to relegate to the realm of conjecture its true invariant status. This is not the case, however. Substituting (17) into (16) (distinguishing among the four a's), and then (16) into (18), proves explicitly that the form is invariant for all real c. That this latter type of calculation is impractical in many cases without the use of symbolic computing does not detract from the certainty of the result.

By counting the degrees of freedom of the spectrum **f**, it can be seen that this is one of only two possible invariants for this model. Since **f** is confined to a space of 7 dimensions and there are 5 parameters, the invariant constitutes one of the two additional constraints possible on **f**. Our results ensure that the other cannot be a quadratic invariant.

It can be seen that our invariant is a phylogenetic invariant, since not only is it identically zero for $\mathbf{T_I}$ in Fig.2, but it is non-zero and varies with the edge matrix parameters when applied to spectra **f** for trees $\mathbf{T_{II}}$ and $\mathbf{T_{III}}$. Since $\mathbf{T_{IV}}$ represents a degenerate case of $\mathbf{T_I}$ (in the matrix associated with ρE, a = 0, representing a zero edge "length"), the formula is also invariant for this case.

APPLICATION TO FOUR SPECIES

We now turn to the more difficult case of N = 4 organisms.

We start with tree $\mathbf{T_1}$ in Fig. 3, which depicts the 26 rooted trees with N = 4 terminal vertices. The spectrum **f** has 16 terms for which we use the following abbreviated notation:

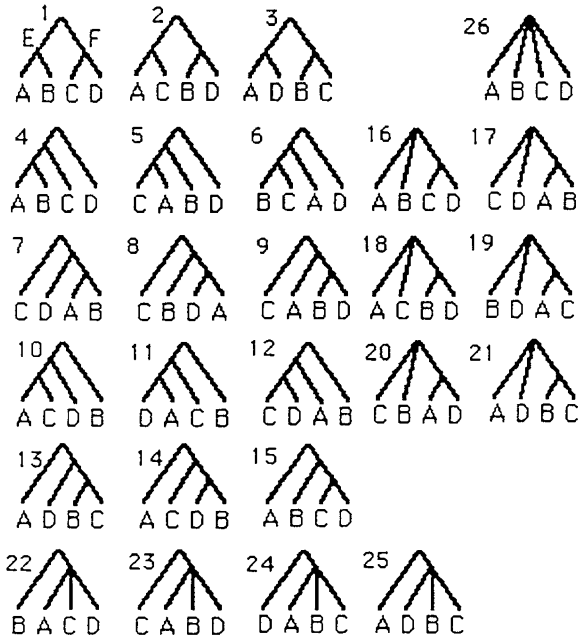

FIG. 3. The 26 rooted trees on $N = 4$ terminal vertices.

$$
\begin{aligned}
f_1 &= f(1,1,1,1) & f_9 &= f(2,1,1,1)\\
f_2 &= f(1,1,1,2) & f_{10} &= f(2,1,1,2)\\
f_3 &= f(1,1,2,1) & f_{11} &= f(2,1,2,1)\\
f_4 &= f(1,1,2,2) & f_{12} &= f(2,1,2,2)\\
f_5 &= f(1,2,1,1) & f_{13} &= f(2,2,1,1)\\
f_6 &= f(1,2,1,2) & f_{14} &= f(2,2,1,2)\\
f_7 &= f(1,2,2,1) & f_{15} &= f(2,2,2,1)\\
f_8 &= f(1,2,2,2) & f_{16} &= f(2,2,2,2)
\end{aligned}
\tag{19}
$$

where, for $\alpha,\beta,\gamma,\delta\epsilon\{1,2\}$,

$$f(\alpha,\beta,\gamma,\delta) = \Sigma_{\epsilon,\phi,\psi\epsilon\{1,2\}}\ \pi_\psi M_{\rho E}(\psi,\epsilon)M_{\rho F}(\psi,\phi)M_{EA}(\epsilon,\alpha)M_{EB}(\epsilon,\beta)\ M_{FC}(\phi,\gamma)M_{FD}(\phi,\delta) \tag{20}$$

