ELSEVIER

A Remarkable Nonlinear Invariant for Evolution with Heterogeneous Rates

VINCENT FERRETTI AND DAVID SANKOFF
Centre de recherches mathématiques, Université de Montréal, Montréal, Québec, Canada H3C 3J7

Received 15 March 1995; revised 3 August 1995

ABSTRACT

A model for DNA or protein sequence evolution is proposed where each position belongs to one of two distinct classes. The two classes evolve at different rates. For a phylogeny on four species, we find a cubic function of 4-tuple occurrence frequencies that is nontrivially invariant no matter what the proportion of positions in each rate class. This result refutes the major criticism of nonlinear polynomial invariants.

1. INTRODUCTION

The use of macromolecular sequences as data for the inference of evolution has given impetus to the study of stochastic models of evolution. Each position in an alignment of the sequences from N organisms whose phylogeny is sought contains information about one sample trajectory of the process, and the totality of this information over all n positions in the alignment should enable us to infer the form of the phylogeny. An evolutionary model is a class of $k \times k$ stochastic matrices representing the nucleotide ($k = 4$) or amino acid ($k = 20$) substitution probabilities over some period of time, such as proposed, for example by Jukes and Cantor [1], Kimura [2, 3], and Cavender [4].[1]

The invariants approach to phylogenetic inference tries to construct an inventory of indicator functions (phylogenetic invariants) of the "spectrum" of the process, different functions for different possible phylogenies, which can then be applied to aggregates of the position-by-position information in the alignment in order to identify which phylogeny actually gave rise to the sequences in this alignment.

Early work on invariants for the case $N = 4$ was based on the Kimura two-parameter model, where $k = 4$, for which linear invariants were

[1] For proteins, the Dayhoff PAM matrices are not stochastic but can be derived from, or can be used to derive, substitution matrices.

MATHEMATICAL BIOSCIENCES 134:71–83 (1996)

655 Avenue of the Americas, New York, NY 10010
0025-5564/96/$15.00
SSDI 0025-5564(95)00108-5

proposed [5], or the Jukes–Cantor model for $k = 2$, for which quadratic invariants were discovered [6]. Both of these models are symmetric; that is, the substitution matrices are symmetric. Each of the n positions was considered to evolve independently, using the same model.

Two distinct traditions have emerged in this field. The class of linear invariants has been characterized exhaustively for several models in a series of papers by Lake [5], Cavender [4, 7], Fu and Li [8, 9], Nguyen and Speed [10], Fu [11], and Steel and Fu [12]. The results on other polynomial invariants, because of their nonlinearity, are less systematic, but many problems have been studied. The effort has been to increase the biological pertinence of the method by relaxing the unrealistic constraints that were imposed to obtain mathematically tractable models in the early studies. Drolet and Sankoff [13], Sankoff [14], and Felsenstein [15] widened the phylogenetic comparison beyond $N = 4$ and $k = 2$ for the Jukes–Cantor model, and a wide variety of other models have been investigated, for many of which there are no linear invariants. This includes models that are asymmetric [16, 17], others where evolution in adjacent positions is not independent [18, 19], and others that can be described as random walks on Abelian groups [18, 20, 21].

One of the advantages often cited for linear invariants in practical applications is that they are not sensitive to inhomogeneities in rates of evolution at different sequence positions while polynomial invariants are valid only for sequences where homogeneity is strictly observed. In this paper, however, we set up a model for evolution where the positions fall into two distinct classes (e.g., RNA secondary structure stems versus single-stranded regions, first two positions of a codon versus the third, mRNA versus noncoding RNA) and find a cubic invariant that is valid no matter what the proportion of positions in each class. This result refutes one of the major criticisms of the utility of polynomial invariants in general and suggests several new lines of inquiry.

2. LINEAR AND NONLINEAR INVARIANTS

To summarize the problem, we want to be able to infer the branching structure of the evolutionary tree $\boldsymbol{T}$ of a group of N observed species. All we know about $\boldsymbol{T}$ is that it contains N terminal vertices, each representing one of the N observed species, and at least one nonterminal vertex, its root, denoted by ρ, such that the flow of time is directed away from ρ on all edges on the paths leading to the terminal vertices.

