J. theor. Biol. (1990) **144**, 117-129

Quadratic Tree Invariants for Multivalued Characters

SUZANNE DROLET AND DAVID SANKOFF†

Centre de recherches mathématiques, Université de Montréal, C.P. 6128 *Succursale A, Montreal, Quebec, Canada H3C 3J7*

(*Received on* 14 *October* 1988, *Accepted in revised form on* 20 *November* 1989)

Studying Markov models for binary character evolution along the branches of unrooted four-trees, Cavender & Felsenstein found a set of branch-length invariants in the case of symmetric transition probabilities. This involved three expressions, K_1, K_2 and K_3, quadratic in the predicted frequencies of occurrence of each possible configuration of character values on the four-tree: $f(0000)$, $f(1000)$, ... Denoting by 1, 2 and 3 the three possible completely resolved unrooted four-trees, K_i is predicted to be always zero (invariant) only if tree i generated the data, *independent of the branch length of the tree.*

Generalization to characters other than binary is difficult because of the computational size of the problem—when the Cavender-Felsenstein method is applied directly to the case of three-valued characters, a quartic polynomial involving 22 050 terms results. Algebraic manipulation with the help of *MACSYMA*, however, shows that there are quadratic branch-length invariants in this case as well.

Similarities in the form of the binary and trinary character invariants suggests a form for the case of four-valued characters and numerous tests confirm this. It is this case which will be of use in phylogenetic reconstruction based on nucleotide sequence data.

We discuss quadratic invariants produced by other methods as well as linear invariants such as those of Lake. Generalizations to larger numbers of character values, larger trees, and wider classes of transition matrices are discussed.

Introduction

Recent debates on early evolution (Lake, 1987, 1988; Olsen, 1987) have hinged in part on the mathematical properties of the methodologies used to infer phylogenetic relationships of a number of organisms, when applied to nucleotide sequence data from these organisms. Under simple probabilistic models for the evolution over time of an undirected character (where changes in character value are freely reversible, as is the case with nucleotide substitutions), it can be shown that the most economical explanation of the data (the "parsimony" criterion) contains a serious bias (Felsenstein, 1983; Lake, 1987): when the underlying, or true, phylogenetic tree is such that the overall rate of evolutionary change differs greatly from one branch to another, the most parsimonious explanation of the data may consist of an incorrect tree where slowly evolving lineages ("short branches") are grouped

†Address correspondence to: Centre de recherches mathématiques, Université de Montréal, C.P. 6128, Succursale "A", Montréal H3C 3J7. This work was supported by grants from the National Science and Engineering Research Council of Canada. David Sankoff is a Fellow of the Canadian Institute for Advanced Research.

0022-5193/90/090117+13 $03.00/0

together, despite their not being closely related, in opposition to more rapidly evolving lineages ("long branches").

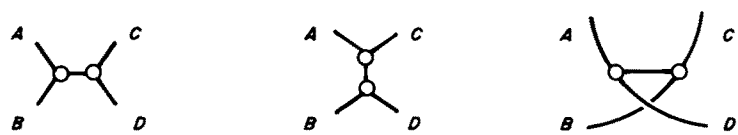
Calculation of the maximum likelihood solution for the tree structure is one approach to circumventing this bias whenever probabilistic models seem justified. This task, however, is computationally costly: not only does the true tree topology need to be reconstructed or approximated, which is a major difficulty in all phylogenetic inference, but each topology tested must also be optimized simultaneously with respect to all branch lengths, since the likelihood depends on these lengths.

One approach to simplifying the problem is exemplified by the work of Cavender & Felsenstein (1987), Lake (1987) and Cavender (1989). The idea is to find a function of the data and of tree topology which is predicted to be invariant with respect to branch length *for the correct tree.* Thus Lake found invariant linear combinations of character value configuration frequencies at the terminal nodes of a "four-tree", an unrooted binary tree with four terminal nodes and two interior nodes, using a probabilistic model of nucleotide replacement which distinguished between transition and transversion mutations among the four nucleotides. Cavender characterized the entire set of such linear invariants under still more general models of nucleotide replacement. Similarly Cavender & Felsenstein found quadratic invariant functions of the same sort of frequency data on four-trees, but for two-valued characters only.

In this paper we generalize the work of Cavender & Felsenstein to multivalued characters under fully symmetric replacement models as a first step towards the analysis of quadratic invariants in more general models.

