J. theor. Biol. (1995) **173,** 147–162

Phylogenetic Invariants for More General Evolutionary Models

V. Ferretti and D. Sankoff†

CRM, Université de Montréal, Montréal, Quebec, H3C3J7, Canada

(*Received on* 16 *March* 1994, *Accepted on* 13 *September* 1994)

An invariant Q of a tree $\mathbf{T}$ under a k-state Markov model, where a generalized time parameter is identified with the E edges of $\mathbf{T}$, allows us to recognize whether data on N observed species (usually, N DNA sequences, one from each species) can be associated with the N leaves of $\mathbf{T}$ in the sense of having been generated on $\mathbf{T}$ rather than on any other N-leaf tree. The form of the generalized time parameter is a positive determinant matrix in some semigroup $\mathbf{S}$ of Markov matrices. The invariance is with respect to the choice of the set of E matrices in $\mathbf{S}$, one associated with each of the E edges of $\mathbf{T}$. The parametric form of $\mathbf{S}$ represents a model of the evolutionary process. In this paper, we apply a general method of finding invariants of a parametrized functional form to find low-degree polynomial invariants for different models. Quadratic invariants are obtained for the Kimura two-parameter model, for a model allowing evolutionary dependence between positions in the sequences and for an asymmetric model that allows for A + T versus G + C asymmetries in DNA base composition. Those invariants are found for trees (unrooted in case of the Kimura model and rooted for the others) with $N = 3$ or $N = 4$ terminal vertices. We also find cubic invariants for a ten-parameter model with $k = 4$ states, for rooted trees with $N = 4$. In each case, we use implicit function theory to predict the number of algebraically independent invariants and then use this prediction to guide a systematic search for algebraic dependence within the set of invariants produced by our method.

1. Introduction

The method of phylogenetic invariants aims to solve the problem of evolutionary inference, for a given semigroup $\mathbf{S}$ of $k \times k$ substitution matrices ($k = 4$ in the case of nucleic acids), by discovering, for each possible evolutionary tree $\mathbf{T}$, a function $Q_{\mathbf{T}}$ of the data that can be expected to take on the value zero if and only if these data are generated on $\mathbf{T}$. Data are presumed to be generated through the association of some matrix in $\mathbf{S}$ with each edge of $\mathbf{T}$. The parametric form of $\mathbf{S}$ represents a model of the evolutionary process, and each matrix in $\mathbf{S}$ may be considered a generalized length for the edges (branches) of trees. The invariance is with respect to the choice of matrices in $\mathbf{S}$ associated with the edges of $\mathbf{T}$; $Q_{\mathbf{T}} \equiv 0$ whatever $\mathbf{M}_{XY} \in \mathbf{S}$ is associated with edge XY, for each XY.

† Author to whom correspondence should be addressed.

The major successes in the search for phylogenetic invariants to data have taken advantage of specific properties of $\mathbf{S}$: Cavender & Felsenstein (1987) for symmetric 2×2 matrices; Lake (1987) for the two-parameter Kimura (1980) model $\mathbf{S_{2K}}$ (4×4 matrices); Drolet & Sankoff (1990) for the Jukes & Cantor (1969) model $\mathbf{S_{JC}}$ in the case of $k \times k$ matrices, $k \geqslant 2$; Fu & Li (1992*b*), Cavender (1989) and Nguyen & Speed (1992) for classes of six-parameter 4×4 matrices; Evans & Speed (1993) and Steel *et al.* (1993*c*) for the three-parameter Kimura (1981) model $\mathbf{S_{3K}}$ (4×4 matrices); and Székely *et al.* (1993) for models that can be described by random walks on abelian groups.

The invariants themselves are functions of the frequencies (among the positions of aligned nucleic acid or other k-ary sequences) of each (of the k^N) possible N-tuple of states observed at the same sequence position at the N terminal vertices of $\mathbf{T}$,

0022–5193/95/060147 + 16 $08.00/0

representing the N contemporary organisms for which the phylogeny is sought.

We have introduced a method (Ferretti & Sankoff, 1993) for finding invariants without any explicit analysis based on specific qualities of $\mathbf{S}$. In the present paper we apply this to a number of models more general than those previously studied, and discuss the task of finding invariants for the "ultimate" model, where $\mathbf{S}$ consists of all stochastic 4×4 matrices with positive determinant.

2. The Model and the Inference Problem

We denote by $\mathbf{T}$ 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 $\mathbf{T}$ contains at least one non-terminal vertex, its root, denoted ρ, such that the flow of time is directed away from ρ on all edge on the paths joining ρ to the terminal vertices. Each of the non-terminal vertices represents a idealized speciation event, and the edge-length $|XY|$ corresponds to the time elapsed between the speciation (non-terminal) or observation (terminal) events represented by vertices X and Y. We stress that the details of $\mathbf{T}$, including the branching structure connecting its vertices as well as the edge-lengths, is unknown. What we do know, or assume, is an evolutionary model, namely a semigroup $\mathbf{S}$ of Markov matrices, each with positive determinant, on the state space $\{1, \ldots, k\}$, where some $\mathbf{M}_{XY} \in \mathbf{S}$ is to be associated with each edge XY of $\mathbf{T}$. (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 $\mathbf{M}_{XY}$ obviates any reference to the edge-lengths $|XY|$, and constitutes a somewhat more general approach. In all these models, one can identify time with $-\log\det \mathbf{M}$.

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

$$\pi \mathbf{M}_{v_0 v_1} \mathbf{M}_{v_1 v_2} \cdots \mathbf{M}_{v_{r-1} v_r},$$

where $\rho = v_0, v_1, \ldots, v_r = Y$ is the sequence of vertices on the path between ρ and Y. Note that $\mathbf{M}_{v_0 v_1} \mathbf{M}_{v_1 v_2} \cdots \mathbf{M}_{v_{r-1} v_r} \in \mathbf{S}$. The paths from ρ to two different terminal vertices Y_1 and Y_2 necessarily contain some of the same non-terminal vertices $\rho = v_0, v_1, \ldots, v_q = X$ (possibly with $q = 0$). Then the structure of $\mathbf{T}$ 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 $\pi_1 = e_{i_0} \mathbf{M}_{v_0 v_1}$ 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 non-terminal 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. 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 $\mathbf{T}$. 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 & Felsenstein (1987) and Lake (1987, 1988), focuses on estimating the branching structure of $\mathbf{T}$ 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, for example 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 $\mathbf{T}$ and its edge lengths (Felsenstein, 1991).

The idea is to find a function Q_{T} of the data (the spectrum) and of tree topology $\mathbf{T}$ 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 $\mathbf{M} \in \mathbf{S}$ (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 $\mathbf{U}, \mathbf{V}, \ldots$. Then by evaluating the functions Q_{T}, Q_{U}, Q_{V}, etc 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.

3. The Nature and Number of Invariants

For a given tree **T** and a given evolutionary model represented by the semigroup **S**, how many invariants are there? Are invariants necessarily polynomial? If there are μ invariants, can a set of μ algebraically independent invariants be constructed only of polynomials? These questions remain open in general, though partial answers are available. To date, whenever a set of μ invariants have been found for a model, these have been only linear, quadratic and other polynomial invariants.

Felsenstein (1991) suggested, by counting parameters, that the number of invariants should be equal to $k^N - hE$, where k^N is the number of frequencies f_i making up the spectrum **f**, and hE is the number of parameters in the $k \times k$ evolutionary model generating these frequencies, h being the number of parameters in each matrix, $1 \leqslant h \leqslant k(k-1)$, and E the number of edges in the tree **T**†. Intuitively, we should terms of the other hE frequencies by eliminating the parameters from the hE of the equations relating the frequencies to the parameters. The $k^N - hE$ "solutions" would be the invariants.

Formalizing these ideas, the number of invariants in the neighbourhood of a point in parameter space should be $k^N - \text{rank}(D)$, where the D is the matrix of partial derivatives of the frequencies with respect to the parameters. Thus, in certain trees containing vertices of valence 2, it can be shown that $\text{rank}(D) < hE$, so that the intuitive analysis is misleading. Note that the calculation of the rank can be difficult, and that a closed form expression for it valid over the parameter space is not available. In addition, general theory does not assure us that the invariants in one neighbourhood are the same as those in another, even if $\text{rank}(D)$ is the same. If we knew that all invariants were polynomials, answers to these questions could be cleared up.

$$\begin{pmatrix} 1-a-b-c & a & b & c \\ a & 1-a-b-c & c & b \\ b & c & 1-a-b-c & a \\ c & b & a & 1-a-b-c \end{pmatrix},$$

3. Previous Work

A variety of semigroups **S** have been studied, each representing some compromise between the biological reality they are supposed to model and the mathematical feasibility of solving the inference problem. The invariants associated with these models have been discovered using very diverse approaches and are all polynomial functions of the k^N components of the expected spectrum **f**.

The simplest (and hence the least realistic) of the models is that of Jukes & Cantor (1969) generalized to k states, where the $k \times k$ substitution matrices form a semigroup $\mathbf{S_{JC}}$ and have form

$$\mathbf{M}_{XY} = \begin{pmatrix} 1-(k-1)a & a & \cdots & a \\ a & 1-(k-1)a & \cdots & a \\ \cdots & \cdots & \cdots & \cdots \\ a & a & \cdots & 1-(k-1)a \end{pmatrix},$$

where $0 < a < 1-(k-1)a$. When $k = 4$, $\mathbf{S_{JC}}$ is the original Jukes & Cantor (1969) model. For a fixed k, the parameter a completely determines the matrix. Setting $t = -\log(1-ka)$, 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 $\mathbf{M}_{v_0v_1}\mathbf{M}_{v_1v_2}\cdots\mathbf{M}_{v_{r-1}v_r}$. Also $-\log\det\mathbf{M} = t(k-1)$.