As data, we have N aligned nucleic sequences (or any other N k-ary sequences form the state space $\boldsymbol{\sigma} = \{1, \ldots, k\}$) of length n, one from each species. For each position i, $1 \leqslant i \leqslant n$, of this alignment, we have a

stochastic model for the observed N-tuple of states, characterized by an (unknown) set $M_i = \{m_{XY}: XY \text{ is an edge of } T\}$, where m_{XY} is the $k \times k$ substitution matrix of a continuous-time Markov process on the state space represented by the edge XY of T. The initial state of the evolutionary process is the state at the sequence position i at the root ρ of T. This state is selected according to a distribution π, the root distribution, on σ. We assume that the set M_i, which we refer to as the set of rates of evolution at position i, is included in a semigroup S of $k \times k$ positive determinant matrices, and each matrix m_{XY} may be considered as a generalized length for the edge XY of the tree. The parametric form of S represents a model for the evolutionary process.

Let $p_M = (p(1), \ldots, p(k^N))$ be the probability distribution for the k possible N-tuples (or configurations) of states at one position given a set M of substitution matrices, one for each edge of T. The vector f containing the expected frequencies, across all positions of the alignment, of each possible configuration of states is called the expected spectrum and is given by the equation

$$f = \frac{1}{n} \sum_{i=1}^{n} p_{M_i}. \tag{1}$$

Note that f depends on the tree T and on the choice of the set M_i for each i, $1 \leqslant i \leqslant n$.

Recall that a phylogenetic invariant for a tree T is a polynomial function Q of the spectrum f that is invariant (e.g., identically zero) with respect to the choice of sets M_i, $1 \leqslant i \leqslant n$, and variable (and nonzero) for all other trees with N terminal vertices. If the value of Q based on the observed frequencies of the configurations along the alignment as estimators of the expected spectrum f is clearly closer to zero than the invariants corresponding to other trees, then T is probably the correct tree, in the sense that the data are generated on T according to the model of evolution adopted.

The usual way of finding an invariant for a tree T is to look for a polynomial function Q of the distribution p_M such that $Q(p_M) = 0$ for all $M \subset S$. If such a function exists, and if it is linear, that is, if it has the form of a scalar product $c \cdot p_M$, where c is a vector of coefficients, then, according to Equation (1),

$$Q(f) = c \cdot f = \frac{1}{n} c \cdot \sum_{i=1}^{n} p_{M_i} = \frac{1}{n} \sum_{i=1}^{n} Q(p_{M_i}) = 0,$$

and thus Q can be seen to be an invariant. If Q, on the other hand, is a nonlinear polynomial, then in order to be sure of finding an invariant we can assume that all positions of the sequence evolve according to the

same choice of matrices, that is, that $M_1 = M_2 = \cdots = M_n = M$. In this case, $f = p_M$ and the function Q is by definition a (nonlinear) invariant. If the M_i are not all the same, we have, for example, in the quadratic case that

$$Q(f) = \sum_{\alpha,\beta} c_{\alpha,\beta} f(\alpha) f(\beta) = \sum_{\alpha,\beta} c_{\alpha,\beta} \frac{1}{n} \sum_{i=1}^{n} p_{M_i}(\alpha) \frac{1}{n} \sum_{j=1}^{n} p_{M_j}(\beta),$$

where α and β range over all configurations (N-tuples) from $\{1,\ldots,k\}$. In order for Q to be invariant, the coefficients would have to be such that all the terms involving configurations at two or more positions cancel each other out for all choices of M_i. Though there is nothing that rules out such a Q, it would seem far from obvious how to construct one.

3. A TWO-RATE PROBLEM

We fix a limited objective: to find nonlinear invariants for a model in which there are just two different "rates," actually two different sets M_1 and M_2 of matrices, and where evolution proceeds according to either one or the other. For example, all the matrices in M_1 may have a large determinant, that is, represent slowly evolving positions, and M_2 may represent the faster rates at other positions, though the "fast–slow" distinction between the two sets of positions need not to be the same for all branches of T. For each θ, $0 \leqslant \theta \leqslant 1$, we will seek a nonlinear invariant for a spectrum f of form

$$\theta p_{M_1} + (1-\theta) p_{M_2},$$

where M_1 and M_2 are two different sets of substitution matrices. We will examine tree T_1 of Figure 1 and adopt the Jukes–Cantor model of evolution as described in the next section.

An explanatory note: We are interested in nonlinear invariants that are not simply functions of linear invariants. The square of an invariant is trivially an invariant, but it tells us nothing additional about the model and does not enable us to infer the phylogeny any better than the original linear invariant does.

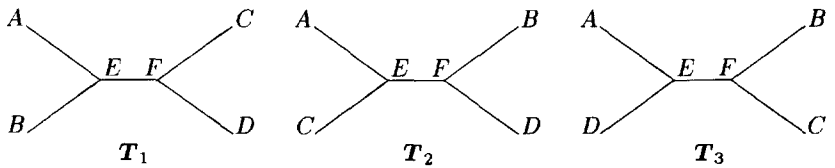

FIG. 1. The three unrooted binary trees on four species.