The Model

We have only four terminal nodes (i.e. species) A, B, C and D. Thus we only need distinguish among three possible binary unrooted trees, which we label 1, 2 and 3.



1 2 3

In our probabilistic model of evolution of n-ary characters (i.e. each character can take on values in the set $\{1, 2, \ldots, n\}$) along the five edges (branches) of the four-tree, we make three simplifying assumptions. We assume first that random character value changes follow a continuous-time Markov process where all characters have the same rate parameter, which may, however, vary over time. Second, we assume complete symmetry among the character values, i.e. that at any given moment in time there are equal transition rates between any two character values (cf. Jukes & Cantor, 1969). Finally, we assume all character values are equally probable (probability $= 1/n$) at one of the nodes of the tree. The last two assumptions together imply that all character values are equally probable at all points in the tree.

Consider tree 1 above. Let E denote the interior node adjacent to A and B, and F denote the interior node adjacent to C and D. Let $\mathbf{P}_{XY}$ be the $n \times n$ matrix of transition probabilities among the different character values between node X and node Y. It is a consequence of the symmetry assumptions in the model that $\mathbf{P}_{YX} = \mathbf{P}_{XY}$, and that this is true whether or not X, Y or some other node or point on a branch of the tree is designated as the root. Then the Chapman-Kolmogorov equation (Feller, 1957: 424) assures that:

$$\begin{aligned}
\mathbf{P}_{AB} &= \mathbf{P}_{AE}\mathbf{P}_{EB} \\
\mathbf{P}_{AC} &= \mathbf{P}_{AE}\mathbf{P}_{EF}\mathbf{P}_{FC} \\
\mathbf{P}_{AD} &= \mathbf{P}_{AE}\mathbf{P}_{EF}\mathbf{P}_{FD} \\
\mathbf{P}_{BC} &= \mathbf{P}_{BE}\mathbf{P}_{EF}\mathbf{P}_{FC} \\
\mathbf{P}_{BD} &= \mathbf{P}_{BE}\mathbf{P}_{EF}\mathbf{P}_{FD} \\
\mathbf{P}_{CD} &= \mathbf{P}_{CF}\mathbf{P}_{FD}.
\end{aligned} \tag{1}$$

Because of the symmetric transition rates, each of the transition matrices $\mathbf{P}$ in (1) is of the form;

$$\mathbf{P} = s\mathbf{J} + [1 - ns]\mathbf{I}, \tag{2}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{J}$ is the $n \times n$ matrix of 1's, and

$$0 < s < 1 - (n-1)s. \tag{3}$$

In other words the diagonal elements are larger than the off-diagonal elements. In the next sections we will need to know the probability p that two nodes X and Y, whose character values are related by the transition matrix $\mathbf{P}$, have identical values for a character. Using (2) and the equal probabilities property at node X,

$$\begin{aligned}
p &= \sum_{i=1}^{n} \text{Prob (value } i \text{ at node } X)\mathbf{P}(i, i) \\
&= n(1/n)[1 - (n-1)s] \\
&= 1 - (n-1)s.
\end{aligned} \tag{4}$$

Let T be the determinant of $\mathbf{P}$. It can be shown, using (3) and the fact that a matrix of form (2) is a scalar times an "equicorrelation matrix" (cf. Mardia *et al.* 1979: 461-462) that;

$$0 < T = (1 - ns)^{n-1} < 1. \tag{5}$$

From (1),

$$\begin{aligned}
T_{AB} &= T_{AE}T_{EB} \\
T_{AC} &= T_{AE}T_{EF}T_{FC} \\
T_{AD} &= T_{AE}T_{EF}T_{FD} \\
T_{BC} &= T_{BE}T_{EF}T_{FC} \\
T_{BD} &= T_{BE}T_{EF}T_{FD} \\
T_{CD} &= T_{CF}T_{FD},
\end{aligned} \tag{6}$$

since the determinant of a product is the product of the determinants. It follows directly that;

$$T_{AC}T_{BD} - T_{AD}T_{BC} = T_{AE}T_{EF}T_{FC}T_{BE}T_{EF}T_{FD} - T_{AE}T_{EF}T_{FD}T_{BE}T_{EF}T_{FC} = 0, \quad (7)$$

independent of the values of s in (2) for the transition matrices in (1). The quantity $T_{AC}T_{BD} - T_{AD}T_{BC}$ is thus called an *invariant* for tree 1. On the other hand,