Applying our method, we find a 136×136 matrix **G** of rank 125 for each of $c = 1,2,3, \ldots$. The 11 quadratic invariants making up Ker(**G**) can be expressed in terms of c as follows:

$$
\begin{aligned}
Q_1 &= f_2f_5 - f_3f_5 - f_1f_6 + (c-1)f_3f_6 + cf_4f_6 + f_1f_7 - (c-1)f_2f_7 - cf_4f_7 - cf_2f_8 + cf_3f_8\\
Q_2 &= f_2f_9 - f_3f_9 - f_1f_{10} + (c-1)f_3f_{10} + cf_4f_{10} + f_1f_{11} - (c-1)f_2f_{11} - cf_4f_{11} - cf_2f_{12} + cf_3f_{12}\\
Q_3 &= f_2f_5 - f_1f_6 - f_2f_9 + (c-1)f_6f_9 + f_1f_{10} - (c-1)f_5f_{10} + cf_6f_{13} - cf_{10}f_{13} - cf_5f_{14} + cf_9f_{14}\\
Q_4 &= f_3f_5 - f_1f_7 - f_3f_9 + (c-1)f_7f_9 + f_1f_{11} - (c-1)f_5f_{11} + cf_7f_{13} - cf_{11}f_{13} - cf_5f_{15} + cf_9f_{15}\\
Q_5 &= f_4f_6 - f_2f_8 - f_4f_{10} + (c-1)f_8f_{10} + f_2f_{12} - (c-1)f_6f_{12} + cf_8f_{14} - cf_{12}f_{14} - cf_6f_{16} + cf_{10}f_{16}\\
Q_6 &= f_2f_{13} - f_3f_{13} - f_1f_{14} + (c-1)f_3f_{14} + cf_4f_{14} + f_1f_{15} - (c-1)f_2f_{15} - cf_4f_{15} - cf_2f_{16} + cf_3f_{16}\\
Q_7 &= f_6f_9 - f_7f_9 - f_5f_{10} + (c-1)f_7f_{10} + cf_8f_{10} + f_5f_{11} - (c-1)f_6f_{11} - cf_8f_{11} - cf_6f_{12} + cf_7f_{12}\\
Q_8 &= f_6f_{13} - f_7f_{13} - f_5f_{14} + (c-1)f_7f_{14} + cf_8f_{14} + f_5f_{15} - (c-1)f_6f_{15} - cf_8f_{15} - cf_6f_{16} + cf_7f_1\\
Q_9 &= f_3f_6 - f_2f_7 - f_3f_{10} + (c-1)f_7f_{10} + f_2f_{11} - (c-1)f_6f_{11} + cf_7f_{14} - cf_{11}f_{14} - cf_6f_{15} + cf_{10}f_{15}\\
Q_{10} &= f_4f_5 - f_1f_8 - f_4f_9 + (c-1)f_8f_9 + f_1f_{12} - (c-1)f_5f_{12} + cf_8f_{13} - cf_{12}f_{13} - cf_5f_{16} + cf_9f_{16}\\
Q_{11} &= f_2f_5 - f_3f_5 - f_1f_6 + (c-1)f_3f_6 + cf_4f_6 + f_1f_7 - (c-1)f_2f_7 - cf_2f_8 - f_2f_9 + f_3f_9 + (c-1)f_6f_9\\
&\quad - (c-1)f_7f_9 + f_1f_{10} - (c-1)f_3f_{10} - cf_4f_{10} - (c-1)f_5f_{10} + (c-1)(c-1)f_7f_{10}\\
&\quad + c(c-1)f_8f_{10} - f_1f_{11} + (c-1)f_2f_{11} + (c-1)f_5f_{11} - (c-1)^2f_6f_{11} + cf_2f_{12}\\
&\quad - c(c-1)f_6f_{12} + cf_6f_{13} - cf_7f_{13} - cf_5f_{14} + c(c-1)f_7f_{14} + c^2f_8f_{14} + cf_5f_{15}\\
&\quad - c(c-1)f_6f_{15} - c^2f_{12}f_{15} - c^2f_6f_{16} + c^2f_{11}f_{16}
\end{aligned}
\tag{21}
$$

Once again, the invariant status of these11 polynomials was proved for all c using symbolic computing.

REMOVING ALGEBRAIC DEPENDENCE

Because of how they were derived, the polynomials in (21) are linearly independent, *i.e.*, for any quadratic equation of form

$$\Sigma_{j\in[1,11]}A_jQ_j = 0, \tag{22}$$

it must be that $A_j = 0$, for all j, $1 \leq j \leq 11$ and for all probability distributions $\mathbf{f} = (f_1, f_2, \ldots, f_{16})$. Our goal, however, is to find the smallest set of invariants that algebraically spans the set of all invariants. A linearly independent set of invariants could still contain algebraically functionally dependent elements which, ideally, we would like to exclude. For example, does (21) contain elements which are cubically related? *i.e.*, are there coefficients A_{ij} such that

$$\Sigma_{i\in[1,16]j\in[1,11]}A_{ij}f_iQ_j = 0? \tag{23}$$

This question may also be investigated "empirically." For fixed c, we evaluate for m randomly generated probability distributions $\mathbf{f}(h)$, $1 \leq h \leq m$ (not necessarily spectra since they are not generated by any process over a fixed tree), the 176 quantities $f_i(h)Q_j[\mathbf{f}(h)]$. The 176 terms derived from each probability distribution form one of the m rows of a mtrix $\mathbf{H}$ (we may take $m = 176$). Then $\mathrm{Ker}(\mathbf{H})$ represents the set of dependencies of form (23) among the polynomials $Q_1, \ldots, Q_{11}$.