4. THE JUKES–CANTOR MODEL OF EVOLUTION

The Jukes–Cantor model on the state space $\sigma = \{1,\dots,k\}$ is defined by the semigroup S_{JC} of transition matrices of the form $m_{XY} = aJ + (1-ka)I$, where I is the $k \times k$ identity matrix, J is the $k \times k$ matrix of 1's, and $0 < a < 1-(k-1)a$, so that

$$m_{XY} = \begin{pmatrix} 1-(k-1)a & a & \cdots & a \\ a & 1-(k-1)a & \cdots & a \\ \vdots & \vdots & \vdots & \vdots \\ a & a & \cdots & 1-(k-1)a \end{pmatrix}. \tag{2}$$

This model was originally proposed by Jukes and Cantor [1] for the particular case $k = 4$. For a fixed k, the parameter a completely determines the matrix. Furthermore, we suppose that the root distribution π is uniform, that is, $\pi(i) = 1/k$ for all $i \in \{1,\dots,k\}$. The literature cited above contains a variety of approaches to phylogenetic invariants for the Jukes–Cantor model. For the present purpose, we simply recall here a set of invariants as found by Ferretti and Sankoff [22].

In T_1 let $p_M(\alpha,\beta,\gamma,\delta)$ be the probability of observing the configuration $(\alpha,\beta,\gamma,\delta)$, that is, of observing states $\alpha,\beta,\gamma,\delta$ at a given position for four species A, B, C, and D, respectively, given $M = \{m_{AE}, m_{BE}, m_{CF}, m_{DF}, m_{EF}\}$. We have

$$p_M(\alpha,\beta,\gamma,\delta) = \frac{1}{n} \sum_{i,j \in \sigma} m_{AE}(i,\alpha) m_{BE}(i,\beta) m_{CF}(j,\gamma) m_{DF}(j,\delta) m_{EF}(i,j). \tag{3}$$

Denote

$$\begin{aligned} f_0 &= \sum_{i \in \sigma} f(i,i,i,i), & f_1 &= \sum_{i,j \in \sigma} f(i,i,i,j), \\ f_2 &= \sum_{i,j \in \sigma} f(i,i,j,i), & f_3 &= \sum_{i,j \in \sigma} f(i,i,j,j), \\ f_4 &= \sum_{i,j \in \sigma} f(i,j,i,i), & f_5 &= \sum_{i,j \in \sigma} f(i,j,i,j), \\ f_6 &= \sum_{i,j \in \sigma} f(i,j,j,i), & f_7 &= \sum_{i,j \in \sigma} f(i,i,j,j), \\ f_8 &= \sum_{i,j,l \in \sigma} f(i,i,j,l), & f_9 &= \sum_{i,j,l \in \sigma} f(i,j,i,l), \\ f_{10} &= \sum_{i,j,l \in \sigma} f(i,j,l,i), & f_{11} &= \sum_{i,j,l \in \sigma} f(i,j,l,l), \\ f_{12} &= \sum_{i,j,l \in \sigma} f(i,j,l,j), & f_{13} &= \sum_{i,j,l \in \sigma} f(i,j,j,l), \\ & & f_{14} &= \sum_{i,j,l,m \in \sigma} f(i,j,l,m), \end{aligned} \tag{4}$$

where f contains the expected frequencies of the configurations across all positions of the alignment. Note that $\Sigma_i f_i = 1$. Then the following 12 invariants linearly span the set of all quadratic invariants of the tree T_1.

$$
\begin{aligned}
L_1 &= [k-3](f_9 + f_{12} - [k-2]f_5) - f_{14}, \\
L_2 &= [k-3](f_{10} + f_{13} - [k-2]f_6) - f_{14}, \\
Q_1 &= [k-2]([k-1]f_0 f_6 + f_3 f_9 + f_3 f_{12} \\
&\qquad - f_1 f_7 - f_2 f_4 - [k-3]f_3 f_5) - f_8 f_{11}, \\
Q_2 &= ([k-1]f_0 - f_3)(f_6 - f_5) + (f_1 - f_2)(f_4 - f_7), \\
Q_3 &= ([k-1]f_0 - f_3)(f_{10} - f_9) + (f_1 - f_2)([k-2]f_4 - f_{11}), \\
Q_4 &= [k-2]([k-1]f_0 - f_3)(f_{12} - [k-2]f_5) \\
&\quad + ([k-2]f_4 - f_{11})([k-2]f_1 - f_8), \\
Q_5 &= [k-2]([k-1]f_0 - f_3)(f_9 - [k-2]f_5) \\
&\quad + ([k-2]f_7 - f_{11})([k-2]f_2 - f_8), \\
Q_6 &= (f_{11} - [k-2]f_4)(f_9 - f_{13}) + [k-2]([k-2]f_5 - f_{12})(f_4 - f_7), \\
Q_7 &= (f_1 - f_2)(f_{10} - f_{12}) + ([k-2]f_2 - f_8)(f_6 - f_5), \\
Q_8 &= (f_{11} - [k-2]f_4)(f_6 - f_5) + (f_4 - f_7)(f_{10} - f_9), \\
Q_9 &= [k-2](f_1 - f_2)([k-2]f_5 - f_{12}) + ([k-2]f_1 - f_8)(f_{10} - f_9), \\
Q_{10} &= [k-2](f_6 - f_5)([k-2]f_5 - f_{12}) + (f_{10} - f_9)(f_{13} - f_9).
\end{aligned}
\tag{5}
$$