$$\begin{aligned} T_{AD}T_{BC} - T_{AB}T_{CD} &= T_{AE}T_{EF}T_{FD}T_{BE}T_{EF}T_{FC} - T_{AE}T_{EB}T_{FC}T_{FD} \\ &= T_{AE}T_{EB}T_{FC}T_{FD}(T_{EF}^2 - 1) < 0, \end{aligned} \quad (8)$$

and

$$\begin{aligned} T_{AB}T_{CD} - T_{AC}T_{BD} &= T_{AE}T_{EB}T_{FC}T_{FD} - T_{AE}T_{EF}T_{FC}T_{BE}T_{EF}T_{FD} \\ &= T_{AE}T_{EB}T_{FC}T_{FD}(1 - T_{EF}^2) > 0, \end{aligned} \quad (9)$$

since all the determinants are positive and less than 1. The values of $T_{AB}T_{CD} - T_{AC}T_{BD}$ and of $T_{AD}T_{BC} - T_{AB}T_{CD}$ thus do depend on s in (2) for the various $\mathbf{P}$ in (1), so they are not invariant.

Now, the path between A and C in tree 2, for example, is different from that in tree 1, not traversing the interior branch as it does in the latter, so that (7) will not hold in tree 2. In fact, we have the following relationships:†

	Tree 1	Tree 2	Tree 3
$T_{AC}T_{BD} - T_{AD}T_{BC}$	0	>0	<0
$T_{AD}T_{BC} - T_{AB}T_{CD}$	<0	0	>0
$T_{AB}T_{CD} - T_{AC}T_{BD}$	>0	<0	0

(10)

The expressions $T_{AD}T_{BC} - T_{AB}T_{CD}$ and $T_{AB}T_{CD} - T_{AC}T_{BD}$ are thus invariants for trees 2 and 3, respectively.

†These results are a generalization of the "four point property" (Buneman, 1974; Dobson, 1974) of "additive trees". In tree 1, denote the sum of the distances (or times) along the three branches between A and C, for example, as AC. Then

	Tree 1	Tree 2	Tree 3
$(AC+BD)-(AD+BC)$	0	<0	>0
$(AD+BC)-(AB+CD)$	>0	0	<0
$(AB+CD)-(AC+BD)$	<0	>0	0

The relationship these equations and inequalities and those in (10) can be understood in terms of the fact that $-\log T$ is a generalized measure of time or evolutionary distance.

The Data

How can the invariants help us select the "true" tree, the one which generated the observed data? If we could calculate the three expressions in (10) as functions of the data, we could pick the tree associated with the single one which is equal to zero. The invariants in (10), however, are functions of the transition probabilities, while the data consist of observed configurations of character values at nodes A, B, C and D. What is the connection between the probabilities and these data?

Let f_{uvxy} be the probability that the configuration with value u at A, v at B, x at C and y at D, occurs. As is generally true in probabilistic models, if the data consist of many characters, we can use the proportion g_{uvxy} of these characters for which the same configuration occurs as an estimate for the probability f_{uvxy}.

It remains to connect the configuration probabilities f with the invariant expressions of the T via the transition probabilities in the matrices **P**. This is the crux of the entire problem.

The Cavender–Felsenstein Results

With Cavender & Felsenstein, we first treat the case of binary characters, i.e. where $n=2$. Here, each transition matrix is of the form;

$$\begin{vmatrix} 1-s & s \\ s & 1-s \end{vmatrix} \tag{11}$$

so that, as in (5),

$$T=(1-s)^2-s^2=1-2s. \tag{12}$$

Denote

$$\begin{aligned} f_0&=f_{1111}+f_{2222}\\ f_1&=f_{1112}+f_{2221}\\ f_2&=f_{1121}+f_{2212}\\ f_3&=f_{1122}+f_{2211}\\ f_4&=f_{1211}+f_{2122}\\ f_5&=f_{1212}+f_{2121}\\ f_6&=f_{1221}+f_{2112}\\ f_7&=f_{1222}+f_{2111} \end{aligned} \tag{13}$$

where

$$\sum f_i=1. \tag{14}$$

Consider the path between A and C in tree 1. Counting all those configurations which have the same value at A and C, we can calculate the probability p that A

and C have the same value:

$$p = f_0 + f_1 + f_4 + f_5. \tag{15}$$

Now, from (4) we also have:

$$p = 1 - s \tag{16}$$

so that from (12) and (16),

$$\begin{aligned} T_{AC} &= 2p - 1 \\ &= f_0 + f_1 + f_4 + f_5 - f_2 - f_3 - f_6 - f_7. \end{aligned} \tag{17}$$

by (14) and (15). Similarly we can calculate homogeneous linear expressions in the f for the other determinants in (6).