The results for values of $c = 1,2, \ldots$ can be written in terms of the following five cubic relations:

$$\begin{gathered}
f_{10}Q_1 - f_{11}Q_1 - f_6Q_2 + f_7Q_2 + f_2Q_7 - f_3Q_7 = 0 \\
f_{14}Q_1 - f_{15}Q_1 - f_6Q_6 + f_7Q_6 + f_2Q_8 - f_3Q_8 = 0 \\
f_7Q_3 - f_{11}Q_3 - f_6Q_4 + f_{10}Q_4 + f_5Q_9 - f_9Q_9 = 0 \\
-f_1Q_1 + (c-1)f_9Q_1 + cf_{13}Q_1 + f_1Q_2 - (c-1)f_5Q_2 - cf_{13}Q_2 - cf_5Q_6 + cf_9Q_6 + f_1Q_3 \\
- (c-1)f_3Q_3 - cf_4Q_3 - f_1Q_4 + (c-1)f_2Q_4 + cf_4Q_4 + cf_2Q_{10} - cf_3Q_{10} = 0
\end{gathered} \tag{24}$$

$$\begin{gathered}
(c-1)f_2Q_1 - (c-1)^2f_{10}Q_1 - c(c-1)f_{14}Q_1 - (c-1)f_2Q_2 + (c-1)^2f_6Q_2 + cf_{13}Q_2 - c^2f_{16}Q_2 \\
+ c(c-1)f_6Q_6 - cf_9Q_6 + c^2f_{12}Q_6 - f_1Q_3 + (c-1)f_3Q_3 + cf_4Q_3 + f_1Q_4 - (c-1)f_2Q_4 - cf_4Q_4 \\
- cf_1Q_5 + c(c-1)f_3Q_5 + c^2f_4Q_5 + f_1Q_{11} - (c-1)f_2Q_{11} - cf_4Q_{11} = 0.
\end{gathered}$$

The discovery of these nonlinear dependencies allows us to eliminate the five polynomials Q_7, Q_8, Q_9, Q_{10}, and Q_{11}. No further cubic relations exists among the six remaining invariants.

We must go a step further and show in the same way that there are two nontrivial quartic relations[4] among them, shown on (25) and (26):

$$\begin{gathered}
f_{10}f_{13}Q_1 - f_{11}f_{13}Q_1 - f_9f_{14}Q_1 + (c-1)f_{11}f_{14}Q_1 + cf_{12}f_{14}Q_1 + f_9f_{15}Q_1 - \\
(c-1)f_{10}f_{15}Q_1 - cf_{12}f_{15}Q_1 - cf_{10}f_{16}Q_1 + cf_{11}f_{16}Q_1 - f_6f_{13}Q_2 + f_7f_{13}Q_2 + f_5f_{14}Q_2 \\
- (c-1)f_7f_{14}Q_2 - cf_8f_{14}Q_2 - f_5f_{15}Q_2 + (c-1)f_6f_{15}Q_2 + cf_8f_{15}Q_2 + cf_6f_{16}Q_2 - cf_7f_{16}Q_2 \\
+ f_6f_9Q_6 - f_7f_9Q_6 - f_5f_{10}Q_6 + (c-1)f_7f_{10}Q_6 + cf_8f_{10}Q_6 + f_5f_{11}Q_6 - (c-1)f_6f_{11}Q_6 \\
- cf_8f_{11}Q_6 - cf_6f_{12}Q_6 + cf_7f_{12}Q_6 = 0
\end{gathered} \tag{25}$$