We stress that $Q_1, \ldots, Q_{10}$ are invariant only under the hypothesis of rate homogeneity across all sequence positions. Note also that the 12 polynomials in (5) are not algebraically independent [22].

5. THE METHOD

Let

$$
f = \theta p_{M_1} + (1-\theta) p_{M_2} \tag{6}
$$

be the spectrum for the tree T_1, given a coefficient θ, $0 \leqslant \theta \leqslant 1$, a pair (M_1, M_2) of sets of transition matrices $M_1 \subset S_{JC}$ and $M_2 \subset S_{JC}$, and a distribution p_m as defined by Equation (3). We wish to find polynomial invariants Q of degree d, $d \geqslant 1$, that is, invariants of form

$$
Q = Q(f, \lambda) = \sum_{0 \leqslant i_1 \leqslant \cdots \leqslant i_d \leqslant 14} \lambda_{i_1 \cdots i_d} f_{i_1} \cdots f_{i_d}, \tag{7}
$$

where $f_0, \ldots, f_{14}$ are defined by (4) and $\lambda = (\lambda_{i_1 \ldots i_d})_{0 \leqslant i_1 \leqslant \cdots \leqslant i_d \leqslant 14}$ is a vector of m coefficients, $m = (15 + d - 1)!/(d!14!)$.

The problem becomes that of determining all $\boldsymbol{\lambda}$ for which the function Q is invariant over all $(\boldsymbol{M}_1, \boldsymbol{M}_2)$, that is, identically equal to zero, independent of the specific parameters associated with each of the edges.

Since Q is to be invariant with respect to the parameters of the model, we simply choose m pairs of sets $(\boldsymbol{M}_1(i), \boldsymbol{M}_2(i))$ of matrices for $\boldsymbol{T}_1$ at random, $1 \leqslant i \leqslant m$, calculate explicitly the distribution $f(i)$ for each set, and set up the system:

$$\begin{aligned} &\sum_{0 \leqslant i_1 \leqslant \cdots \leqslant i_d \leqslant 14} \lambda_{i_1 \ldots i_d} f_{i_1}(1) \cdots f_{i_d}(1) = 0, \\ &\qquad \vdots \\ &\sum_{0 \leqslant i_1 \leqslant \cdots \leqslant i_d \leqslant 14} \lambda_{i_1 \cdots i_d} f_{i_1}(m) \cdots f_{i_d}(m) = 0, \end{aligned} \tag{8}$$

which is a system of homogeneous linear equations in the unknown $\lambda_{i_1 \cdots i_d}$,

$$\boldsymbol{G}\boldsymbol{\lambda} = \boldsymbol{O}, \tag{9}$$

where $\boldsymbol{G}$ is a $m \times m$ matrix with elements

$$g_{hr} = f_{i_1}(h) \cdots f_{i_d}(h),$$

$\lambda_{i_1 \cdots i_d}$ being the rth component of $\boldsymbol{\lambda}$. The set of invariants of form Q is necessarily contained in the set of nontrivial solutions of this system. The set of solutions to $\boldsymbol{G}\boldsymbol{\lambda} = \boldsymbol{O}$ defines the kernel of the matrix $\boldsymbol{G}$, denoted $\mathrm{Ker}(\boldsymbol{G})$. This is a vector subspace of dimension equal to $m - \mathrm{rank}(\boldsymbol{G})$, for which the simplest basis is a set of vectors expressing the linear dependences among the columns of $\boldsymbol{G}$.

The key to the choice of the m pairs $(\boldsymbol{M}_1(i), \boldsymbol{M}_2(i)$ is to ensure that there are no extra solutions to $\boldsymbol{G}\boldsymbol{\lambda} = \boldsymbol{O}$ because of accidental dependences among its columns. In practice, of course, with pseudorandom generators this cannot be ensured, but this is of no mathematical importance, because spurious invariants are, as we shall see, easily detected and discarded.