Using these expressions Cavender & Felsenstein calculated the three invariants in (10) and found the following quadratic expressions:

$$\begin{aligned} T_{AC}T_{BD} - T_{AD}T_{BC} &= 4[(f_4 - f_7)(f_2 - f_1) - (f_6 - f_5)(f_0 - f_3)] \\ T_{AD}T_{BC} - T_{AB}T_{CD} &= 4[(f_2 - f_7)(f_1 - f_4) - (f_3 - f_6)(f_0 - f_5)] \\ T_{AB}T_{CD} - T_{AC}T_{BD} &= 4[(f_1 - f_7)(f_4 - f_2) - (f_5 - f_3)(f_0 - f_6)]. \end{aligned} \tag{18}$$

Note that for each expression, $2(8 \times 8) = 128$ terms are produced when the two pairs of determinants are multiplied together and subtracted, but that only eight remain after simplification.

We can now calculate the three expressions in (18), using the g_{uvxy} as estimates for the f_{uvxy}, and choose the tree for which the corresponding invariant is closest to zero, preferably as validated by a statistical analysis such as the one to be discussed below.

Multivalued Characters

Working with nucleotide sequences, the case $n = 2$ is not too useful. What is needed is the case $n = 4$, since there are four "values" A, C, G and U (or T) in a nucleotide sequence.

As a first step, we study the case $n = 3$.

In this case, each transition matrix is of the form

$$\begin{vmatrix} 1-2s & s & 2 \\ s & 1-2s & s \\ s & s & 1-2s \end{vmatrix} \tag{19}$$

so that, from (4) and (5),

$$\begin{aligned} T &= (1 - 3s)^2 \\ &= (3p - 1)^2/4. \end{aligned} \tag{20}$$

Note that this is quadratic instead of linear as in the case $n = 2$. Moreover the

configuration probabilities are more numerous and complicated:

$$\begin{aligned}
f_0 &= f_{1111}+f_{2222}+f_{3333}\\
f_1 &= f_{1112}+f_{2221}+f_{2223}+f_{3332}+f_{1113}+f_{3331}\\
f_2 &= f_{1121}+f_{2212}+f_{2232}+f_{3323}+f_{1131}+f_{3313}\\
f_3 &= f_{1122}+f_{2211}+f_{2233}+f_{3322}+f_{1133}+f_{3311}\\
f_4 &= f_{1211}+f_{2122}+f_{2322}+f_{3233}+f_{1311}+f_{3133}\\
f_5 &= f_{1212}+f_{2121}+f_{2323}+f_{3232}+f_{1313}+f_{3131}\\
f_6 &= f_{1221}+f_{2112}+f_{2332}+f_{3223}+f_{1331}+f_{3113}\\
f_7 &= f_{1222}+f_{2111}+f_{2333}+f_{3222}+f_{1333}+f_{3111}\\
f_8 &= f_{1123}+f_{1132}+f_{2213}+f_{2231}+f_{3312}+f_{3321}\\
f_9 &= f_{1213}+f_{1312}+f_{2123}+f_{2321}+f_{3132}+f_{3231}\\
f_{10} &= f_{1231}+f_{1321}+f_{2132}+f_{2312}+f_{3123}+f_{3213}\\
f_{11} &= f_{2311}+f_{3211}+f_{1322}+f_{3122}+f_{1233}+f_{2133}\\
f_{12} &= f_{2131}+f_{3121}+f_{1232}+f_{3212}+f_{1323}+f_{2313}\\
f_{13} &= f_{2113}+f_{3112}+f_{1223}+f_{3221}+f_{1332}+f_{2331}.
\end{aligned} \tag{21}$$

For each determinant such as T_{AC}, if we try to express it in a way analogous to (17) in the case of $n=2$, we obtain a 14-term homogeneous linear expression squared:

$$T_{AC}=[2(f_0+f_1+f_4+f_5+f_9)-(f_2+f_3+f_6+f_7+f_8+f_{10}+f_{11}+f_{12}+f_{13})]^2/4. \tag{22}$$

This involves 105 quadratic terms. Substituting these determinants into one of the invariants in (10) leads to an expression with 22 050 terms, which cannot easily be simplified.