$$\begin{gathered}
f_1f_2Q_1 - (c-1)f_1f_6Q_1 - (c-1)f_2f_9Q_1 + (c-1)^2f_6f_9Q_1 - cf_2f_{13}Q_1 + c(c-1)f_6f_{13}Q_1 \\
- cf_1f_{14}Q_1 + c(c-1)f_9f_{14}Q_1 + c^2f_{13}f_{14}Q_1 - f_1f_2Q_2 + (c-1)f_2f_5Q_2 + (c-1)f_1f_6Q_2 \\
- (c-1)^2f_5f_6Q_2 + cf_2f_{13}Q_2 - c(c-1)f_6f_{13}Q_2 + cf_1f_{14}Q_2 - c(c-1)f_5f_{14}Q_2 - c^2f_{13}f_{14}Q_2 \\
+ cf_2f_5Q_3 - c(c-1)f_5f_6Q_3 - cf_2f_9Q_3 + c(c-1)f_6f_9Q_3 - c^2f_5f_{14}Q_3 + c^2f_9f_{14}Q_3 - f_1f_2Q_4 \\
+ (c-1)f_2f_3Q_4 + cf_3f_4Q_4 + (c-1)f_1f_6Q_4 - (c-1)^2f_3f_6Q_4 - c(c-1)f_4f_6Q_4 + c(c-1)f_2f_8Q_4 \\
- c(c-1)f_3f_8Q_4 + cf_1f_{14}Q_4 - c(c-1)f_3f_{14}Q_4 - c^2f_4f_{14}Q_4 + c^2f_2f_{16}Q_4 - c^2f_3f_{16}Q_4 + f_1f_2Q_5 \\
- (c-1)f_2f_2Q_5 - cf_2f_4Q_5 - (c-1)f_1f_6Q_5 + (c-1)^2f_2f_6Q_5 + c(c-1)f_4f_6Q_5 - cf_1f_{14}Q_5 \\
+ c(c-1)f_2f_{14}Q_5 + c^2f_4f_{14}Q_5 - cf_1f_2Q_6 + cf_1f_3Q_6 + c(c-1)f_2f_5Q_6 \\
- c(c-1)f_3f_5Q_6 + c^2f_2f_{13}Q_6 - c^2f_3f_{13}Q_6 = 0.
\end{gathered} \tag{26}$$

[4]By non-trivial quartic relation, we mean a relation of the form $\Sigma_{k\in[1,6]}(\Sigma_{1\leq i\leq j\leq 16}A_{ij}f_if_j)Q_k = 0$ in which each polynomial $(\Sigma_{1\leq i\leq j\leq 16}A_{ij}f_if_j)$, $1 \leq k \leq 6$, is not itself an invariant. This is to avoid trivial relations such $(Q_i)Q_j - (Q_j)Q_i = 0$.

This involves a 816 × 816 matrix of rank 798.

Using (25), we can eliminate Q_6. Furthermore, we can substitute for Q_6 in (26), giving an algebraic relation among the five remaining invariants, so that we can also eliminate Q_5. This leaves the four invariants:

$$\begin{aligned}
Q_1 &= f_2f_5 - f_3f_5 - f_1f_6 + (c-1)f_3f_6 + cf_4f_6 + f_1f_7 - (c-1)f_2f_7 - cf_4f_7 - cf_2f_8 + cf_3f_8 \\
Q_2 &= f_2f_9 - f_3f_9 - f_1f_{10} + (c-1)f_3f_{10} + cf_4f_{10} + f_1f_{11} - (c-1)f_2f_{11} - cf_4f_{11} \\
&\quad - cf_2f_{12} + cf_3f_{12} \\
Q_3 &= f_2f_5 - f_1f_6 - f_2f_9 + (c-1)f_6f_9 + f_1f_{10} - (c-1)f_5f_{10} + cf_6f_{13} - cf_{10}f_{13} \\
&\quad - cf_5f_{14} + cf_9f_{14} \\
Q_4 &= f_3f_5 - f_1f_7 - f_3f_9 + (c-1)f_7f_9 + f_1f_{11} - (c-1)f_5f_{11} + cf_7f_{13} - cf_{11}f_{13} \\
&\quad - cf_5f_{15} + cf_9f_{15}
\end{aligned} \tag{27}$$

There is certainly one and possibly two additional dependencies among these four functions, expressible as linear combinations where the coefficients are quotients of polynomials in the **f**, but for the purposes of testing our approach on real data, we will retain all four invariants. We have not yet established the number (one or two) of these dependencies and, even more important, when working with finite amounts of data, there is no harm in having some "backup" invariants. No function is likely to take on the value zero exactly, so comparing two or more invariants for one tree against a comparable set for another tree can lend an additional measure of confidence to the analysis.

The quadratic formulae in (27) are phylogenetic invariants. Each is invariant for certain of the trees in Fig. 3 and not for others. This information is summarized in Table 1, where the invariance (or not) of each of $Q_1, \ldots, Q_4$ is given for a spectrum **f** calculated according to the 26 topologies of Fig. 3.