6. RESULTS

We first apply our method to look for all quadratic invariants for $\boldsymbol{T}_1$, that is, invariants of form (8) with $d = 2$.

For a fixed $k \geqslant 4$ and θ, $0 \leqslant \theta \leqslant 1$, we randomly construct $m = 120$ pairs $(\boldsymbol{M}_1(i), \boldsymbol{M}_2(i))$, for each of which we calculate $f_0(i), \ldots, f_{14}(i)$. With

these we construct the 120×120 matrix $\boldsymbol{G}$ of the system of equations (9). In solving this, we find that $\text{rank}(\boldsymbol{G}) = 91$ for state spaces of size $k = 4,\ldots,20$, so that the canonical basis of the subspace $\text{Ker}(\boldsymbol{G})$ contains 29 elements. It turns out that all of the 29 invariants thus determined are only trivially invariant, as they can be factored as $f_i \times L_1$ or $f_i \times L_2$, $0 \leqslant i \leqslant 15$, where L_1 and L_2 are the linear invariants for $\boldsymbol{T}_1$ in (5).

We are thus obliged to carry our search for invariants to the next higher degree of polynomial, namely $d = 3$, or cubic. This entails the construction of a 680×680 matrix $\boldsymbol{G}$, for which we find, for $k = 4,\ldots,20$, a rank of 454. Among the $680-454 = 196$ invariants thus obtained, 191 are of form $f_i f_j L_1$ or $f_i f_j L_2$ or RL_i, for some other quadratic polynomial R. We are left with five invariants, which, for $k = 4$, can be written

$$\begin{aligned}
C_1 = {} & f_1f_{11}f_{12} - f_1f_{10}f_{11} + f_{10}f_{11}f_2 - f_{11}f_{12}f_2 - f_{10}^2f_3 + f_{10}f_{12}f_3 \\
& + f_{10}f_4f_8 - f_{11}f_5f_8 + f_{11}f_6f_8 - f_{10}f_7f_8 + f_{10}f_3f_9 - f_{12}f_3f_9 \\
& - f_4f_8f_9 + f_7f_8f_9 - 3(f_0f_{10}f_{12} + f_0f_{10}f_9 - f_0f_{12}f_9 - f_0f_{10}^2) \\
& + 2(f_{11}f_2f_5 + f_{12}f_2f_7 + f_{12}f_3f_5 + f_{10}f_3f_6 + f_1f_4f_9 + f_3f_5f_9 \\
& \qquad + f_1f_{10}f_7 + f_4f_5f_8 - f_{10}f_2f_4 - f_{10}f_3f_5 - f_{11}f_2f_6 \\
& \qquad - f_{12}f_3f_6 - f_1f_{12}f_7 - f_4f_6f_8 - f_3f_6f_9 - f_1f_7f_9) \\
& + 4(f_3f_5f_6 + f_1f_5f_7 + f_2f_4f_6 - f_1f_4f_5 - f_3f_5^2 - f_2f_5f_7) \\
& + 6(f_0f_{10}f_5 + f_0f_{12}f_6 + f_0f_6f_9 - f_0f_{12}f_5 - f_0f_{10}f_6 \\
& \qquad - f_0f_5f_9 + 2f_0f_5^2 - 2f_0f_5f_6),
\end{aligned} \tag{10a}$$

$$\begin{aligned}
C_2 = {} & f_4f_8f_9 - f_4f_8f_{10} - f_3f_9f_{10} + f_3f_{10}^2 + f_5f_8f_{11} - f_6f_8f_{11} \\
& + f_1f_{10}f_{11} - f_2f_{10}f_{11} + f_7f_8f_{12} + f_3f_9f_{12} \\
& - f_3f_{10}f_{12} - f_1f_{11}f_{12} + f_2f_{11}f_{12} - f_7f_8f_{13} \\
& + 4(f_1f_4f_5 + f_3f_5^2 - f_2f_4f_6 - f_3f_5f_6 - f_1f_5f_7 + f_2f_5f_7) \\
& + 3(f_0f_9f_{10} - f_0f_{10}^2 + f_0f_{10}f_{12} - f_0f_9f_{12}) \\
& + 2(f_4f_6f_8 + f_6f_7f_8 + f_3f_6f_9 + f_1f_7f_9 + f_2f_4f_{10} + f_2f_6f_{11} \\
& \qquad + f_3f_5f_{10} + f_1f_7f_{12} + f_3f_6f_{12} - f_4f_5f_8 - f_5f_7f_8 - f_1f_4f_9 \\
& \qquad - f_3f_5f_9 - f_3f_6f_{10} - f_1f_7f_{10} - f_2f_7f_{12} - f_3f_5f_{12} - f_2f_5f_{11}) \\
& + 6(f_0f_5f_{12} - f_0f_6f_{12} + f_0f_6f_{10} + f_0f_5f_9 + 2f_0f_5f_6 \\
& \qquad - f_0f_5f_{10} - f_0f_6f_9 - 2f_0f_5^2),
\end{aligned} \tag{10b}$$