A more realistic, three-parameter model for $k = 4$ was proposed by Kimura (1981). We denote this $\mathbf{S_{3K}}$. Its matrices have form:

$$\begin{pmatrix} 1-a-b-c & a & b & c \\ a & 1-a-b-c & c & b \\ b & c & 1-a-b-c & a \\ c & b & a & 1-a-b-c \end{pmatrix},$$

where $a > b$ and $a > c$. Matrices of this form satisfying the additional constraint $b = c$ make up the Kimura (1980) two-parameter model $\mathbf{S_{2K}}$. Another notable model for $k = 4$ was proposed by Cavender (1989) in the context of invariant analysis. The matrices in this model, denoted $\mathbf{S_{CAV}}$, have six independent parameters and are generally asymmetric:

$$\begin{pmatrix} 1-a-2b & a & b & b \\ c & 1-c-2d & d & d \\ p & p & 1-q-2p & q \\ r & r & s & 1-s-2r \end{pmatrix},$$

where $a+b = c+d$ and $s+r = p+q$. $\mathbf{S_{CAV}}$ is a particular case of the model considered by Nguyen &

† For simplicity we assume that the initial distribution is known; otherwise we simply add $k-1$ to the number of model parameters in the following analysis.

Speed (1992) where the semigroup $\mathbf{S}_{\mathrm{NS}}$ is the set of matrices of form

$$\begin{pmatrix} 1-a-(1+\alpha)b & a & b & \alpha b \\ c & 1-c-(1+\alpha)d & d & \alpha d \\ p & \beta p & 1-q-(1+\beta)p & q \\ r & \beta r & s & 1-s-(1+\beta)r \end{pmatrix},$$

where α and β are constants.

Table 1 summarizes the principal results obtained to date. It indicates, for each reference, the number of invariants, and their polynomial degrees, found in various contexts characterized by N, k, π, $\mathbf{T}$ and $\mathbf{S}$.

Some existence results are also available. Cavender (1991) and Fu & Li (1992*a*) independently established necessary and sufficient conditions on $\mathbf{S}$ for the existence of linear invariants. Fu & Li (1991) have also obtained this kind of result for the existence of certain types of quadratic invariants.

4. Method

4.1. THE SEARCH FOR INVARIANTS

We denote by $\mathbf{f} = (f_1, \ldots, f_w)$ the probability distribution of $w = k^N$ N-tuples for a given rooted tree $\mathbf{T}$, a given root distribution π, and a given set of matrices $\mathbf{M} = \{\mathbf{M}_1, \ldots, \mathbf{M}_{2N-2}\}$ from $\mathbf{S}$ associated with the $2N-2$ edges of $\mathbf{T}$. We wish to find all invariants Q having a specific parametric form

$$Q = Q(\mathbf{f}, \boldsymbol{\lambda}), \tag{1}$$

TABLE 1
Summary of invariant literature

	Model					Invariant	
Citation	N	k	$\mathbf{T}$	π	$\mathbf{S}$	d	Methodology
Lakes (1987)	4	4	−R	U	$\mathbf{S}_{2K}$	1	Heuristic
Cavender & Felsenstein (1987)	4	2	−R	U	$\mathbf{S}_{JC}$	2	Independence; four-point metric
Felsenstein (1991)	3, 4	4	−R &	U	$\mathbf{S}_{JC}$	3	Heuristic
Drolet & Sankoff (1990)	4	$\geqslant 2$	−R	U	$\mathbf{S}_{JC}$	2	Independence; four-point metric
Sankoff (1990)	>2	$\geqslant 2$	−R	U	Symmetric	$\geqslant 2$	Independence
Cavender (1989)	4	4	R	A	$\mathbf{S}_{CAV}$	1	Vectorial analysis; numerical method
Nguyen & Speed (1992)	$\geqslant 2$	4	R	A	$\mathbf{S}_{NS}$	1	Vectorial analysis
Ferretti & Sankoff (1993)	4	$\geqslant 2$	−R	U	$\mathbf{S}_{JC}$	1, 2	Empirical
Fu & Li (1992*b*)	>2	4	A	A	$\mathbf{S}_{CAV}$	1	Analytical method
Steel *et al*. (1993*b*)	>2	$\geqslant 2$	R	A	$\lvert\mathbf{M}\rvert \neq 0, \pm 1$	>2	Four-point metric
Steel *et al*. (1993*c*)	>2	4	R	A	$\mathbf{S}_{3K}$	$\geqslant 2$	Random walk on abelian group; Fourier analysis
Evans & Speed (1993)	>2	4	R	A	$\mathbf{S}_{3K}$	$\geqslant 1$	Random walk on abelian group; Fourier analysis
Ferretti & Sankoff (1994)	3, 4	2, 4	R	C	$\mathbf{M}$ asymmetric	1, 2	Empirical
This paper	4	2	R	C	$\mathbf{S}_{2K}$	2	Empirical
This paper	4	4	R	A	10-param.	3	Empirical

N: number of terminals on tree.
k: size of state space.
$\mathbf{T}$: rooted (R), Unrooted (−R) or Arbitrary (A).
$\mathbf{S}$: semigroup.
π: root distribution: Uniform (U), Arbitrary (A) or Constrained (C).
d: degree of polynomial.
Independence: independence of events in disjoint paths in $\mathbf{T}$.
Four-point metric: property of lengths (like $-\det \mathbf{M}$) in "additive" tree structures.

where $\boldsymbol{\lambda}$ represents a vector of parameters. The problem becomes that of determining all $\boldsymbol{\lambda}$ for which the function Q is invariant over all $\mathbf{M}$ 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 $\mathbf{M}(i)$ of matrices, $1 \leqslant i \leqslant m$, for $\mathbf{T}$ at random, calculate explicitly the distribution $\mathbf{f}(i)$ for each set, and set up the system:

$$Q(\mathbf{f}(1), \boldsymbol{\lambda}) = 0$$
$$\cdots$$
$$Q(\mathbf{f}(m), \boldsymbol{\lambda}) = 0. \tag{2}$$

The set of invariants having of form Q is necessarily contained in the set of non-trivial solutions of this sytem.

Consider the important case of quadratic invariants. The function Q in (1) is of form

$$Q(\mathbf{f}, \boldsymbol{\lambda}) = \sum_{1 \leqslant i \leqslant j \leqslant w} \lambda_{ij} f_i f_j, \tag{3}$$

and then the equations in (2) can be written as a system of homogeneous linear equations in the unknown λ_{ij}

$$\mathbf{G} \cdot \boldsymbol{\lambda} = \mathbf{0},$$

where $\mathbf{G}$ is a $m \times v$ matrix, $v = w(w+1)/2$, with elements

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

λ_{ij} being the n-th component of $\boldsymbol{\lambda}$. The set of solutions to $\mathbf{G} \cdot \boldsymbol{\lambda} = \mathbf{0}$ defines the kernel of the matrix $\mathbf{G}$, denoted ker($\mathbf{G}$). 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 $\mathbf{G}$.

The key to the choice of m, π and $\mathbf{M}$ is to ensure that there are no extra solutions to $\mathbf{G} \cdot \boldsymbol{\lambda} = \mathbf{0}$ owing to accidental dependences among its columns. This can be ensured 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 since 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 $\mathbf{G}$ is a real matrix, the precise solution of $\mathbf{G} \cdot \boldsymbol{\lambda} = \mathbf{0}$, and of the larger problems of this sort we encounter with our method, becomes computationally cumbersome. It is easier to embed $\mathbf{G} \cdot \boldsymbol{\lambda} = \mathbf{0}$ in a multiple regression problem

$$\mathbf{G} \cdot \boldsymbol{\lambda} = \mathbf{0} + \boldsymbol{\epsilon},$$

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

4.2. REMOVING ALGEBRAIC DEPENDENCE

Suppose that we obtain the set $\{Q_1, \ldots, Q_p\}$ of quadratic invariants as linear basis for Ker($\mathbf{G}$), the solutions subspace of system (2) for a given model. By definition, the polynomials $Q_1, \ldots, Q_p$ are linearly independent, i.e. for any quadratic equation of form

$$\sum_{i=1}^{p} A_i Q_i = 0,$$

it must be that $A_i = 0$, for all i, $1 \leqslant i \leqslant p$, and for all probability distributions $\mathbf{f} = (f_1, \ldots, f_w)$. Our goal, however, is to find the smallest set of invariants which *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 $\{Q_1, \ldots, Q_p\}$ contain elements which are cubically related? In other words, are there coefficients A_{ij} such that

$$\sum_{i=1}^{w} \sum_{j=1}^{p} A_{ij} f_i Q_j = 0? \tag{4}$$

This question may also be investigated "empirically". We evaluate for r randomly generated probability distributions $\mathbf{f}(h)$, $1 \leqslant h \leqslant r$ (not necessarily spectra since they are not generated by any process over a fixed tree), the $p \times w$ quantities $f_i(h) Q_j(\mathbf{f}(h))$. The $p \times w$ terms derived from each probability distribution form one of the r rows of a matrix $\mathbf{H}$ (we may take $r = p \times w$). Then Ker($\mathbf{H}$) represents the set of dependences of form (4) among the polynomials $Q_1, \ldots, Q_p$.

We may go a step further and find in the same way quartic or higher degree algebraic relations between $Q_1, \ldots, Q_p$. However, as we shall see, there is a computational limit to this kind of investigation, owing to the rapid growth of the size of matrix $\mathbf{H}$ as the degree of the relation increases.

5. Three Applications

5.1. QUADRATIC INVARIANTS FOR S_{2K}

We now apply our method to find all quadratic invariants for tree $\mathbf{T}_1$ in Fig. 1 for the model $\mathbf{S_{2K}}$

$$\begin{pmatrix} 1-a-2b & a & b & b \\ a & 1-a-2b & b & b \\ b & b & 1-a-2b & a \\ b & b & a & 1-a-2b \end{pmatrix}, \tag{5}$$

with the uniform root distribution $\pi = (1/4, 1/4, 1/4, 1/4)$. There are thus ten unknown parameters in this inference problem, namely two for each edge. We write $\mathbf{f}(q)$ for the probability of observing the four-tuple $q = (q_1, q_2, q_3, q_4)$ at a given position. We have

$$\mathbf{f}(q) = \sum_{i=1}^{4}\sum_{j=1}^{4} \mathbf{M}_{EF}(i,j)\mathbf{M}_{EA}(i,q_1) \times \mathbf{M}_{EB}(i,q_2)\mathbf{M}_{FC}(j,q_3)\mathbf{M}_{FD}(j,q_4). \tag{6}$$

Given q, we define the four-tuple $\sigma(q) = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ recursively as follows: set $\sigma_1 = 1$ and, given $\sigma_1, \ldots, \sigma_j$, $1 \leqslant j \leqslant 3$,

$\sigma_{j+1} = \sigma_{j'}$ if there exists $j' \leqslant j$ such that $q_{j'} = q_{j+1}$

otherwise,

$\sigma_{j+1} = 2$ if the substitution (q_1, q_{j+1}) is a transition

$= 3$ if (q_1, q_{j+1}) is a transversion and $\sigma_{j'} \neq 3$, $1 \leqslant j' \leqslant j$

$= 4$ if (q_1, q_{j+1}) is a transversion and if $\exists j' \leqslant j$ such that $\sigma_{j'} = 3$.