The invariants for trees $\mathbf{T_2}$ and $\mathbf{T_3}$ can be obtained from $Q_1, \ldots, Q_4$ by simple permutation of the frequencies f_i. For example, the role of f_4 in the context of $\mathbf{T_1}$ is played by f_6 in $\mathbf{T_2}$. To find the invariants for $\mathbf{T_4}$, however, we must carry out the complete procedure as for $\mathbf{T_1}$. After constructing the matrix **G**, solving the system (10), and eliminating the algebraic dependencies, we finally arrive at just two

TABLE 1. INVARIANT STATUS OF QUADRATIC FUNCTIONS FOR THE TREES IN FIG. 3

	Q_1	Q_2	Q_3	Q_4
$\mathbf{T_1}$	I	I	I	I
$\mathbf{T_2}$	NI	NI	NI	NI
$\mathbf{T_3}$	NI	NI	NI	NI
$\mathbf{T_4}$	NI	NI	I	I
$\mathbf{T_5}$	NI	NI	NI	NI
$\mathbf{T_6}$	NI	NI	NI	NI
$\mathbf{T_7}$	NI	NI	I	I
$\mathbf{T_8}$	NI	NI	NI	NI
$\mathbf{T_9}$	NI	NI	NI	NI
$\mathbf{T_{10}}$	NI	NI	NI	NI
$\mathbf{T_{11}}$	NI	NI	NI	NI
$\mathbf{T_{12}}$	I	I	NI	NI
$\mathbf{T_{13}}$	NI	NI	NI	NI
$\mathbf{T_{14}}$	NI	NI	NI	NI
$\mathbf{T_{15}}$	I	I	NI	NI
$\mathbf{T_{16}}$	I	I	I	I
$\mathbf{T_{17}}$	I	I	I	I
$\mathbf{T_{18}}$	NI	NI	NI	NI
$\mathbf{T_{19}}$	NI	NI	NI	NI
$\mathbf{T_{20}}$	NI	NI	NI	NI
$\mathbf{T_{21}}$	NI	NI	NI	NI
$\mathbf{T_{22}}$	I	I	NI	NI
$\mathbf{T_{23}}$	NI	NI	I	I
$\mathbf{T_{24}}$	NI	NI	I	I
$\mathbf{T_{25}}$	I	I	NI	NI
$\mathbf{T_{26}}$	I	I	I	I

I, Invariant; NI, not invariant.

invariants, which turn out to be Q_3 and Q_4. The invariants for the 11 trees of the same shape as $\mathbf{T}_4$ are obtained by appropriately permuting the frequencies f_i in Q_3 and Q_4.

It can be seen that each of the invariants corresponds to one pair of terminal vertices, either A and B, or C and D; that is, each Q is invariant in all and only those trees where the pair is closely grouped. Thus, Q1 is invariant for just those trees in Table 1 where C and D are at least as closely grouped as either is with A or B; $\mathbf{T}_1$, $\mathbf{T}_{12}$, $\mathbf{T}_{15}$, $\mathbf{T}_{16}$, $\mathbf{T}_{17}$, $\mathbf{T}_{22}$, $\mathbf{T}_{25}$, and $\mathbf{T}_{26}$.

APPLICATION TO SKEWED BASE COMPOSITION

To explore the use of the invariants derived in the previous sections, we examined two groups of four organisms, and attempted to determine the correct tree in each case under our asymmetric model. The data consisted of small subunit ribosomal RNA sequences drawn from a previously constructed data base (D.F. Spencer, unpublished). The sequences have all been previously aligned and we used only positions containing nucleotides in all four organisms, *i.e.*, sequence position containing gaps were eliminated. Each nucleotide was then scored as 1 (A and T) or 2 (C and G) so that our $k = 2$ model could be applied.

For each set of observed values of the spectrum $\mathbf{f}$, the invariants $Q_1, \ldots, Q_4$ are linear functions of c. In the following analysis, we will plot these functions for $0 \leq c \leq 4$, more than covering the range of moderately or strongly asymmetric substitution rates. There are three ways of grouping four organisms to form a maximally resolved tree, namely as trees 1, 2, and 3 in Fig. 3. For each of these candidate trees, we will plot the four corresponding invariants $Q_1, \ldots, Q_4$ in a separate graph. Were the model in (17) perfectly correct, and one of the trees 1, 2, or 3 the "true" one, and sequence length n arbitrarily large, then in the appropriate graph all four functions of c would cross the abscissa at the same point, namely at the value of c in (17). In the other two groups, this condition would not be met. Indeed, we would in general expect none of the invariants for the "wrong" trees to take on the value zero within the range of reasonable values of c. On the other hand, since n is not enormous, and since the model in (17) is at best an approximation, we might expect only some of the invariants for the correct tree to cross the abscissa near the same value of c, perhaps the pair of invariants pertinent to the most clearly related pair of organisms.

We stress that the following analyses are meant uniquely to illustrate the use of the asymmetric invariants. Based on only four organisms each, they do not represent an attempt to provide definitive phylogenies, but rather a way of understanding the potential of our method by confronting it with fairly predictable data sets.