$$C_3 = 8(3f_5^2f_6 + f_6^3 - f_5^3 - 3f_5f_6^2) + 4(3f_5^2f_9 - 6f_5f_6f_9 + 3f_6^2f_9 + 3f_5^2f_{12}$$
$$+ 6f_5f_6f_{10} + 3f_6^2f_{12} - 3f_6^2f_{10} - 6f_5f_6f_{12} - 3f_5^2f_{10})$$
$$+ 2(3f_6f_9^2 + 6f_5f_9f_{10} + 6f_5f_{10}f_{12} + 3f_6f_{10}^2 + 6f_6f_9f_{12} + 3f_6f_{12}^2 + 3f_5f_9^2$$
$$- 6f_6f_9f_{10} - 3f_5f_{10}^2 - 6f_5f_9f_{12} - 6f_6f_{10}f_{12} - 3f_5f_{12}^2) \quad (10c)$$
$$+ f_9^3 - 3f_9^2f_{10} + 3f_9f_{10}^2 - f_{10}^3 + 3f_9^2f_{12} - 6f_9f_{10}f_{12}$$
$$+ 3f_{10}^2f_{12} + 3f_9f_{12}^2 - 3f_{10}f_{12}^2 + f_{12}^3 - f_{13}^3,$$

$$C_4 = -8(f_5^3 + f_5f_6^2 - 2f_5^2f_6)$$
$$+ 2(4f_5f_{10}f_{12} + 4f_6f_9f_{12} + 2f_6f_9^2 - 3f_5f_9^2 + 4f_5f_9f_{10} - 2f_6f_9f_{10}$$
$$- 3f_5f_{12}^2 - f_5f_{10}^2 - 2f_6f_{10}f_{12} - 6f_5f_9f_{12} + 2f_6f_{12}^2)$$
$$+ 4(3f_5^2f_9 + f_6^2f_9 - 2f_5^2f_{10} - 4f_5f_6f_9 + 2f_5f_6f_{10} \quad (10d)$$
$$+ f_6^2f_{12} + 3f_5^2f_{12} - 4f_5f_6f_{12})$$
$$+ f_9^3 - 2f_9^2f_{10} + f_9f_{10}^2 + 3f_9^2f_{12} - 4f_9f_{10}f_{12} + f_{10}^2f_{12}$$
$$+ 3f_9f_{12}^2 - 2f_{10}f_{12}^2 + f_{12}^3 - f_{13}^2f_{14},$$

and

$$C_5 = -8f_5^3 + 8f_5^2f_6 + 2(6f_5^2f_9 - 4f_5f_6f_9 - 2f_5^2f_{10} + 6f_5^2f_{12} - 4f_5f_6f_{12})$$
$$+ 2(f_6f_9^2 + 2f_5f_9f_{10} + 2f_6f_9f_{12} + 2f_5f_{10}f_{12} - 3f_5f_{12}^2 - 6f_5f_9f_{12}$$
$$- 3f_5f_9^2 + f_6f_{12}^2) + f_9^3 - f_9^2f_{10} + f_9^2f_{12} \quad (10e)$$
$$- 2f_9f_{10}f_{12} + 3f_9f_{12}^2 - f_{10}f_{12}^2 + f_{12}^3 - f_{13}f_{14}^2.$$

A remarkable aspect of these invariants $C_1,\ldots,C_5$ is that they do not depend on the proportion θ. At first this was only an empirical observation based on carrying out our method for several values of θ. But in substituting (2) and (3) in (6), and then (4) in (10), symbolic computing provides an analytic proof that the polynomials $C_1,\ldots,C_5$ are truly invariant for T_1 for all θ, $0 \leqslant \theta \leqslant 1$.