For example, $\sigma((3,4,2,2)) = (1,2,3,3)$, $\sigma((2,4,3,1)) = (1,3,4,2)$, and so on. By the symmetries in the model, one has that $\mathbf{f}(q) = \mathbf{f}(q')$ if and only if $\sigma(q) = \sigma(q')$. Then, it will be simpler to treat only one representative from each of the 36 classes in the partition induced by this equality, which we denote as follow:

$$\begin{aligned} f_1 &= S(1,1,1,1) & f_{19} &= S(1,2,1,3) \\ f_2 &= S(1,1,1,2) & f_{20} &= S(1,3,1,2) \\ f_3 &= S(1,1,1,3) & f_{21} &= S(1,3,1,4) \\ f_4 &= S(1,1,2,1) & f_{22} &= S(1,2,3,1) \\ f_5 &= S(1,1,3,1) & f_{23} &= S(1,3,2,1) \\ f_6 &= S(1,1,2,2) & f_{24} &= S(1,3,4,1) \\ f_7 &= S(1,1,3,3) & f_{25} &= S(1,2,3,3) \\ f_8 &= S(1,2,1,1) & f_{26} &= S(1,3,2,2) \\ f_9 &= S(1,3,1,1) & f_{27} &= S(1,3,4,4) \\ f_{10} &= S(1,2,1,2) & f_{28} &= S(1,2,3,2) \\ f_{11} &= S(1,3,1,3) & f_{29} &= S(1,3,2,3) \\ f_{12} &= S(1,2,2,1) & f_{30} &= S(1,3,4,3) \\ f_{13} &= S(1,3,3,1) & f_{31} &= S(1,2,2,3) \\ f_{14} &= S(1,2,2,2) & f_{32} &= S(1,3,3,2) \\ f_{15} &= S(1,3,3,3) & f_{33} &= S(1,3,3,4) \\ f_{16} &= S(1,1,2,3) & f_{34} &= S(1,2,3,4) \\ f_{17} &= S(1,1,3,2) & f_{35} &= S(1,3,2,4) \\ f_{18} &= S(1,1,3,4) & f_{36} &= S(1,3,4,2), \end{aligned} \tag{7}$$

where

$$S(\sigma_1,\sigma_2,\sigma_3,\sigma_4) = \sum_{q:\,\sigma(q)=(\sigma_1,\sigma_2,\sigma_3,\sigma_4)} f(q).$$

Note that $\Sigma_{i=1}^{36} f_i = 1$.

With this notation, the problem is then to determinate all the invariants of the form (3) with $w = 36$. We first construct the matrix $\mathbf{G}$. To do this we randomly choose $m = v = 666$ points $\mathbf{p}(1), \ldots, \mathbf{p}(666)$ in ten-dimensional space to be the parameters in the various $\mathbf{M}$. We can then accurately calculate the $36f_j$ for each of the 666 spectra $\mathbf{f}(1), \ldots, \mathbf{f}(666)$. Using all of these values we can construct the 666×666 matrix $\mathbf{G}$. The set of quadratic invariants must be in ker($\mathbf{G}$).

5.1.2. *Results*

We find that rank($\mathbf{G}$) = 541 so that the canonical basis of ker($\mathbf{G}$) contains 125 elements. Among these invariants, 53 can be factored as f_iL_1 or f_iL_2, for some i, $1 \leqslant i \leqslant 36$, where L_1 and L_2 are linear combinations of the f_i. As a by-product, then, we have found the set of linear invariants for $\mathbf{S_{2K}}$. L_1 and L_2 are none other than the two invariants discovered by Lake (1987), namely

$$\begin{aligned} L_1 &= f_{21} + f_{29} - f_{11} - f_{35} \\ L_2 &= f_{24} + f_{32} - f_{13} - f_{36}. \end{aligned}$$

Another 18 of the 125 invariants can be discarded as being of the form $f_iL_j + Q$, where Q is itself a

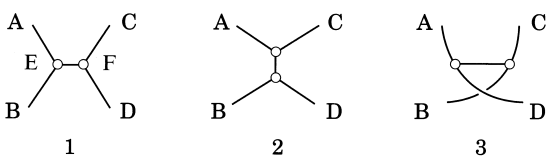

FIG. 1. The three unrooted binary trees on four species.

quadratic invariant. Thus there remain 54 quadratic invariants, which we denote $Q_1, \ldots, Q_{54}$, and which are given in Appendix A.

Applying the arguments about the number of invariants at the end of Section 2, with $h = 2$ and π given, we know that there must exist $256 - 10 = 246$ independent invariants for this model. But 220 of these invariants are simply the symmetric relations summarized in (7), and we have in addition the trivial invariant $\Sigma f = 1$. There must then be at least $56 - (246 - 220 - 1) = 31$ algebraic relations among the invariants $L_1, L_2, Q_1, \ldots, Q_{54}$ and we can apply the method of Section 4.2 to find all the cubic relations (4) among $Q_1, \ldots, Q_{54}$. In this case, $p = 54$ and $w = 36$, and we must then find the kernel of a matrix $\mathbf{H}$ of dimension 1944×1944.

We find 23 relations of form (4) (see equations in Appendix B) which allow us to eliminate the invariants $Q_{32}, Q_{33}, \ldots, Q_{54}$. There are no more cubic relations among the invariants $Q_1, \ldots, Q_{31}$, but there may be among $L_1, L_2, Q_1, \ldots, Q_{31}$. That is, there may be coefficients A_{ij}, B_{ij} and C_{ij}, not all zero, such that

$$\sum_{i \leqslant j \in [1, 36]} A_{ij} f_i f_j L_1 + \sum_{i \leqslant j \in [1, 36]} B_{ij} f_i f_j L_2 + \sum_{i \in [1, 36], j \in [1, 31]} C_{ij} f_i Q_j = 0. \quad (8)$$

Our method would require, in this case, a matrix $\mathbf{H}$ of dimension 2448×2448. It can be seen that the search for dependencies (8) as well as those of higher order, e.g. quartic, is limited by the computational size of the problem. For the moment then, we remain with $33 - 25 = 8$ invariants too many.

The fact that our $\boldsymbol{\lambda}$ are, strictly speaking, only estimated by the computer program on the basis of a (pseudo)-random sample might seem to relegate to the realm conjecture the invariant status of the forms in (A.1) in Appendix A. This is not the case, however. Substituting (5) and (6) into (A.1) proves explicitly that the forms are all invariant. That this latter calculation is impractical in many cases without the use of symbolic computing does not detract from the certainty of the result.

The quadratic forms $Q_1, \ldots, Q_{31}$ are not only invariants, but are true phylogenetic invariants in that they are non-zero for spectra $\mathbf{f}$ calculated according to the topologies of trees $\mathbf{T}_2$ or $\mathbf{T}_3$ in Fig. 1, and indeed vary with the edge matrix parameters associated with these trees. The linear forms L_1 and L_2, however, are each invariant for two of the three trees and variable with the edge matrix parameters of the third one.

5.2. QUADRATIC INVARIANTS FOR A MODEL WITH DEPENDENCE BETWEEN POSITIONS

The hypothesis of independent evolution of the values at different positions of a sequence is not really necessary in our search for invariants because our spectra $\mathbf{f}$ consist of expected frequencies at one position and estimations of $\mathbf{f}$ by averaging over positions is not affected by dependence, since summation and expectation commute. However, non-independence may severely interfere with the convergence of the observed spectra to the expected value. Moreover, non-independence is widespread, since positions that are close together in either primary or secondary structure may co-evolve. In this section, we investigate a simple model of evolution that incorporates a degree of non-independence between pairs (e.g. adjacent pairs) of positions. Steel *et al*. (1993*c*) have also proposed non-linear invariants for evolutionary models where the positions do not correspond to independent identically distributed (i.i.d.) random variables

We first divide the N aligned sequences of length n into $n/2$ pairs of positions, and define our semigroup $\mathbf{S}$ of $k^2 \times k^2$ matrices on the state space $\{1, \ldots, k\} \times \{1, \ldots, k\}$. We can then obtain spectra $\mathbf{f}$ on pairs of positions, and apply our method to obtain polynomial invariants. The assumption that the observed N-tuples are produced by i.i.d. random variables at each of n positions no longer holds, but is bequeathed to the $n/2$ pairs instead. Note that the positions constituting each pair are not necessarily adjacent in the sequence, which allows us to model, say, secondary structure constraints.

In the simplest case, the original state space is $\{1, 2\}$. Each pair of positions can then take on the following values: 11, 22, 12, 21, which we renumber as 1, 2, 3, 4, respectively. Because of the symmetry between states 1 and 2 and between states 3 and 4, it seems appropriate to use the $\mathbf{S_{2K}}$ model where 1↔2 and 3↔4 are "transitions" (rare in this context) and the other substitutions are "transversions". The root distribution is π where $\pi(1) = \pi(2)$ and $\pi(3) = \pi(4)$, which is consistent, for example, with random binary sequences.