Budding yeasts versus *filamentous Ascomycetes*

The small-subunit rRNA gene in fungal mitochondria is AT-rich, despite the necessity of a great deal of C-G base-pairing in the secondary structure of the RNA. This reflects the even greater AT-richness of the mitochondrial genome as a whole. We compared the two budding yeasts, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* and the two filamentous Ascomycetes, *Podospora anserina* and *Aspergillus nidulans*. After removing gaps, the sequences were of length $n = 947$, and the A + T content was 64, 68, 59, and 58%, respectively (compared to the A + T content of the entire first three genomes of 70, 82, and 70%, respectively). The results of estimating the 4-tuple frequencies $\mathbf{f}$, and substituting them in (27) are portrayed as linear functions of c in Fig. 4a–c. Each of these figures considers two compatible pairs of organisms and portrays four invariants, two for each pair.

We observe first that except in Fig. 4a, one of the Q functions for each pair of organisms is comparatively remote from zero for all reasonable values of c (comparatively, because we are only estimating the $\mathbf{f}$ and hence we cannot rely on Q to be precisely zero, even for the true tree). The relatively small values of all the functions in Fig. 4a, suggests that $\mathbf{T}_1$ is the correct inference. Only the groupings *Podospora–Aspergillus* and *S. pombe–S. cerevisiae* are possibly validated by the invariant analysis. Only for *Podospora–Aspergillus*, however, do both associated functions intersect the c axis at reasonable values of c, namely 1.5–2.5. This would suggest that the more refined asymmetric model in (17) only supports one of the weaker results in $\mathbf{T}_{12}$, $\mathbf{T}_{15}$, $\mathbf{T}_{22}$, or $\mathbf{T}_{25}$, where $c \approx 2$. The latter two trees can be eliminated because each would have required that functions corresponding to certain other pairs cross the c axis in Fig. 4, b and c.

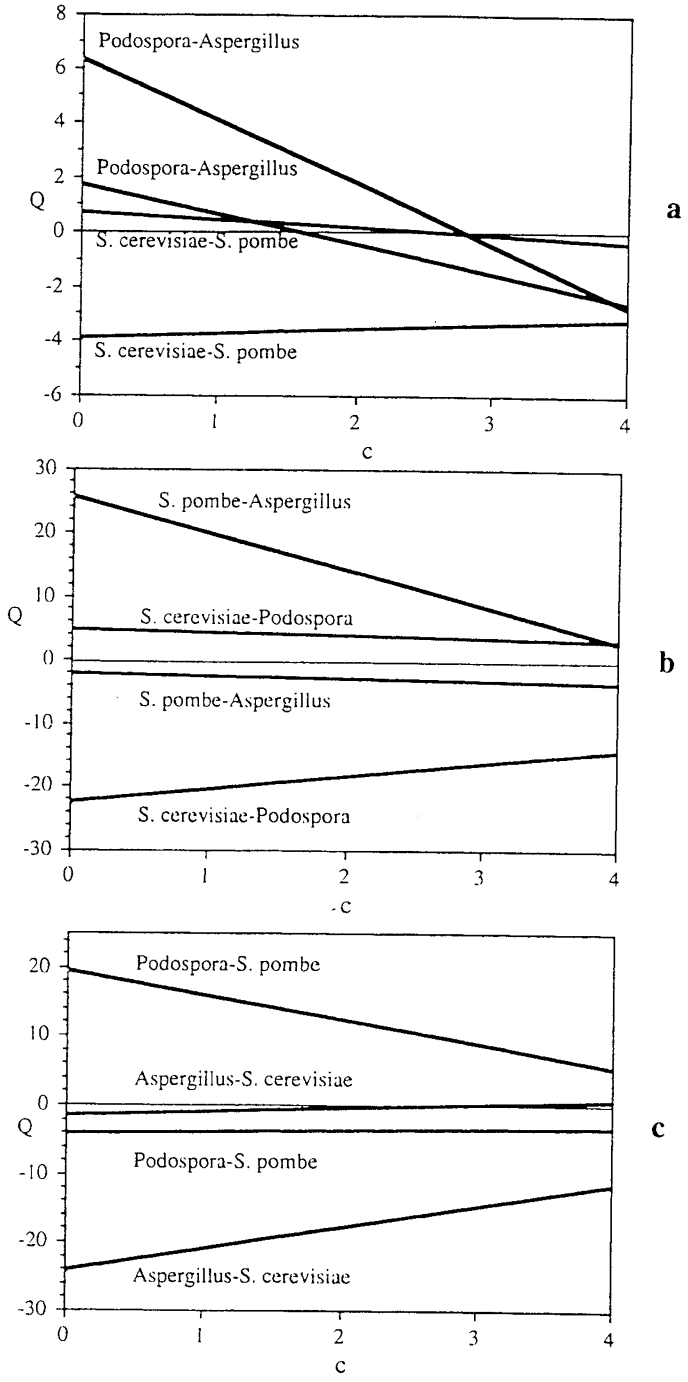

FIG. 4. Invariant formulae as functions of c, calculated from fungal data.