7. A QUESTION OF TRIVIALITY

The key question remaining is whether invariance of the polynomials $C_1,\ldots,C_5$ for a spectrum combining two rates of evolution is not simply a consequence of some functional relation between them and the linear invariants L_1 and L_2. For example, are there non-null functions u_i and v_i of the variables $f_0,\ldots,f_{14}$ such that, for each $1 \leqslant i \leqslant 5$,

$$C_i = u_i L_1 + v_i L_2?$$

There does not seem to be any particular algebraic method for answering this question in its most general form. If, however, we look only at the special case where u_i and v_i are polynomials in $f_0,\ldots,f_{14}$, this question is the same as asking whether the polynomial C_i is an element of the ideal[2] generated by L_1 and L_2. To answer this, we can use an algebraic method based on Gröbner bases [23, 24].

This method allows us to determine whether a polynomial p belongs to the ideal generated by a set of polynomials $q_1,\ldots,q_r$, by "reducing" it with respect to $q_1,\ldots,q_r$ by means of an algorithm $A = A[p; q_1,\ldots,q_r]$.

The Gröbner basis $g(q_1,\ldots,q_r)$ of the ideal $\langle q_1,\ldots,q_r\rangle$ is, by definition, the unique basis of $\langle q_1,\ldots,q_r\rangle$ for which $A[p; g(q_1,\ldots,q_r)] = 0$ if and only if $p \in \langle q_1,\ldots,q_r\rangle$. Thus, to answer our question, we must first calculate $g(L_1, L_2)$ in order to be then able to evaluate $A[C_i; g(L_1, L_2)]$ for each i, $1 \leqslant i \leqslant 5$. This can be accomplished using a symbolic computing package like Maple, in which there are two procedures, gbasis and normalf, appropriate for carrying out these two steps.

Thus, by evaluating $A[C_i; g(L_1, L_2)]$ for each i, $1 \leqslant i \leqslant 5$, we were able to find that C_3, C_4, and C_5 are elements of $\langle L_1, L_2\rangle$ but C_1 and C_2 are not. In addition, we find that $A[C_2; g(L_1, L_2, C_1)] = 0$, meaning that the invariance of C_1 implies that of C_2 and vice versa. It suffices then to consider only C_1, which can be written by comparing its form for different values of k, as

$$\begin{aligned} C_1 = {} & Q_3(L_2 - L_1) + (f_2 - f_1)Q_6 + (f_8 - (k-2)f_1)Q_8 \\ & + (f_3 - (k-1)f_0)Q_{10}, \end{aligned}$$

---

[2] Given a finite set of polynomials $W = \{w_1,\ldots,w_r\}$ in a field $F[x_1,\ldots,x_n]$, the set

$$\left\{ \sum_{i=1}^{r} \kappa_i w_i : \kappa_i \in F[x_1,\ldots,x_n] \right\}$$

is called the *ideal* generated by W, and it will be denoted by $\langle W\rangle$ or $\langle w_1,\ldots,w_r\rangle$. The set W is then said to form a basis for this ideal.

where L_1, L_2, Q_3, Q_6, Q_8, and Q_{10} are given by (5). Symbolic computing again proves, for values of k up to 20, that this generalized form of C_1 is effectively an invariant. The term $Q_3(L_2 - L_1)$ being itself an invariant, we arrive finally at the conclusion that the cubic polynomial

$$C = (f_2 - f_1)Q_6 + (f_8 - (k-2)f_1)Q_8 + (f_3 - (k-1)f_0)Q_{10} \quad (11)$$

is also an invariant for the tree T_1 and a spectrum of form (6).

To solve our problem in greater generality, we ask whether there is any function $Z(L_1, L_2)$ that is zero whenever L_1 and L_2 are, and for which

$$C = Z(L_1, L_2). \quad (12)$$

We argue as follows. Consider a spectrum made up of three evolutionary rates, that is, a spectrum of the form

$$\theta \boldsymbol{p}_{M_1} + \omega \boldsymbol{p}_{M_2} + (1 - \theta - \omega)\boldsymbol{p}_{M_3}, \quad (13)$$

for $\theta, \omega \geqslant 0$, $\theta + \omega \leqslant 1$, the tree T_1, and the Jukes–Cantor model. Given that L_1 and L_2 remain invariants for a spectrum of this form, the existence of a relation of type (12) would imply the invariance of C with respect to the choice of sets of matrices $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$. Now, by substituting (2), (3), and (13) in (11), we can show that this does not hold true; C is not an invariant for a spectrum made up of three evolutionary rates. This proves that there can exist no relation of form (12) among C, L_1, and L_2.

The polynomial C is thus a nontrivial invariant. Furthermore, through making the appropriate substitutions, it is easy to show that C is a phylogenetic invariant in that C is nonzero for spectra generated on the two other unrooted binary trees T_2 and T_3 in Figure 1.