We will first look for quadratic invariants for the tree $\mathbf{T}_1$ on the three species A, B and C in Fig. 2. Given the symmetries in $\mathbf{S}$ and in π, we classify the components of the spectrum $\mathbf{f}$ as follows:

$$f_1 = f(1, 1, 1) + f(2, 2, 2)$$

$$f_2 = f(1, 1, 3) + f(2, 2, 3) + f(1, 1, 4) + f(2, 2, 4)$$

$$f_3 = f(1, 1, 2) + f(2, 2, 1)$$

$$f_4 = f(1, 3, 1) + f(2, 3, 2) + f(1, 4, 1) + f(2, 4, 2)$$

$$f_5 = f(1, 3, 3) + f(1, 4, 4) + f(2, 3, 3) + f(2, 4, 4)$$

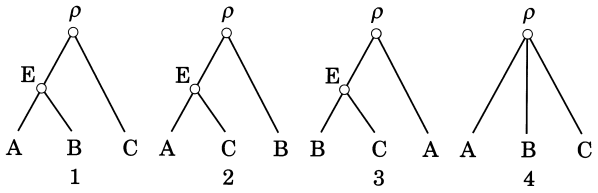

FIG. 2. The four rooted trees on three terminal vertices.

$$f_6 = f(1,3,4) + f(2,3,4) + f(1,4,3) + f(2,4,3)$$
$$f_7 = f(1,3,2) + f(1,4,2) + f(2,3,1) + f(2,4,1)$$
$$f_8 = f(1,2,1) + f(2,1,2)$$
$$f_9 = f(1,2,3) + f(1,2,4) + f(2,1,3) + f(2,1,4)$$
$$f_{10} = f(1,2,2) + f(2,1,1)$$
$$f_{11} = f(3,1,1) + f(3,2,2) + f(4,1,1) + f(4,2,2)$$
$$f_{12} = f(3,1,3) + f(3,2,3) + f(4,1,4) + f(4,2,4)$$
$$f_{13} = f(3,1,4) + f(3,2,4) + f(4,1,3) + f(4,2,3)$$
$$f_{14} = f(3,1,2) + f(3,2,1) + f(4,1,2) + f(4,2,1)$$
$$f_{15} = f(3,3,1) + f(3,3,2) + f(4,4,1) + f(4,4,2)$$
$$f_{16} = f(3,3,3) + f(4,4,4)$$
$$f_{17} = f(3,3,4) + f(4,4,3)$$
$$f_{18} = f(3,4,1) + f(3,4,2) + f(4,3,1) + f(4,3,2)$$
$$f_{19} = f(3,4,3) + f(4,3,4)$$
$$f_{20} = f(3,4,4) + f(4,3,3).$$

The symmetries ensure that each of the components of f_i are equal, for each matrix in $\mathbf{S}$. We have also that $\Sigma f_i = 1$. Applying our method, we find a matrix $\mathbf{G}$ of dimension 210×210 and of rank 198. After eliminating cubic dependences among the 12 invariants thus found, we are left with the following nine quadratic invariants:

$$\begin{aligned} Q_1 = & -f_2f_{11} + 2f_3f_{12} + 2f_1f_{13} - f_2f_{14} + f_5f_{15} \\ & + f_6f_{15} - 2f_7f_{16} - 2f_4f_{17} \end{aligned}$$
$$\begin{aligned} Q_2 = & -f_2f_{11} + 2f_1f_{12} + 2f_3f_{13} - f_2f_{14} + f_5f_{15} \\ & + f_6f_{15} - 2f_4f_{16} - 2f_7f_{17} \end{aligned}$$
$$\begin{aligned} Q_3 = & f_9f_{11} - 2f_{10}f_{12} - 2f_8f_{13} + f_9f_{14} - f_5f_{18} \\ & - f_6f_{18} + 2f_7f_{19} + 2f_4f_{20} \end{aligned}$$
$$\begin{aligned} Q_4 = & f_9f_{11} - 2f_8f_{12} - 2f_{10}f_{13} + f_9f_{14} - f_5f_{18} \\ & - f_6f_{18} + 2f_4f_{19} + 2f_7f_{20} \end{aligned}$$
$$\begin{aligned} Q_5 = & f_9f_{15} - 2f_{10}f_{16} - 2f_8f_{17} - f_2f_{18} \\ & + 2f_3f_{19} + 2f_1f_{20} \end{aligned}$$
$$\begin{aligned} Q_6 = & f_4f_5 - f_4f_6 - f_5f_7 + f_6f_7 - f_{11}f_{12} \\ & + f_{11}f_{13} + f_{12}f_{14} - f_{13}f_{14} \end{aligned}$$
$$\begin{aligned} Q_7 = & -f_2f_4 + 2f_3f_5 + 2f_1f_6 - f_2f_7 + f_{12}f_{15} \\ & + f_{13}f_{15} - 2f_{14}f_{16} - 2f_{11}f_{17} \end{aligned}$$
$$\begin{aligned} Q_8 = & 2f_6f_8 - f_4f_9 - f_7f_9 + 2f_5f_{10} + f_{12}f_{18} \\ & + f_{13}f_{18} - 2f_{14}f_{19} - 2f_{11}f_{20} \end{aligned}$$
$$\begin{aligned} Q_9 = & -f_2f_8 + f_1f_9 + f_3f_9 - f_2f_{10} - f_{16}f_{18} \\ & - f_{17}f_{18} + f_{15}f_{19} + f_{15}f_{20}. \end{aligned}$$

It is within the limits of computational feasibility to find all possible quartic relations involving the invariants for this model; i.e. relations of form $\Sigma_{1 \leqslant i \leqslant j \leqslant 20} \Sigma_{k \in [1,9]} A_{ijk} f_i f_j Q_k = 0$. To do this, we must construct a matrix $\mathbf{H}$ of dimension 1890×1890. We found that $\text{rank}(\mathbf{H}) = 1845$. Each of the $1890 - 1845 = 45$ relations turns out to be one of the 45 trivials relations of form $Q_iQ_j - Q_jQ_i = 0$. Thus there are no non-trivial quartic relations among the polynomials $Q_1, \ldots, Q_9$.

All these functions, whose invariant status has been confirmed with the help of symbolic computing, are not invariant when applied to spectra $\mathbf{f}$ generated by the trees $\mathbf{T}_2$ and $\mathbf{T}_3$ of Fig. 2.

Invoking the arguments of Section 3 above, there should be ten algebraically independent invariants for this model. There must thus be one invariant in $\mathbf{f}$ for which the functional form is not quadratic.

5.3. SKEWED BASE COMPOSITION

The evolution of nucleotide sequences is most frequently modeled so that $\mathbf{S}$ contains symmetric substitution matrices on the set of four nucleotide bases {A, G, C, T}, with uniform initial distribution π. Such models are inappropriate when the observed base compositions deviate strongly from uniform, because evolutionary inference methods, assuming uniformity and symmetry, are likely to group species on the basis of similar base compositions rather than true homology. Ferretti & Sankoff (1994) investigated asymmetric 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 three difficulties faced in this study were: first, that it was not trivial to define a constrained semigroup modeling skewed base compositions; second, that even when such semigroups were found, there were not necessarily any low-degree polynomial phylogenetic invariants; and third, that the empirical methodology quickly became computationally difficult as the number of parameters increased.

In the search for a tractable model for the evolution of skewed base composition, the following scenario was postulated. 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 chances of changing state in one direction is a constant multiplied by that of the other direction, this constant being universal across the group of organisms. Thus, given a constant $c > 1$, consider matrices of the form:

$$\mathbf{M} = \begin{pmatrix} 1-a & a \\ ca & 1-ca \end{pmatrix},$$

where a varies between 0 and $1/2c$. It is easy to prove that these matrices form a semigroup, which we denote $\mathbf{S}_c$. 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 $\mathbf{T}_1$ of Fig. 2. The parameter space now is five dimensional (the probability π_1 and the four matrix parameters a) and the matrix $\mathbf{G}$, as computed from $v = 2^3(2^3 + 1)/2$, is 36×36.

Calculation shows that $\text{rank}(\mathbf{G}) = 35$ for $c = 1, 2, 3, \ldots$ so that a basis of the subspace $\ker(\mathbf{G})$ contains only one element. For all values of c, this invariant could be expressed as

$$f_2f_3 - f_1f_4 - f_2f_5 + (c-1)f_4f_5 + f_1f_6 - (c-1)f_3f_6 + cf_4f_7 - cf_6f_7 - cf_3f_8 + cf_5f_8,$$

where

$$f_1 = f(1,1,1) \quad f_2 = f(1,1,2) \quad f_3 = f(1,2,1) \quad f_4 = f(1,2,2)$$
$$f_5 = f(2,1,1) \quad f_6 = f(2,1,2) \quad f_7 = f(2,2,1) \quad f_8 = f(2,2,2).$$

With the help of symbolic computing, it was proved explicitly that this form is invariant for all real c. From the discussion in Section 3, it can be seen that this is one of only two possible invariants for this model. However, the other cannot be a quadratic invariant.

The invariant found is a phylogenetic invariant, since not only is it identically zero for $\mathbf{T}_1$ in Fig. 2, but it is non-zero and varies with the edge matrix parameters when applied to spectra $\mathbf{f}$ for trees $\mathbf{T}_2$ and $\mathbf{T}_3$. Since $\mathbf{T}_4$ represents a degenerate case of $\mathbf{T}_1$ (in the matrix associated with ρE, $a = 0$, representing a zero edge length), the formula is also invariant for this case.