Alternatively, we may work with the non-homogeneous form obtained from (20) using the same kind of reasoning as we did to find p in (15):

$$T_{AC}=[3(f_0+f_1+f_4+f_5+f_9)-1]^2/4, \tag{23}$$

giving rise to only 882 terms in each of the unsimplified quartic invariants obtained by substituting in (10).

This size of expression can be handled by symbolic manipulation programs such as *MACSYMA*, which enabled us to find a factorization of each quartic invariant into two quadratic factors of more manageable size†. Finally, efforts to find structural parallels in these two factors led to the following formulations:

$$\begin{aligned}
T_{AC}T_{BD}-T_{AD}T_{BC} &= (F_2+F_3)(F_2-F_3)\\
T_{AD}T_{BC}-T_{AB}T_{CD} &= (F_3+F_1)(F_3-F_1)\\
T_{AB}T_{CD}-T_{AC}T_{BD} &= (F_1+F_2)(F_1-F_2)
\end{aligned} \tag{24}$$

†We thank David Rand for carrying out the *MACSYMA* computations.

where

$$\begin{aligned}
F_2 &= (3(f_0+f_5+f_9+f_{12})-1)(3(f_0+f_1+f_2+f_4+f_7+f_5)-1) \\
&\quad +9(f_1+f_4-f_{12})(f_2+f_7-f_9) \\
F_3 &= (3(f_0+f_6+f_{10}+f_{13})-1)(3(f_0+f_1+f_2+f_4+f_7+f_6)-1) \\
&\quad +9(f_2+f_4-f_{13})(f_1+f_7-f_{10}) \\
F_1 &= (3(f_0+f_3+f_8+f_{11})-1)(3(f_0+f_1+f_2+f_4+f_7+f_3)-1) \\
&\quad +9(f_7+f_4-f_8)(f_2+f_1-f_{11}).
\end{aligned} \tag{25}$$

For one of the quartic polynomials in (24) to be identically zero over the domain of values of f which can be generated by a given tree shape (tree 1, 2 or 3, as appropriate) means that at least one of its quadratic factors must also be identically zero over this domain. Now, in the invariant for tree 1, we can show that the minimum value for F_2+F_3 is zero, as follows. Taking the first derivatives of F_2+F_3 with respect to the f_i and setting them equal to zero produces a set of 14 linear equations in the f_i. This set of equations only has rank four and may be simplified to

$$\begin{aligned}
f_9 &= 1/3-f_0-f_1-f_4-f_5 \\
f_{10} &= 1/3-f_0-f_2-f_4-f_6 \\
f_{12} &= 1/3-f_0-f_2-f_7-f_5 \\
f_{13} &= 1/3-f_0-f_1-f_7-f_6.
\end{aligned} \tag{26}$$

The second derivatives of the quadratic function F_2+F_3 are all non-negative so that any solution of (26) must be a minimum for F_2+F_3. Substituting (26) in (25) produces $F_2=F_3=0$, so that when F_2+F_3 attains its minimum value (zero) the value of F_2-F_3 is also zero. Elsewhere F_2+F_3 is strictly positive. In other words, it is the F_2-F_3 factor which must be identically zero over the domain of f values generated by tree 1, so that this factor is a quadratic invariant for that tree. Thus

$$F_2-F_3,\ F_3-F_1,\ F_1-F_2 \tag{27}$$

are the $n=3$ analogs of the Cavender-Felsenstein invariants.

What about $n=4$? It would seem beyond the capacity of existing programs to deal with the results of expanding the determinant of the 4×4 version of (2), i.e. with invariants consisting of polynomials of order six containing millions of terms.

That there are quadratic invariants for both $n=2$ and $n=3$, however, suggests that there may be for $n=4$ as well.

We can rewrite the $n=2$ invariants in the same form as (27):

$$\begin{aligned}
T_{AC}T_{BD}-T_{AD}T_{BC} &= F_2-F_3 \\
T_{AD}T_{BC}-T_{AB}T_{CD} &= F_3-F_1 \\
T_{AB}T_{CD}-T_{AC}T_{BD} &= F_1-F_2,
\end{aligned} \tag{28}$$

but where, for example,

$$F_1 = [2(f_0+f_3)-1][2(f_0+f_1+f_2+f_4+f_7+f_3)-1]+4(f_7+f_4)(f_2+f_1). \quad (29)$$