How does this result compare with other phylogenetic analyses of the fungi? *Schizosaccharomyces pombe*, while usually grouped with the budding yeasts, typified by *Saccharomyces cerevisiae*, is usually found to diverge from this group very early, so it is not surprising that with the limited data at hand this grouping does not emerge clearly. On the other hand the two filamentous Ascomycetes, *Aspergillus nidulans* and *Podospora anserina*, are clearly differentiable as a closely related group when compared with either of the other two under any available analysis. We can conclude then that the invariants-based analysis of this particular data set using $Q_1, \ldots, Q_4$ produces meaningful results.

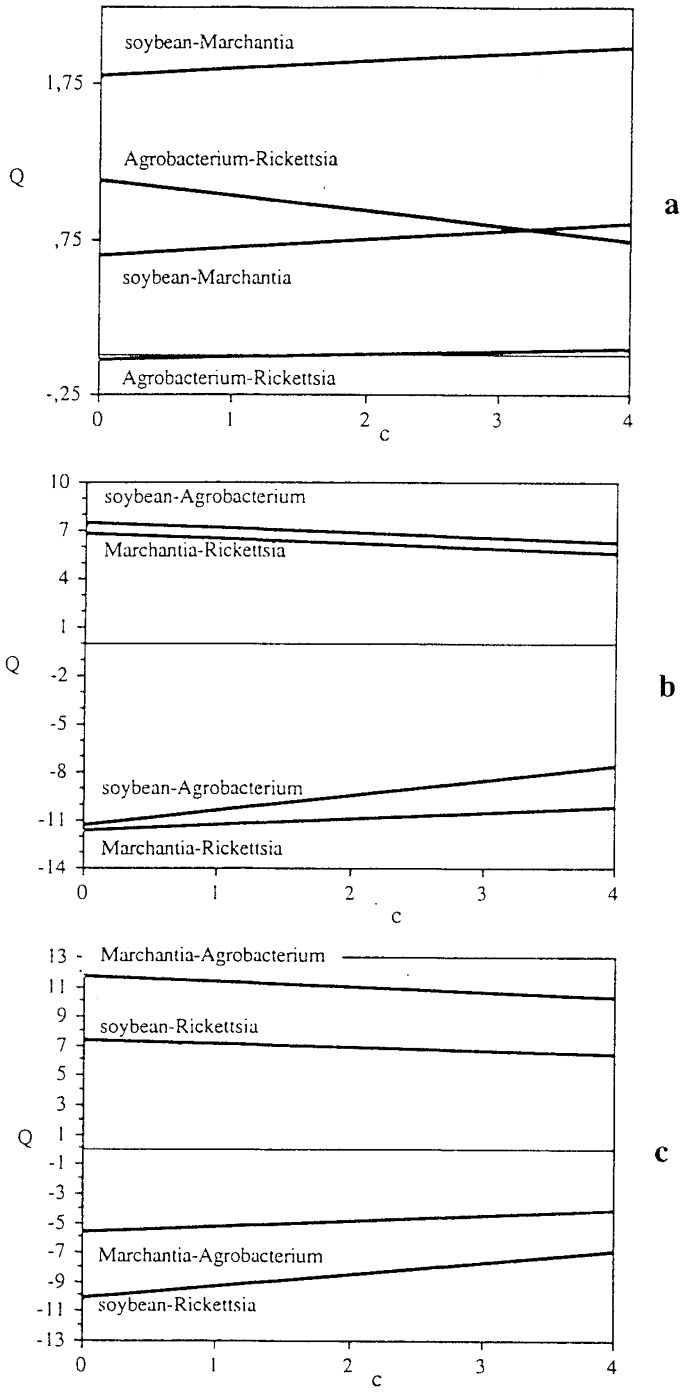

FIG. 5. Invariant formulae as functions of c, calculated from mitochondrial and eubacterial data.

An AT-poor example

The second example compares genes in the mitochondrial genomes of plants, which are known to be evolutionarily very conservative, with their homologs in two eubacterial genomes closely related to the bacterial endosymbiont thought to be ancestral to present-day mitochondria. The mitochondria in question are those of soybean and of *Marchantia polymorpha,* the eubacteria *Agrobacterium tumefaciens* and *Rickettsia rickettsii*. The $n = 960$ ungapped positions of the four sequences show A + T content of 44, 47, 45, and 46%, respectively. While it is clear from the highly reduced scale of Fig. 5a

compared to Fig.5, b and c, that the correct grouping is likely the two plant mitochondria *versus* the two eubacteria, there is no evidence that an asymmetric model pertains. Given the direction of the skewness in the base composition, we might have expected some of the Q to intercept c at values less than 1. This is not the case, and the low A + T ratios may simply reflect an even base composition in the entire genome, with an excess of G − C in the rRNA gene for structural reasons. Indeed, the only one of the four genomes to be entirely sequences to date, *Marchantia,* has A + T content of 58%.