Ferretti & Sankoff (1994) also studied the more difficult case of $N = 4$ organisms. Consider the tree $\mathbf{T}_1$ in Fig. 3, which depicts the 26 rooted trees with $N = 4$ terminal vertices. The spectrum $\mathbf{f}$ has 16 terms for which we use the following abbreviated notation:

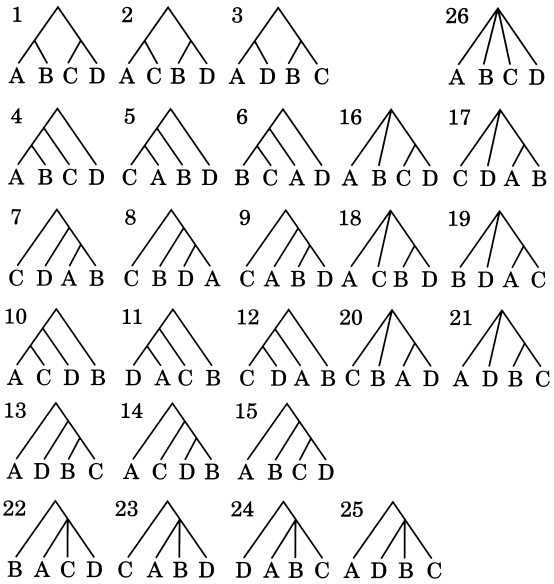

FIG. 3. The 26 rooted trees on $N = 4$ terminal vertices.

$$f_1 = f(1,1,1,1) \quad f_9 = f(2,1,1,1)$$
$$f_2 = f(1,1,1,2) \quad f_{10} = f(2,1,1,2)$$
$$f_3 = f(1,1,2,1) \quad f_{11} = f(2,1,2,1)$$
$$f_4 = f(1,1,2,2) \quad f_{12} = f(2,1,2,2)$$
$$f_5 = f(1,2,1,1) \quad f_{13} = f(2,2,1,1)$$
$$f_6 = f(1,2,1,2) \quad f_{14} = f(2,2,1,2)$$
$$f_7 = f(1,2,2,1) \quad f_{15} = f(2,2,2,1)$$
$$f_8 = f(1,2,2,2) \quad f_{16} = f(2,2,2,2).$$

Applying our method, a 136×136 matrix $\mathbf{G}$ of rank 125 was constructed for each of $c = 1, 2, 3, \ldots$. The 11 quadratic invariants making up $\ker(\mathbf{G})$ were found. Five cubic and two quartic relations were then found relating these invariants so that all except the following four could be eliminated as being algebraically dependent:

$$Q_1 = f_2f_5 - f_3f_5 - f_1f_6 + (c-1)f_3f_6 + cf_4f_6 + f_1f_7 - (c-1)f_2f_7 - cf_4f_7 - cf_2f_8 + cf_3f_8$$

$$Q_2 = f_2f_9 - f_3f_9 - f_1f_{10} + (c-1)f_3f_{10} + cf_4f_{10} + f_1f_{11} - (c-1)f_2f_{11} - cf_4f_{11} - cf_2f_{12} + cf_3f_{12}$$

$$Q_3 = f_2f_5 - f_1f_6 - f_2f_9 + (c-1)f_6f_9 + f_1f_{10} - (c-1)f_5f_{10} + cf_6f_{13} - cf_{10}f_{13} - cf_5f_{14} + cf_9f_{14}$$

$$Q_4 = f_3f_5 - f_1f_7 - f_3f_9 + (c-1)f_7f_9 + f_1f_{11} - (c-1)f_5f_{11} + cf_7f_{13} - cf_{11}f_{13} - cf_5f_{15} + cf_9f_{15}.$$

The arguments in Section 3 ensure that there must be two additional dependencies among these four functions, expressible as linear combinations where the coefficients are quotients of polynomials in the components of **f**, but these are not yet known.

The Q are quadratic phylogenetic invariants; each is invariant for certain of the trees in Fig. 3 and not for others. This information is summarized in Table 2, 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**, applying our method and eliminating the algebraic dependencies, we finally arrive at just two 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 Q_1 is invariant for just those trees in Table 2 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 $\mathbf{T}_{26}$.

TABLE 2
Invariants and topologies

	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.

Using small subunit ribosomal RNA sequences, Ferretti *et al*. (1994) used these invariants to confirm the phylogenetic relationships differentiating among four AT-rich fungi and among a set of AT-poor organisms including two plant mitochondria and two eubacteria closely related to the bacterial endosymbiont thought to be ancestral to present-day mitochondria.

Note that $\mathbf{S}_{\mathbf{CAV}}$, which is asymmetric, has only been shown to have linear invariants. A different approach to correcting for skewed base composition, using simulation, has recently been suggested by Steel *et al*. (1993*a*).

6. Computational Considerations

In the light of the three preceding examples, it is clear that our method is limited by the computational size necessary for its application, not only in the search for invariants, but also in the detection of algebraic dependence among them. This computational size depends directly on the dimension v^2 of **G** or **H**. v varies as a function of the number w of components in the spectrum **f** and on the nature of the parametric form Q. For example, to find linear invariants, $v = w$; for quadratic invariants, $v = w(w+1)/2$; and generally, for polynomial invariants of degree d, $v = (w+d-1)!/d!(w-1)!$ At worst, $w = k^N$, but we have seen how the existence of symmetries in the evolutionary model allows us to reduce this number. The existence of symmetries in a model is related to the number of parameters necessary to define its semigroup **S** and its root distribution π: the less the number of parameters, the more symmetries there are and the greater the possibility for working with a smaller **G**.

Ideally, we would like to have invariants for the most general model possible, which would then be applicable to all possible situations. In such a model, we would have $(k-1)(k+1)$ independent parameters ($k-1$ for π and $k(k-1)$ for **S**) and, from Cavender (1991) we already know that there can be no linear invariant for this. As for higher order invariants, our method rapidly becomes impracticable as k and N increase, because, in this case, $w = k^N$.

One way of getting around this difficulty is to develop methods that allow us to derive new invariants with the help of known invariants for less general models. This is the great interest in results such as Proposition 4 of Cavender (1991) which stipulates that a semigroup containing $\mathbf{S}_{\mathbf{CAV}}$ generates the same invariants as $\mathbf{S}_{\mathbf{CAV}}$ (or none). Similarly, methods like that of Fu & Li (1992*b*) that find all invariants for N

species based on the invariants for $N-1$. Unfortunately for our purposes, these two techniques are only pertinent to linear invariants.

In the next section, we combine our method with an ingenious technique from Felseinstein (1991) to obtain, based on invariants for a two-state, two-parameter per edge model, invariants for $k=4$ with ten independent parameters per edge.

7. Cubic Invariants for a Ten-parameter Model

Consider the tree $\mathbf{T}_1$ with $N=4$ terminals in Fig. 3 and the model $\mathbb{M}_2$ with $k=2$ states $\{\alpha, \beta\}$, characterized by the semigroup of matrices of form

$$\begin{array}{c} \\ \alpha \\ \beta \end{array} \begin{array}{c} \begin{array}{cc} \alpha & \beta \end{array} \\ \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \end{array} \tag{9}$$

and an arbitrary root distribution π. Let

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

be the 16 components of the spectrum $\mathbf{f}=(f_1,\ldots,f_{16})$. Applying our method, we find that there exists no linear and no quadratic invariants for this model. We do find, however, a set of 16 cubic phylogenetic invariants, which we reduce to six after determining the dependences of degree four and five in f_i:

$$\begin{aligned} Q_1 &= -f_3f_6f_9 + f_2f_7f_9 + f_3f_5f_{10} - f_1f_7f_{10} \\ &\quad - f_2f_5f_{11} + f_1f_6f_{11} \\ Q_2 &= -f_4f_6f_9 + f_2f_8f_9 + f_4f_5f_{10} - f_1f_8f_{10} \\ &\quad - f_2f_5f_{12} + f_1f_6f_{12} \\ Q_3 &= -f_3f_6f_{13} + f_2f_7f_{13} + f_3f_5f_{14} - f_1f_7f_{14} \\ &\quad - f_2f_5f_{15} + f_1f_6f_{15} \\ Q_4 &= -f_4f_6f_{13} + f_2f_8f_{13} + f_4f_5f_{14} - f_1f_8f_{14} \\ &\quad - f_2f_5f_{16} + f_1f_6f_{16} \\ Q_5 &= -f_4f_{11}f_{13} + f_3f_{12}f_{13} + f_4f_9f_{15} - f_1f_{12}f_{15} \\ &\quad - f_3f_9f_{16} + f_1f_{11}f_{16} \\ Q_6 &= -f_8f_{11}f_{14} + f_7f_{12}f_{14} + f_8f_{10}f_{15} - f_6f_{12}f_{15} \\ &\quad - f_7f_{10}f_{16} + f_6f_{11}f_{16} \end{aligned} \tag{11}$$

These six polynomials are invariant for the spectra $\mathbf{f}$ calculated according to the trees $\mathbf{T}_1$, $\mathbf{T}_4$, $\mathbf{T}_7$, $\mathbf{T}_{10}$, $\mathbf{T}_{13}$, $\mathbf{T}_{16}$, $\mathbf{T}_{17}$, $\mathbf{T}_{22}$, $\mathbf{T}_{23}$, $\mathbf{T}_{24}$, $\mathbf{T}_{25}$ and $\mathbf{T}_{26}$ of Fig. 3, but they are not for the other trees in the same figure. Once again, we can show that there exists at least four other independent algebraic relations among $Q_1,\ldots,Q_6$.

Now consider the state space $\{1, 2, 3, 4\}$. Suppose that the event "to be in state α" is defined as the event "to be in state 1 or 2" and the event "to be in state β" is defined as the event "to be in state 3 or 4". With these definitions, we can extend our model $\mathbb{M}_2$ to a stochastic model $\mathbb{M}_4$ with four states, in which, for example, each of the four substitutions (1, 3), (1, 4), (2, 3) and (2, 4) is represented by a substitution (α, β) in model $\mathbb{M}_2$. Let

$$\begin{pmatrix} c & d & e & f \\ g & h & i & j \\ l & n & o & p \\ q & r & s & t \end{pmatrix} \tag{12}$$

be a transition matrix of $\mathbb{M}_4$. By (9), $P[(\alpha,\beta)]=a$ and $P[(\beta,\alpha)]=b$, so the following constraints should hold:

$$\begin{aligned} e+f &= i+j = a \\ 1+n &= q+r = b, \end{aligned} \tag{13}$$

and hence the matrix (12) contains ten independent parameters. It is easily verified that the set of matrices of form (12) with constraints (13) do constitute a semigroup. With the correspondence we defined between $\mathbb{M}_2$ and $\mathbb{M}_4$, each of the frequencies f_i in (10) can be written as a linear combination of 16 different components of the spectrum $\mathbf{f}$ of model $\mathbb{M}_4$. For example,

$$\begin{aligned} f_3 &= f(1,1,3,1) + f(1,1,3,2) + f(1,2,3,1) \\ &\quad + f(1,2,3,2) + f(2,1,3,1) + f(2,1,3,2) \\ &\quad + f(2,2,3,1) + f(2,2,3,2) + f(1,1,4,1) \\ &\quad + f(1,1,4,2) + f(1,2,4,1) + f(1,2,4,2) \\ &\quad + f(2,1,4,1) + f(2,1,4,2) \\ &\quad + f(2,2,4,1) + f(2,2,4,2). \end{aligned}$$