Comparing this to the $n=3$ case, where;

$$F_1 = [3(f_0+f_3+f_8+f_{11})-1][3(f_0+f_1+f_2+f_4+f_7+f_3)-1] + 9(f_7+f_4-f_8)(f_2+f_1-f_{11}), \quad (30)$$

suggests that the $n=4$ case might involve;

$$F_1 = [4(f_0+f_3+f_8+f_{11})-1][4(f_0+f_1+f_2+f_4+f_7+f_3)-1] + 16(f_7+f_4-f_8)(f_2+f_1-f_{11}) \quad (31)$$

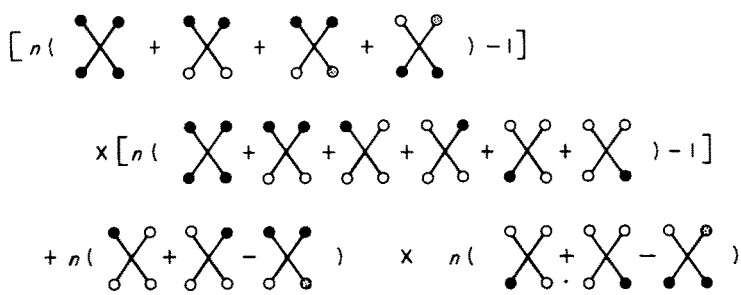
with the possible inclusion of f_{14} somewhere in the formula, where;

$$f_{14} = f_{1234}+f_{1243}+\ldots+f_{4321}, \quad (32)$$

and where the other f_i's are adjusted in obvious ways to take into account configurations containing the fourth value of the character.

In fact, the expression in (31) is correct as it stands, with f_{14} only implicitly present, in the "-1" terms [cf. (14)]. No analytic proof is available, but it has been confirmed on a large number of examples spanning the space of four-trees, so that it may be considered true for all practical purposes.

A pictorial representation of the F components of the invariants for the various n is given in (33). The large X's (indicating unresolved four-trees) represent the probabilities of the various configurations. Two nodes are shaded in the same way on a tree if they have the same value for the character. With the nodes labeled A, B, C and D as in our earlier discussion of the model, (33) represents F_2, while F_1 and F_3 can be obtained by suitably permuting the labels. The n can be 2, 3 or 4. Indeed we conjecture that this representation is valid for all larger n, as well.



(33)

Other Quadratic Invariants

Cavender & Felsenstein found a second set of three quadratic invariants. Counting all the cases in (13) where the character values are the same at A and B, all the cases where they are the same at C and D, and then just the cases where they are both the same at A and B and the same at C and D, gives $f_0+f_3+f_1+f_2$, $f_0+f_3+f_4+f_7$

and f_0+f_3, respectively, as the probabilities of these three events. Now, under the symmetric model for the transition probabilities and the probability distributions at each node, what happens on the path between A and B in tree 1 is probabilistically independent of what happens on the path between C and D. (This is not true in trees 2 and 3.) Thus, for tree 1,

$$(f_0+f_3+f_1+f_2)(f_0+f_3+f_4+f_7)=f_0+f_3, \tag{34}$$

from which, using (14), we derive

$$L_1=(f_1+f_2)(f_4+f_7)-(f_0+f_3)(f_5+f_6)=0. \tag{35}$$

Similarly for trees 2 and 3,

$$L_2=(f_1+f_4)(f_2+f_7)-(f_0+f_5)(f_3+f_6)=0 \tag{36}$$

and

$$L_3=(f_1+f_7)(f_2+f_4)-(f_0+f_6)(f_3+f_5)=0, \tag{37}$$

respectively. The tree invariants L_1, L_2 and L_3 are related to those discussed earlier by:

$$\begin{aligned} F_2-F_3&=4(L_2-L_3)\\ F_3-F_1&=4(L_3-L_1)\\ F_1-F_2&=4(L_1-L_2). \end{aligned} \tag{38}$$

so that when $L_1=0$ (and $F_2-F_3=0$), then $L_2=L_3$.

What of the case $n=3$? Here, the same arguments leading to (34) and to (35-37) result in:

$$\begin{aligned} L_1&=(f_1+f_2+f_8)(f_4+f_7+f_{11})-(f_0+f_3)(f_5+f_6+f_9+f_{10}+f_{12}+f_{13})\\ L_2&=(f_1+f_4+f_9)(f_2+f_7+f_{12})-(f_0+f_5)(f_3+f_6+f_8+f_{10}+f_{11}+f_{13})\\ L_3&=(f_1+f_7+f_{13})(f_4+f_2+f_{10})-(f_0+f_6)(f_3+f_5+f_8+f_9+f_{11}+f_{12}). \end{aligned} \tag{39}$$