8. DISCUSSION AND CONCLUSION

The importance of the invariant C does not lie in its potential use for phylogenetic inference, since we already have linear invariants that do the same thing. Its existence, however, refutes the common assumption that nonlinear invariants are restricted by the unrealistic assumption of homogeneity of rates. Steel et al. [18] also studied nonlinear invariants allowing different choice of matrices $\boldsymbol{M}$ at all positions of the sequence. For the Kimura [3] three-parameter model, they found invariant polynomials of degree 32 for trees on four species.

The fact that C does not depend on θ is unexpected. Not only do we not need to know which positions have the same rate, we do not even have to know how many positions there are of each type.

Our method could be used to search for invariants for spectra of form (13). Or the distribution of rates could be modeled by a one- or two-parameter distribution and invariants found for this model. The next step would be to find a heterogeneous rate context in which there are no linear invariants but one or more nonlinear ones.

Research supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. D. S. is a fellow of the Canadian Institute for Advanced Research.

REFERENCES

1 T. H. Jukes and C. R. Cantor, Evolution of protein molecules, in *Mammalian Protein Metabolism*, H. N. Munro, Ed., Academic, New York, 1969, pp. 21–132.
2 M. Kimura, A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, *J. Mol. Evol.* 16:111–120 (1980).
3 M. Kimura, Estimation of evolutionary sequences between homologous nucleotide sequences, *Proc. Natl. Acad. Sci. U.S.A.* 78:454–458 (1981).
4 J. A. Cavender, Mechanized derivation of linear invariants, *Mol. Biol. Evol.* 6:301–316 (1989).
5 J. A. Lake, A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony, *Mol. Biol. Evol.* 4:167–191 (1987).
6 J. A. Cavender and J. Felsenstein, Invariants of phylogenies: simple case with discrete states, *J. Classif.* 4:57–71 (1987).
7 J. A. Cavender, Necessary conditions for the method of inferring phylogeny by linear invariants, *Math. Biosci.* 103:69–75 (1991).
8 Y. X. Fu and W. H. Li, Necesasry and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree, *Math. Biosci.* 108:203–218 (1992).
9 Y. X. Fu and W. H. Li, Construction of linear invariants in phylogenetic inference, *Math. Biosci.* 109:201–228 (1992).
10 T. Nguyen and T. P. Speed, A derivation of all linear invariants for a nonbalanced transversion model, *J. Mol. Evol.* 35:60–76 (1992).
11 Y. X. Fu, Linear invariants under Jukes and Cantor's one parameter model, *J. Theor. Biol.* 173(4): 353–360 (1995).
12 M. A. Steel and Y. X. Fu, Classifying and counting linear phylogenetic invariants for the Jukes & Cantor model, *J. Comput. Biol.* 2:39–48 (1995).
13 S. Drolet and D. Sankoff, Quadratic tree invariants for multivalued characters, *J. Theor. Biol.* 144:117–129 (1990).
14 D. Sankoff, Designer invariants for large phylogenies, *Mol. Biol. Evol.* 7:255–269 (1990).
15 J. Felsenstein, Counting phylogenetic invariants in some simple cases, *J. Theor. Biol.* 152:357–376 (1991).
16 V. Ferretti, B. F. Lang, and D. Sankoff, Skewed base compositions, asymmetric transition matrices, and phylogenetic invariants, *J. Comput. Biol.* 1:77–92 (1994).

17 M. Steel, Recovering a tree from the leaf colorations it generates under a Markov model, *Appl. Math. Lett.* 7:19–23 (1994).
18 M. A. Steel, L. A. Szekely, P. L. Erdos, and P. Waddell, A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model, *N.Z. J. Bot.* 31:289–296 (1993).
19 V. Ferretti and D. Sankoff, Phylogenetic invariants for more general evolutionary models, *J. Theor. Biol.* 173:147–162 (1995).
20 S. N. Evans and T. P. Speed, Invariants of some probability models used in phylogenetic inference, *Ann. Stat.* 21:355–377 (1993).
21 L. A. Szekely, M. A. Steel, and P. L. Erdos, Fourier calculus on evolutionary trees, *Adv. Appl. Math.* 14:200–216 (1993).
22 V. Ferretti and D. Sankoff, The empirical discovery of phylogenetic invariants, *Adv. Appl. Probab.* 25:290–302 (1993).
23 W. Adams and P. Loustaunau, *An Introduction to Gröbner Bases*, American Mathematical Society, Providence, RI, 1994.
24 B. Buchberger, Gröbner bases: an algorithmic method in polynomial ideal theory, in *Recent Trends in Multidimensional Systems Theory*, N. K. Bose, Ed., Reidel, Dordrecht, 1985.