Substituting this equation and the 15 others of the same type in (11), we obtain six cubic invariants for the model $\mathbb{M}_4$. Given the length of these expressions, we present only the first one.

$$\begin{aligned} Q_1' = -(&f(1,1,3,1) + f(1,1,3,2) + f(1,1,4,1) \\ &+ f(1,1,4,2) + f(1,2,3,1) + f(1,2,3,2) \\ &+ f(1,2,4,1) + f(1,2,4,2) + f(2,1,3,1) \\ &+ f(2,1,3,2) + f(2,1,4,1) + f(2,1,4,2) \end{aligned}$$

$$
\begin{aligned}
&+f(2,2,3,1)+f(2,2,3,2)+f(2,2,4,1)\\
&+f(2,2,4,2))(f(1,3,1,3)+f(1,3,1,4)\\
&+f(1,3,2,3)+f(1,3,2,4)+f(1,4,1,3)\\
&+f(1,4,1,4)+f(1,4,2,3)+f(1,4,2,4)\\
&+f(2,3,1,3)+f(2,3,1,4)+f(2,3,2,3)\\
&+f(2,3,2,4)+f(2,4,1,3)+f(2,4,1,4)\\
&+f(2,4,2,3)+f(2,4,2,4))(f(3,1,1,1)\\
&+f(3,1,1,2)+f(3,1,2,1)+f(3,1,2,2)\\
&+f(3,2,1,1)+f(3,2,1,2)+f(3,2,2,1)\\
&+f(3,2,2,2)+f(4,1,1,1)+f(4,1,1,2)\\
&+f(4,1,2,1)+f(4,1,2,2)+f(4,2,1,1)\\
&+f(4,2,1,2)+f(4,2,2,1)+f(4,2,2,2))\\
&+(f(1,1,1,3)+f(1,1,1,4)+f(1,1,2,3)\\
&+f(1,1,2,4)+f(1,2,1,3)+f(1,2,1,4)\\
&+f(1,2,2,3)+f(1,2,2,4)+f(2,1,1,3)\\
&+f(2,1,1,4)+f(2,1,2,3)+f(2,1,2,4)\\
&+f(2,2,1,3)+f(2,2,1,4)+f(2,2,2,3)\\
&+f(2,2,2,4))(f(1,3,3,1)+f(1,3,3,2)\\
&+f(1,3,4,1)+f(1,3,4,2)+f(1,4,3,1)\\
&+f(1,4,3,2)+f(1,4,4,1)+f(1,4,4,2)\\
&+f(2,3,3,1)+f(2,3,3,2)+f(2,3,4,1)\\
&+f(2,3,4,2)+f(2,4,3,1)+f(2,4,3,2)\\
&+f(2,4,4,1)+f(2,4,4,2))(f(3,1,1,1)\\
&+f(3,1,1,2)+f(3,1,2,1)+f(3,1,2,2)\\
&+f(3,2,1,1)+f(3,2,1,2)+f(3,2,2,1)\\
&+f(3,2,2,2)+f(4,1,1,1)+f(4,1,1,2)\\
&+f(4,1,2,1)+f(4,1,2,2)+f(4,2,1,1)\\
&+f(4,2,1,2)+f(4,2,2,1)+f(4,2,2,2))\\
&+(f(1,1,3,1)+f(1,1,3,2)+f(1,1,4,1)\\
&+f(1,1,4,2)+f(1,2,3,1)+f(1,2,3,2)\\
&+f(1,2,4,1)+f(1,2,4,2)+f(2,1,3,1)\\
&+f(2,1,3,2)+f(2,1,4,1)+f(2,1,4,2)\\
&+f(2,2,3,1)+f(2,2,3,2)+f(2,2,4,1)\\
&+f(2,2,4,2))(f(1,3,1,1)+f(1,3,1,2)\\
&+f(1,3,2,1)+f(1,3,2,2)+f(1,4,1,1)\\
&+f(1,4,1,2)+f(1,4,2,1)+f(1,4,2,2)\\
&+f(2,3,1,1)+f(2,3,1,2)+f(2,3,2,1)\\
&+f(2,3,2,2)+f(2,4,1,1)+f(2,4,1,2)\\
&+f(2,4,2,1)+f(2,4,2,2))(f(3,1,1,3)\\
&+f(3,1,1,4)+f(3,1,2,3)+f(3,1,2,4)\\
&+f(3,2,1,3)+f(3,2,1,4)+f(3,2,2,3)\\
&+f(3,2,2,4)+f(4,1,1,3)+f(4,1,1,4)\\
&+f(4,1,2,3)+f(4,1,2,4)+f(4,2,1,3)\\
&+f(4,2,1,4)+f(4,2,2,3)+f(4,2,2,4))\\
&-(f(1,1,1,1)+f(1,1,1,2)+f(1,1,2,1)\\
&+f(1,1,2,2)+f(1,2,1,1)+f(1,2,1,2)\\
&+f(1,2,2,1)+f(1,2,2,2)+f(2,1,1,1)\\
&+f(2,1,1,2)+f(2,1,2,1)+f(2,1,2,2)\\
&+f(2,2,1,1)+f(2,2,1,2)+f(2,2,2,1)\\
&+f(2,2,2,2))(f(1,3,3,1)+f(1,3,3,2)\\
&+f(1,3,4,1)+f(1,3,4,2)+f(1,4,3,1)\\
&+f(1,4,3,2)+f(1,4,4,1)+f(1,4,4,2)\\
&+f(2,3,3,1)+f(2,3,3,2)+f(2,3,4,1)\\
&+f(2,3,4,2)+f(2,4,3,1)+f(2,4,3,2)\\
&+f(2,4,4,1)+f(2,4,4,2))(f(3,1,1,3)\\
&+f(3,1,1,4)+f(3,1,2,3)+f(3,1,2,4)\\
&+f(3,2,1,3)+f(3,2,1,4)+f(3,2,2,3)\\
&+f(3,2,2,4)+f(4,1,1,3)+f(4,1,1,4)\\
&+f(4,1,2,3)+f(4,1,2,4)+f(4,2,1,3)\\
&+f(4,2,1,4)+f(4,2,2,3)+f(4,2,2,4))\\
&-(f(1,1,1,1)+f(1,1,1,2)+f(1,1,2,1)\\
&+f(1,1,2,2)+f(1,2,1,1)+f(1,2,1,2)\\
&+f(1,2,2,1)+f(1,2,2,2)+f(2,1,1,1)\\
&+f(2,1,1,2)+(2,1,2,1)+f(2,1,2,2)\\
&+f(2,2,1,1)+f(2,2,1,2)+f(2,2,2,1)\\
&+f(2,2,2,2))(f(1,3,3,1)+f(1,3,3,2)\\
&+f(1,3,4,1)+f(1,3,4,2)+f(1,4,3,1)\\
&+f(1,4,3,2)+f(1,4,4,1)+f(1,4,4,2)\\
&+f(2,3,3,1)+f(2,3,3,2)+f(2,3,4,1)\\
&+f(2,3,4,2)+f(2,4,3,1)+f(2,4,3,2)\\
&+f(2,4,4,1)+f(2,4,4,2))(f(3,1,1,3)\\
&+f(3,1,1,4)+f(3,1,2,3)+(3,1,2,4)\\
&+f(3,2,1,3)+f(3,2,1,4)+f(3,2,2,3)
\end{aligned}
$$

$$+f(3,2,2,4)+f(4,1,1,3)+f(4,1,1,4)$$
$$+f(4,1,2,3)+f(4,1,2,4)+f(4,2,1,3)$$
$$+f(4,2,1,4)+f(4,2,2,3)+f(4,2,2,4))$$
$$-(f(1,1,1,3)+f(1,1,1,4)+f(1,1,2,3)$$
$$+f(1,1,2,4)+f(1,2,1,3)+f(1,2,1,4)$$
$$+f(1,2,2,3)+f(1,2,2,4)+f(2,1,1,3)$$
$$+f(2,1,1,4)+f(2,1,2,3)+f(2,1,2,4)$$
$$+f(2,2,1,3)+f(2,2,1,4)+f(2,2,2,3)$$
$$+f(2,2,2,4))(f(1,3,1,1)+f(1,3,1,2)$$
$$+f(1,3,2,1)+f(1,3,2,2)+f(1,4,1,1)$$
$$+f(1,4,1,2)+f(1,4,2,1)+f(1,4,2,2)$$
$$+f(2,3,1,1)+f(2,3,1,2)+f(2,3,2,1)$$
$$+f(2,3,2,2)+f(2,4,1,1)+f(2,4,1,2)$$
$$+f(2,4,2,1)+f(2,4,2,2))(f(3,1,3,1)$$
$$+f(3,1,3,2)+f(3,1,4,1)+f(3,1,4,2)$$
$$+f(3,2,3,1)+f(3,2,3,2)+f(3,2,4,1)$$
$$+f(3,2,4,2)+f(4,1,3,1)+f(4,1,3,2)$$
$$+f(4,1,4,1)+f(4,1,4,2)+f(4,2,3,1)$$
$$+f(4,2,3,2)+f(4,2,4,1)+f(4,2,4,2))$$
$$+(f(1,1,1,1)+f(1,1,1,2)+f(1,1,2,1)$$
$$+f(1,1,2,2)+f(1,2,1,1)+f(1,2,1,2)$$
$$+f(1,2,2,1)+f(1,2,2,2)+f(2,1,1,1)$$
$$+f(2,1,1,2)+f(2,1,2,1)+f(2,1,2,2)$$
$$+f(2,2,1,1)+f(2,2,1,2)+f(2,2,2,1)$$
$$+f(2,2,2,2))(f(1,3,1,3)+f(1,3,1,4)$$
$$+f(1,3,2,3)+f(1,3,2,4)+f(1,4,1,3)$$
$$+f(1,4,1,4)+f(1,4,2,3)+f(1,4,2,4)$$
$$+f(2,3,1,3)+f(2,3,1,4)+f(2,3,2,3)$$
$$+f(2,3,2,4)+f(2,4,1,3)+f(2,4,1,4)$$
$$+f(2,4,2,3)+f(2,4,2,4))(f(3,1,3,1)$$
$$+f(3,1,3,2)+f(3,1,4,1)+f(3,1,4,2)$$
$$+f(3,2,3,1)+f(3,2,3,2)+f(3,2,4,1)$$
$$+f(3,2,4,2)+f(4,1,3,1)+f(4,1,3,2)$$
$$+f(4,1,4,1)+f(4,1,4,2)+f(4,2,3,1)$$
$$+f(4,2,3,2)+f(4,2,4,1)+f(4,2,4,2)).$$

By construction, the six polynomials thus derived are invariants for the same trees as Q_1, Q_2, Q_3, Q_4, Q_5 and Q_6.