In cases $n\geq 4$, we have:

$$\begin{aligned} L_1&=(f_1+f_2+f_8)(f_4+f_7+f_{11})-(f_0+f_3)(f_5+f_6+f_9+f_{10}+f_{12}+f_{13}+f_{14})\\ L_2&=(f_1+f_4+f_9)(f_2+f_7+f_{12})-(f_0+f_5)(f_3+f_6+f_8+f_{10}+f_{11}+f_{13}+f_{14})\\ L_3&=(f_1+f_7+f_{13})(f_4+f_2+f_{10})-(f_0+f_6)(f_3+f_5+f_8+f_9+f_{11}+f_{12}+f_{14}). \end{aligned} \tag{40}$$

In fact, with the definition of the f_i appropriate to the number of character values as in (13), (21) and (32), for example, the expressions in (40) are valid for all $n\geq 2$.

The relationships in (38) do not hold for all n. Rather, comparing (40) with (33) leads to the following general expressions:

$$\begin{aligned} F_2-F_3&=n^2(L_2-L_3)-n[f_{12}+f_9-(n-2)f_5+(n-2)f_6-f_{10}-f_{13}]\\ F_3-F_1&=n^2(L_3-L_1)-n[f_{13}+f_{10}-(n-2)f_6+(n-2)f_3-f_8-f_{11}]\\ F_1-F_2&=n^2(L_1-L_2)-n[f_{11}+f_8-(n-2)f_3+(n-2)f_5-f_9-f_{12}]. \end{aligned} \tag{41}$$

Recall that the invariant status of the expressions in (41) is only conjectured for $n>3$.

Statistical Considerations

Given that we must use the proportions g_{uvxy} and not the actual probabilities f_{uvxy} in calculating the invariants, the question may arise of whether one of three values is significantly closer to zero than the other two. To test hypotheses about quadratic functions of multinomial frequencies (the f_i), we must first be able to estimate the expectations and the variances of these formulae when the g_i are substituted for the f_i. Here we will review some of the pertinent facts about these quantities. Estimating an invariant such as (27), (39) or (40) as;

$$\sum\sum a_{ij}g_ig_j+\sum b_ig_i+c, \tag{42}$$

its expected value is;

$$\sum\sum a_{ij}E(g_ig_j)+\sum b_iE(g_i)+c, \tag{43}$$

and its variance is;

$$\begin{aligned}&\sum\sum a_{ij}^2 \operatorname{var}(g_ig_j)+\sum b_i^2 \operatorname{var}(g_i)+\sum\sum\sum\sum a_{ij}a_{hk} \operatorname{covar}(g_ig_j, g_hg_k)\\&\quad+\sum\sum\sum a_{ij}b_h \operatorname{covar}(g_ig_j, g_h)+\sum\sum b_ib_j \operatorname{covar}(g_i, g_j).\end{aligned} \tag{44}$$

Now, we may use the formulae;

$$\operatorname{var}(X)=E(X^2)-[E(X)]^2, \qquad \operatorname{covar}(X, Y)=E(XY)-E(X)E(Y), \tag{45}$$

together with the following properties of multinomial frequencies:

$$\begin{aligned}E(g_i)&=f_i\\E(g_i^2)&=f_i^2(N-1)/N+f_i/N\\E(g_i^3)&=f_i^3(N-1)(N-2)/N^2+3f_i^2(N-1)/N^2+f_i/N^2\\E(g_i^4)&=f_i^4(N-1)(N-2)(N-3)/N^3+6f_i^3(N-1)(N-2)/N^3\\&\quad+7f_i^2(N-1)/N^3+f_i/N^3\\E(g_ig_j)&=f_if_j(N-1)/N\\E(g_i^2g_j)&=f_i^2f_j(N-1)(N-2)/N^2+f_if_j(N-1)/N^2\\E(g_i^3g_j)&=f_i^3f_j(N-1)(N-2)(N-3)/N^3+f_i^2f_j(N-1)(N-2)/N^3\\&\quad+f_if_j(N-1)/N^3\\E(g_i^2g_j^2)&=f_i^2f_j^2(N-1)(N-2)(N-3)/N^3+[f_i^2f_j+f_if_j^2](N-1)(N-2)/N^3\\&\quad+f_if_j(N-1)/N^3\\E(g_ig_jg_k)&=f_if_jf_k(N-1)(N-2)/N^2\\E(g_i^2g_jg_k)&=f_i^2f_jf_k(N-1)(N-2)(N-3)/N^3+f_if_jf_k(N-1)(N-2)/N^3\\E(g_ig_jg_hg_k)&=f_if_jf_hf_k(N-1)(N-2)(N-3)/N^3,\end{aligned} \tag{46}$$

to find the variances of the invariant formulae when the g_i are substituted for the

f_i. Then these variances may be estimated by substituting the g_i for the f_i in (46), and approximate tests of significance can be applied to the estimated values of the invariant formulae based on the g_i.