CONCLUSIONS

The study of asymmetric models highlights the versatility of our "empirical" method for finding invariants. The two main drawbacks are the excessive computational requirements for larger problems, and the current lack of a systematic procedure for finding and dealing with all algebraic dependencies.

Work is underway to find invariants for more general asymmetric models. While S_{PD} for $k = 4$ seems out of range of current computational capacities, we have recently identified polynomial invariants for a 10-parameter 4×4 matrix (Ferretti and Sankoff, 1994). (Added in proof: Steel *et al.* 1993b have found invariants of degree 8 for S_{PD}.)

ACKNOWLEDGMENTS

We thank David F. Spencer and Michael W. Gray for providing the aligned SSU rRNA sequences. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada. D.S. and B.F.L. are fellows of the Canadian Institute for Advanced Research.

REFERENCES

Cavender, J.A. 1989. Mechanized derivation of linear invariants. *Mol. Biol. Evol.* 6, 301–316.

Cavender, J.A. 1991. Necessary conditions for the method of inferring phylogeny by linear invariants. *Math. Biosci.* 103, 69–75.

Cavender, J.A., and Felsenstein, J. 1987. Invariants of phylogenies: Simple case with discrete states. *J. Classif.* 4, 57–71.

Clark-Walker, G.D. 1992. Evolution of mitochondrial genomes in fungi. *Int. Rev. Cytol.* 141, 89–127

Drolet, S., and Sankoff, D. 1990. Quadratic tree invariants for multivalued characters. *J. Theoret. Biol.* 144, 117–129.

Evans, S.N., and Speed, T.P. 1993. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.* 21, 355–377.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.

Felsenstein, J. 1983. Inferring evolutionary trees from DNA sequences, 133–150. *In* Weir, B.S., ed., *Statistical Analysis of DNA Sequences,* Marcel Dekker, New York.

Felsenstein, J. 1991. Counting phylogenetic invariants in some simple cases. *J. Theoret. Biol.* 152, 357–376.

Ferretti, V., and Sankoff, D. 1993. The empirical discovery of phylogenetic invariants. *Adv. Appl. Prob.* 25, 290–302.

Ferretti, V. and Sankoff, D. 1994. Phylogenetic invariants for more general evolutionary models. Technical Report 2162, Centre de recherches mathématiques, Université de Montréal.

Fu, Y.X., and LI, W.H. 1992a. Necessary and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree. *Math. Biosci.* 108, 203–218.

Fu, Y.X., and Li, W.H. 1992b. Construction of linear invariants in phylogenetic inference. *Math. Biosci.* 109, 201–228.

Gray, N.W. 1992. The endosymbiont hypothesis revisited. *Int. Rev. Cytol.* 141, 233–357.

Jukes, T.H., and Cantor, C.R. 1969. Evolution of protein molecules, 21–132. *In* Munro, H.N., ed., *Mammalian Protein Metabolism,* Academic Press, New York.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.

Kimura, M. 1981. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 78, 454–458.

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

Lake, J.A. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* 331, 184–186.

Mardia, K.V., Kent, J.T., and Bibby, J.M. 1979. *Multivariate Analysis*. Academic Press, London.

Nguyen, T. and Speed, T.P. 1992. A derivation of all linear invariants for a non-balanced transversion model. *J. Mol. Evol.* 35, 128–143.

Sankoff, D. 1990. Designer invariants for large phylogenies. *Mol. Biol. Evol.* 7, 255–269.

Steel, M.A., Hendy, M.D., Székely, L.A., and Erdös, P.L. 1992. Spectral analysis and a closest tree method for genetic sequences. *Appl. Math. Lett.* 5, 63–67.

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

Steel, M.A., Hendy, M.D., and Penny, D. 1993b. Invertible models of sequence evolution. Dept. of Math., Massey University, New Zealand. Manuscript.

Székely, L.A., Steel, M.A., and Erdos, P.L. 1993. Fourier calculus on evolutionary trees. *Adv. Appl. Math.* 14, 200–216.

Address reprint requests to:
Dr. D. Sankoff
Université de Montreal
CPG128, OUCC Centre-Ville
Montreal, Quebec
H3C 3J7 Canada

Received for publication February 12, 1993; accepted December 3, 1993.