8. Conclusion

The empirical method is fundamentally a discovery tool for identifying and relating invariants. It is conceptually simple and is capable of finding all invariants of a given functional form. In earlier work (Ferretti & Sankoff, 1993), however, it appeared to have two major difficulties that threatened to limit its utility for all but the smallest problems. One of these is the computational cost of dealing with the large matrices generated during the analysis. The second is the problem of reducing the large number of forms found to a minimal set of algebraically independent functions, or even knowing the size of this set. Still another problem facing any attempts to use invariants other than linear is the i.i.d. assumption that permit us to analyze one position as representing all positions.

In this paper we have attempted to make some headway against these problems. For example, we have demonstrated that there is a wide variety of models that can be anlyzed using moderate computational resources. We were required to manipulate matrices of size 2000×2000. It does not seem impossible that the matrices of size $32\,000 \times 32\,000$ necessary for an attack on the 12-parameter model with our method should be feasible in the not-too-distant future.

The problem of determining the maximum number of algebraically independent invariants seems to hinge on whether it suffices to consider only polynomial invariants, which seems likely given the results to date. In this case, rank(D) should be constant over parameter space, not only in a neighbourhood around each point (which is all we can say from general implicit function considerations), and it would suffice to calculate it for one point only. This is quite feasible at the present time.

In addition, until we know whether it suffices to consider only polynomial invariants that are globally valid in parameter space, we cannot prove our method identifies all invariants. For example, if an invariant had two different polynomial expressions in two sampled regions of parameter space, then our method would in general not identify it, since (at least) two sampled points would not generally be on the polynomial corresponding to the other region.

As for the i.i.d. constraint, we have taken the first steps towards allowing dependence among sequence positions.

Future direction for phylogenetic invariant research will involve not only the search for a complete set of invariants for $\mathbf{S_{PD}}$ but also feasible approaches to incorporating the invariant methodology into practical algorithms for constructing trees on large members of terminal vertices.

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada, the Fonds pour la formation de Chercheurs et l'Aide à la Recherche of the government of Québec, and the Canadian Genome Analysis & Technology Program. D.S. is a fellow of the Canadian Institute for Advanced Research.

REFERENCES

Cavender, J. A. (1989). Mechanized derivation of linear invariants. *Molec. 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. & Felsenstein, J. (1987). Invariants of phylogenies: Simple case with discrete states. *J. Classif.* **4,** 57–71.

Drolet, S. & Sankoff, D. (1990). Quadratic tree invariants for multivalued characters. *J. theor. Biol.* **144,** 117–129.

Evans, S. N. & Speed, T. P. (1994). Invariants of some probability models used in phylogenetic inference. *Amm. Statist.* **21,** 351–377.

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

Ferretti, V. & Sankoff, D. (1993). The empirical discovery of phylogenetic invariants. *Adv. appl. Probab.* **25,** 290–302.

Ferretti, V. Lang, B. F. & Sankoff, D. (1994). Skewed base compositions, asymmetric transition matrices and phylogenetic invariants. *J. Comput. Biol.* **1**.

Fu, Y. X. & Li, W. H. (1991). Necessary and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree. *Math. Biosci.* **105,** 229–238.

Fu, Y. X. & Li, W. H. (1992*a*). Necessary and sufficient conditions for the existence of linear invariants. *Math. Biosci.* **108,** 203–218.

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

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

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

Kimura, M. (1981). Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natn. Acad. Sci. U.S.A.* **78,** 454–458.

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

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

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

Sankoff, D. (1990). Designer invariants for large phylogenies. *Molec. Biol. Evolut.* **7,** 255–269.

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

Steel, M. A., Hendy, M. D. & Penny, D. (1993*b*). Invertible models of sequence evolution. Department of Mathematics Massey University, New Zealand.

Steel, M. A., Szekely, L. A., Erdos, P. L. & Waddell, P. (1993*c*). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *N.Z. Jl Bot.* **31,** 289–296.

Székely, L. A., Steel, M. A. & Erdos, P. L. (1993). Fourier calculus on evolutionary trees. *Adv. appl. Math.* **14,** 200–216.

APPENDIX A

Quadratic Invariants for $\mathbf{S_{2K}}$ and the Tree $\mathbf{T_1}$ of Fig. 1

$$Q_1 = (f_1 - f_6)(f_{12} - f_{10}) + (f_2 - f_4)(f_8 - f_{14})$$

$$Q_2 = (f_1 - f_6)(f_{23} - f_{20}) + (f_2 - f_4)(f_9 - f_{26})$$

$$Q_3 = (f_1 - f_6)(f_{33} - f_{30}) + (f_2 - f_4)(f_{27} - f_{15})$$

$$Q_4 = (f_2 - f_4)(f_{19} - f_{31}) + (f_3 - f_{16})(f_{12} - f_{10})$$

$$Q_5 = (f_2 - f_4)(f_{22} - f_{28}) + (f_5 - f_{17})(f_{12} - f_{10})$$

$$Q_6 = f_1 f_{34} + f_6 f_{34} + f_7 f_{10} + f_7 f_{12} - f_2 f_{25} - f_4 f_{25} - f_8 f_{18} - f_{14} f_{18}$$

$$Q_7 = 2f_1 f_{24} + 2f_1 f_{32} + 2f_6 f_{13} + f_7 f_{21} + f_7 f_{29} - f_2 f_1 f_{13} - f_3 f_{27} - f_5 f_{26} - f_9 f_{17} - f_{15} f_{16}$$

$$Q_8 = 2f_1 f_{21} + 2f_1 f_{29} + 2f_6 f_{11} + f_7 f_{24} + f_7 f_{32} - 2f_1 f_{11} - f_3 f_{26} - f_5 f_{27} - f_9 f_{16} - f_{15} f_{17}$$

$$Q_9 = 2(f_2 - f_4)(f_{11} - f_{29}) + (f_3 - f_{16})(f_{23} - f_{20})$$

$$Q_{10} = 2f_1 f_{28} + 2f_6 f_{22} + f_7 f_{19} + f_7 f_{31} - 2f_5 f_{14} - 2f_8 f_{17} - f_3 f_{25} - f_{16} f_{25}$$

$$Q_{11} = 2f_1 f_{32} + 2f_6 f_{24} + f_7 f_{21} + f_7 f_{29} - f_3 f_{15} - f_5 f_{26} - f_9 f_{17} - f_{16} f_{27}$$

$$Q_{12} = 2(f_2 - f_4)(f_{13} - f_{32}) + (f_5 - f_{17})(f_{23} - f_{20})$$

$$Q_{13} = (f_5 - f_{17})(f_{24} - f_{13}) + (f_3 - f_{16})(f_{11} - f_{21})$$

$$Q_{14} = 2f_1 f_{31} + 2f_6 f_{19} + f_7 f_{22} + f_7 f_{28} - 2f_3 f_{14} - 2f_8 f_{16} - f_5 f_{25} - f_{17} f_{25}$$

$$Q_{15} = 2(f_2 - f_4)(f_{21} - f_{11}) + (f_5 - f_{17})(f_{33} - f_{30})$$

$$Q_{16} = 2f_1 f_{12} + 2f_6 f_{10} + f_7 f_{34} - f_2 f_{14} - 2f_4 f_8 - f_{18} f_{25}$$

$$Q_{17} = 2f_1 f_{33} + 2f_6 f_{30} + f_7 f_{20} + f_7 f_{23} - 2f_2 f_{15} - 2f_4 f_{27} - f_9 f_{18} - f_{18} f_{26}$$

$$Q_{18} = 2f_1 f_{23} + 2f_6 f_{20} + f_7 f_{30} + f_7 f_{33} - 2f_2 f_{26} - 2f_4 f_9 - f_{15} f_{18} - f_{18} f_{27}$$

$$Q_{19} = 2f_3 f_{12} + 2f_{10} f_{16} + f_5 f_{34} + f_{17} f_{34} - 2f_2 f_{31} - 2f_4 f_{19} - f_{18} f_{22} - f_{18} f_{28}$$