Discussion

Are there other quadratic invariants for the model with symmetric transition probabilities and uniform probability distributions at each node, besides those in (27) and those in (40)?† What if we relax the symmetry assumptions in our model? Are there invariants of form other than linear and quadratic? Can we find invariants for five-trees, or even larger trees?‡ We propose a general approach to these and similar questions. Suppose we wish to search for an invariant which is a function of the f_{uvxy} (or f_{uvxyz}, for five-trees, etc), having a certain parametrized form $\Phi(\mathbf{f}, \mathbf{a})$, where $\mathbf{f}$ represents the set of configuration probabilities and $\mathbf{a}$ the unknown parameters of the desired invariant. Given a probabilistic model including the various matrices $\mathbf{P}$ and a probability distribution of the character values at some point in the tree (such as the root), the first step is to calculate all the f, for each of a sufficiently large set of trees $\tau_1, \ldots, \tau_M$ of a given shape (e.g. tree 1). For each of these trees τ_i we equate the desired functional form, with the parameters as unknowns and substituting in $\mathbf{f}_i$ the previously calculated values of the f, to zero:

$$\Phi(\mathbf{f}_1, \mathbf{a}) = 0$$

$$\ldots$$

$$\Phi(\mathbf{f}_M, \mathbf{a}) = 0.$$

The set of non-trivial solutions for $\mathbf{a}$ (if they exist) to this set of equations should constitute the set of all invariants for this tree shape having the desired functional form.

The method of invariants, although not as informative (as to branch lengths) as maximum likelihood inference under the same probabilistic models, has a possible computational advantage. Whereas it may require extensive calculation based on the data in order to evaluate the likelihood of a tree, invariants are precalculated formulae which may be programmed before any data is considered and which require negligible computing time to evaulate.

It should be pointed out that the validity of the quadratic invariant approach is dependent on the assumption that all characters (positions in a nucleotide or amino acid sequence) have the same rate parameter, though this may change from branch to branch. The robustness or sensitivity of the analysis to the breakdown of this assumption has not yet been investigated, and whether it may be circumvented or incorporated into the analysis is an open question. Note that there is no such

†A more general question was answered in the context of linear tree invariants for $n = 4$ by Cavender (1988).

‡Positive answers to this question and the previous one (about larger trees and higher-order invariants) have been given by Sankoff (1990).

problem in the case of linear invariants, where all that is required is that each character's rate parameter is constant. If even this assumption is seriously violated, however, there may be no advantage in postulating a continuous-time Markov model, so that parsimony may well be the most reasonable criterion for evaluating phylogenies in this context.

REFERENCES

BUNEMAN, P. (1974). A note on the metric property of trees. *J. Combinatorial Theor. (B)* **17,** 48-50.

CAVENDER, J. A. (1989). Mechanized derivation of linear invariants. *Molec. Biol. Evol.* **6,** 301-316.

CAVENDER, J. A. & FELSENSTEIN, J. (1987). Invariants of phylogenies: Simple case with discrete states. *J. Classif.* **4,** 57-71.

DOBSON, A. J. (1974). Unrooted trees for numerical taxonomy. *J. appl. Prob.* **11,** 32-42.

FELLER, W. (1957). *An Introduction to Probability Theory and Its Applications* 2nd edn. New York: John Wiley.

FELSENSTEIN, J. (1983). Inferring evolutionary trees from DNA sequences. In: *Statistical Analysis of DNA Sequences* (Weir, B. S., ed.) pp. 133-150. New York: Marcel Dekker.

JUKES, T. H. & CANTOR, C. R. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism III* (Munro, H. N., ed.) pp. 21-123. New York: Academic Press.

LAKE, J. A. (1987). A rate-invariant 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.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis.* London: Academic Press.

OLSEN, G. J. (1987). The earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harb. Symp. quant. Biol.* **LII,** 825-839.

SANKOFF, D. (1990). Designer invariants for large phylogenies. *Molec. Biol. Evol.* 7 (in press).