$$
\begin{aligned}
Q_{20} &= f_3f_{33} + f_5f_{23} + f_{16}f_{30} + f_{17}f_{20} - 2f_2f_{32} \\
&\quad - 2f_4f_{24} - f_{18}f_{21} - f_{18}f_{29} \\
Q_{21} &= 2f_5f_{12} + 2f_{10}f_{17} + f_3f_{34} + f_{16}f_{34} - 2f_2f_{28} \\
&\quad - 2f_4f_{22} - f_{18}f_{19} - f_{18}f_{31} \\
Q_{22} &= f_3f_{23} + f_{16}f_{20} + f_{17}f_{30} - 2f_2f_{29} \\
&\quad - 2f_4f_{21} - f_{18}f_{24} - f_{18}f_{32} \\
Q_{23} &= 2f_8f_{13} + f_9f_{28} + f_{15}f_{31} + f_{19}f_{27} + f_{22}f_{26} \\
&\quad - 2f_8f_{24} - 2f_8f_{32} - 2f_{13}f_{14} - f_{21}f_{25} - f_{25}f_{29} \\
Q_{24} &= 2f_8f_{11} + f_9f_{31} + f_{15}f_{28} + f_{19}f_{26} + f_{22}f_{27} \\
&\quad - 2f_8f_{21} - 2f_8f_{29} - 2f_{11}f_{14} - f_{24}f_{25} - f_{25}f_{32} \\
Q_{25} &= 2(f_8 - f_{14})(f_{13} - f_{32}) + (f_9 - f_{26})(f_{28} - f_{22}) \\
Q_{26} &= 2(f_8 - f_{14})(f_{11} - f_{29}) + (f_9 - f_{26})(f_{31} - f_{19}) \\
Q_{27} &= 2f_8f_{33} + 2f_{14}f_{30} + f_{20}f_{25} + f_{23}f_{25} \\
&\quad - 2f_{10}f_{14} - 2f_{12}f_{27} - f_9f_{34} - f_{26}f_{34} \\
Q_{28} &= 2(f_8 - f_{14})(f_{11} - f_{21}) \\
&\quad + (f_{15} - f_{27})(f_{28} - f_{22}) \\
Q_{29} &= 2f_8f_{23} + 2f_{14}f_{20} + f_{25}f_{30} + f_{25}f_{33} - 2f_9f_{12} \\
&\quad - 2f_{10}f_{26} - f_{15}f_{34} - f_{27}f_{34} \\
Q_{30} &= 2f_{10}f_{32} + 2f_{12}f_{24} + f_{21}f_{34} + f_{29}f_{34} - f_{19}f_{33} \\
&\quad - f_{20}f_{28} - f_{22}f_{23} - f_{30}f_{31} \\
Q_{31} &= f_{19}f_{23} + f_{20}f_{31} + f_{22}f_{33} + f_{28}f_{30} - 2f_{10}f_{29} \\
&\quad - 2f_{12}f_{21} - f_{24}f_{34} - f_{32}f_{34} \\
Q_{32} &= (f_1 - f_6)(f_{19} - f_{31}) + (f_3 - f_{16})(f_{14} - f_8) \\
Q_{33} &= (f_1 - f_6)(f_{22} - f_{28}) + (f_5 - f_{17})(f_{14} - f_8) \\
Q_{34} &= (f_8 - f_{14})(f_{23} - f_{20}) + (f_9 - f_{26})(f_{10} - f_{12}) \\
Q_{35} &= (f_8 - f_{14})(f_{33} - f_{30}) + (f_{27} - f_{15})(f_{10} - f_{12}) \\
Q_{36} &= 2(f_1 - f_6)(f_{11} - f_{29}) + (f_3 - f_{16})(f_{26} - f_9) \\
Q_{37} &= 2(f_2 - f_4)(f_{24} - f_{13}) + (f_3 - f_{16})(f_{33} - f_{30}) \\
Q_{38} &= 2(f_1 - f_6)(f_{13} - f_{32}) + (f_5 - f_{17})(f_{26} - f_9) \\
Q_{39} &= 2f_1f_{29} + 2f_6f_{21} + f_7f_{24} + f_7f_{32} - f_3f_{26} \\
&\quad - f_5f_{15} - f_9f_{16} - f_{17}f_{27} \\
Q_{40} &= (f_3 - f_{16})(f_{13} - f_{32}) + (f_5 - f_{17})(f_{29} - f_{11}) \\
Q_{41} &= (f_3 - f_{16})(f_{22} - f_{28}) + (f_5 - f_{17})(f_{31} - f_{19}) \\
Q_{42} &= 2(f_{10} - f_{12})(f_{13} - f_{32}) + (f_{20} - f_{23})(f_{28} - f_{22}) \\
Q_{43} &= 2(f_{10} - f_{12})(f_{11} - f_{29}) + (f_{19} - f_{31})(f_{23} - f_{20}) \\
Q_{44} &= (f_{11} - f_{21})(f_{19} - f_{31}) + (f_{13} - f_{24})(f_{28} - f_{22}) \\
Q_{45} &= (f_{11} - f_{29})(f_{11} - f_{21}) + (f_{13} - f_{32})(f_{24} - f_{13}) \\
Q_{46} &= (f_9 - f_{26})(f_{33} - f_{30}) + (f_{15} - f_{27})(f_{23} - f_{20}) \\
Q_{47} &= (f_9 - f_{26})(f_{13} - f_{24}) + (f_{11} - f_{29})(f_{27} - f_{15}) \\
Q_{48} &= 2(f_8 - f_{14})(f_{13} - f_{24}) + (f_{15} - f_{27})(f_{31} - f_{19}) \\
Q_{49} &= (f_9 - f_{26})(f_{11} - f_{21}) + (f_{13} - f_{32})(f_{27} - f_{15}) \\
Q_{50} &= 2(f_{10} - f_{12})(f_{21} - f_{11}) + (f_{22} - f_{28})(f_{33} - f_{30}) \\
Q_{51} &= (f_{11} - f_{29})(f_{33} - f_{30}) + (f_{13} - f_{24})(f_{23} - f_{20}) \\
Q_{52} &= (f_{11} - f_{29})(f_{22} - f_{28}) + (f_{13} - f_{32})(f_{31} - f_{19}) \\
Q_{53} &= 2f_{10}f_{24} + 2f_{10}f_{32} + 2f_{12}f_{13} + f_{21}f_{34} + f_{29}f_{34} \\
&\quad - 2f_{10}f_{13} - f_{19}f_{30} - f_{20}f_{28} - f_{22}f_{23} - f_{31}f_{33} \\
Q_{54} &= (f_{11} - f_{21})(f_{23} - f_{20}) + (f_{13} - f_{32})(f_{33} - f_{30})
\end{aligned}
\tag{A.1}
$$

APPENDIX B

Cubic Relations of Form (4) Between $Q_1, \ldots, Q_{54}$

$$
\begin{aligned}
&f_3Q_1 - f_{16}Q_1 - f_1Q_4 + f_6Q_4 + f_2Q_{32} - f_4Q_{32} = 0 \\
&f_5Q_1 - f_{17}Q_1 - f_1Q_5 + f_6Q_5 + f_2Q_{33} - f_4Q_{33} = 0 \\
&f_{20}Q_1 - f_{23}Q_1 - f_{10}Q_2 + f_{12}Q_2 + f_2Q_{34} - f_4Q_{34} = 0 \\
&f_{30}Q_1 - f_{33}Q_1 - f_{10}Q_3 + f_{12}Q_3 + f_2Q_{35} - f_4Q_{35} = 0 \\
&f_3Q_2 - f_{16}Q_2 - f_1Q_9 + f_6Q_9 + f_2Q_{36} - f_4Q_{36} = 0 \\
&f_{16}Q_3 - f_3Q_3 - f_2Q_7 + f_4Q_7 + f_2Q_{11} - f_4Q_{11} \\
&\quad + f_1Q_{37} - f_6Q_{37} = 0 \\
&f_5Q_2 - f_{17}Q_2 - f_1Q_{12} + f_6Q_{12} + f_2Q_{38} - f_4Q_{38} = 0 \\
&f_5Q_7 - f_{17}Q_7 - f_3Q_8 + f_{16}Q_8 - f_5Q_{11} + f_{17}Q_{11} \\
&\quad - 2f_1Q_{13} + 2f_6Q_{13} + f_3Q_{39} - f_{16}Q_{39} = 0 \\
&f_5Q_9 - f_{17}Q_9 - f_3Q_{12} + f_{16}Q_{12} + 2f_2Q_{40} - 2f_4Q_{40} = 0 \\
&f_5Q_4 - f_{17}Q_4 - f_3Q_5 + f_{16}Q_5 + f_2Q_{41} - f_4Q_{41} = 0 \\
&f_{20}Q_5 - f_{23}Q_5 - f_{10}Q_{12} + f_{12}Q_{12} + f_2Q_{42} - f_4Q_{42} = 0 \\
&f_{20}Q_4 - f_{23}Q_4 - f_{10}Q_9 + f_{12}Q_9 + f_2Q_{43} - f_4Q_{43} = 0 \\
&f_{21}Q_4 - f_{11}Q_4 + f_{13}Q_5 - f_{24}Q_5 - f_{10}Q_{13} + f_{12}Q_{13} \\
&\quad + f_2Q_{44} - f_4Q_{44} = 0 \\
&f_{21}Q_9 - f_{11}Q_9 + f_{13}Q_{12} - f_{24}Q_{12} - f_{20}Q_{13} + f_{23}Q_{13} \\
&\quad + 2f_2Q_{45} - 2f_4Q_{45} = 0 \\
&f_{30}Q_2 - f_{33}Q_2 - f_{20}Q_3 + f_{23}Q_3 + f_2Q_{46} - f_4Q_{46} = 0 \\
&2f_{24}Q_2 - 2f_{13}Q_2 + f_{20}Q_7 - f_{23}Q_7 + f_{15}Q_9 - f_{27}Q_9 \\
&\quad - f_{20}Q_{11} + f_{23}Q_{11} + 2f_2Q_{47} - 2f_4Q_{47} = 0
\end{aligned}
$$

$$
\begin{aligned}
&2f_{24}Q_1 - 2f_{13}Q_1 + f_{15}Q_4 - f_{27}Q_4 + f_{10}Q_7 - f_{12}Q_7 \\
&\qquad -f_{10}Q_{11} + f_{12}Q_{11} + f_2Q_{48} - f_4Q_{48} = 0 \\
&f_{15}Q_{25} - f_{27}Q_{25} - f_9Q_{28} + f_{26}Q_{28} + 2f_8Q_{94} - 2f_{14}Q_{49} = 0 \\
&f_{30}Q_5 - f_{33}Q_5 - f_{10}Q_{15} + f_{12}Q_{15} + f_2Q_{50} - f_4Q_{50} = 0 \\
&2f_{29}Q_3 - 2f_{11}Q_3 - f_{20}Q_7 + f_{23}Q_7 - f_{15}Q_9 + f_{27}Q_9 \\
&\qquad + f_{20}Q_{11} - f_{23}Q_{11} + 2f_1Q_{51} - 2f_6Q_{51} = 0 \\
&f_{19}Q_{25} - f_{31}Q_{25} - f_{22}Q_{26} + f_{28}Q_{26} + 2f_8Q_{52} - 2f_{14}Q_{52} = 0 \\
&2f_{21}Q_4 - 2f_{11}Q_4 - 2f_{10}Q_{13} + 2f_{12}Q_{13} - f_{19}Q_{15} + f_{31}Q_{15} \\
&\qquad -f_5Q_{30} + f_{17}Q_{30} + f_5Q_{53} - f_{17}Q_{53} = 0 \\
&f_{30}Q_{12} - f_{33}Q_{12} - f_{20}Q_{15} + f_{23}Q_{15} + 2f_2Q_{54} - 2f_4Q_{54} = 0
\end{aligned}
\tag{B.1}
$$