Counting phylogenetic invariants in some simple cases

Joseph Felsenstein

Department of Genetics SK-50

University of Washington

Seattle, Washington 98195

Running Headline: Counting Phylogenetic Invariants

ABSTRACT

An informal degrees-of-freedom argument is used to count the number of phylogenetic invariants in cases where we have 3 or 4 species and can assume a Jukes-Cantor model of base substitution with or without a molecular clock. A number of simple cases are treated and in each the number of invariants can be found. Two new classes of invariants are found: non-phylogenetic cubic invariants testing independence of evolutionary events in different lineages, and linear phylogenetic invariants which occur when there is a molecular clock. Most of the linear invariants found by Cavender (1989) turn out in the Jukes-Cantor case to be simple tests of symmetry of the substitution model, and not phylogenetic invariants.

1. Introduction

The method of phylogenetic invariants, also known as "evolutionary parsimony", has been introduced into the field of phylogenetic inference by Cavender (Cavender and Felsenstein, 1987) and by Lake (1987). The invariants are polynomial expressions (quadratic in Cavender's case, linear in Lake's) in the expected frequencies of different patterns of characters states. They are invariants if they have the same value (usually zero) for all phylogenies of a given

tree topology. I define *phylogenetic invariants* as the ones that have the same value for all phylogenies of one tree topology, but have a different value for at least one tree of a different topology. They are usually constant only in one tree topology. Non-phylogenetic invariants have the same value in all phylogenies. Testing whether phylogenetic invariants have this value is useful as a test of the tree topology.

In this paper I will count the number of invariants that exist in certain cases, those with a symmetric model of change of character state. The cases I will consider have four character states (the four nucleotides A, C, G, and T) and four species. The model of nucleotide substitution considered is that of Jukes & Cantor (1969). I will also consider similar models with two and three states.

2. Definitions and Notation

Suppose that we have four different species for which we have nucleotide sequences, and consider s sites at which they can be correctly aligned, and in which the nucleotides are available in all four sequences. We will assume that the process of evolution occurs independently at each site, though of course not independently in the four species. The likelihood of the sequences is then a product over sites. If p_{ijkl} is the probability, at a site, of observing the nucleotides i, j, k and l in the four sequences, and if b_{ij} is the base observed at site j of species i, the likelihood is:

$$L = \prod_{i=1}^{s} p_{b_{1i}b_{2i}b_{3i}b_{4i}}. \tag{1}$$

We can rearrange the terms in this product so that terms that have the same four-tuple of bases are adjacent. This leads to the alternate form:

$$L = \prod_{ijkl} p_{ijkl}^{n_{ijkl}}, \tag{2}$$

where the indices i, j, k, and l each run over the four bases in the set $\{A, C, G, T\}$

and n_{ijkl} is the number of terms in (1) that have the bases i, j, k, and l. The sum of the n_{ijkl} must be the total number of sites s.

The 4-tuple of bases $ijkl$ will be called, in agreement with Cavender (1978) a *pattern*. There are 256 possible patterns, AAAA, AAAC, ... TTTT, if we ignore (as we do) ambiguous nucleotides. n_{ijkl} is thus the observed number of occurrences of pattern $ijkl$.

Note that the probabilities p_{ijkl} are functions of the tree topology and the lengths of branches, plus whatever other parameters exist in the model of base change. I have presented here only the case of four species, but the three-species case is entirely analogous, leading to:

$$L = \prod_{ijk} p_{ijk}^{n_{ijk}}. \tag{3}$$

It should be immediately clear that the set of n_{ijkl} are sufficient statistics for estimation of the phylogeny of the species and the testing of assertions about the model of nucleotide substitution. In the three-species case the sufficient statistics are of course the n_{ijk}.

3. The model of base substitution

.

The results of this paper depend on a symmetric model of base substitution, that of Jukes & Cantor (1969). A natural question, left unanswered here, is to what extent these results would generalize to models with fewer symmetries, notably the 2-parameter model of Kimura (1980). The Jukes-Cantor model is the simplest possible symmetric model. The probabilities of base change in a

given branch of a phylogeny are given by the table:

$$\begin{array}{lcccc} & to: & A & C & G & T \\ from: & & & & & \\ & A & 1-q & q/3 & q/3 & q/3 \\ & C & q/3 & 1-q & q/3 & q/3 \\ & G & q/3 & q/3 & 1-q & q/3 \\ & T & q/3 & q/3 & q/3 & 1-q \end{array} \tag{4}$$

where q is a parameter which depends on the product of the length of the branch in time (t) and the substitution rate (u) per nucleotide in that branch:

$$q = \frac{3}{4}(1 - e^{-\frac{4}{3}ut}). \tag{5}$$

Basically q is the net probability of change in that branch, and is the same no matter what the current nucleotide. If a nucleotide changes, it has an equal probability of changing to each of the other three nucleotides. Under the Jukes-Cantor model the equilibrium distribution of nucleotides (the expected nucleotide composition) is $0.25 : 0.25 : 0.25 : 0.25$. The largest biologically reasonable value of q is $3/4$. We assume in all cases that the evolutionary process has reached equilibrium before the start of the divergence of the species.

4. A simple case: 3 species and no clock

The meaning of invariants will be clearest in the simplest cases. It is useful to start with three species, a Jukes-Cantor model, and no molecular clock. The "molecular clock" is in this context simply the assertion that the probability of base substitution (u) is constant per unit time and the same in all lineages. With three species and four nucleotides there are 64 patterns, AAA, AAC, ... TTT. It should immediately be apparent that the symmetry of the Jukes-Cantor model will leave the probability of any given set of data, as expressed by the equation (3), unchanged if we replace all A's in the data by C's and all C's by A's, or for that matter if we carry out any other permutation of the four bases.

Thus if σ is the permutation, so that σ_b is the base into which base b is changed by the permutation,

$$p_{\sigma_i \sigma_j \sigma_k} = p_{ijk} \tag{6}$$

for all bases i, j, k, and l, whether or not those are distinct. An analogous result of course holds for the four-species case.

Symmetry tests. The probability of the pattern AAC on a given tree will be the same as the probability of pattern GGA, and similarly for any pattern of the form xxy, where x and y stand for any two distinct nucleotides. This argument allows us to see that there are in fact only five types of patterns: xxx, xxy, xyx, yxx, and xyz, where x, y, and z are distinct nucleotides. We will call these classes of patterns *pattern types*. Type xxx consists of patterns AAA, GGG, CCC, and TTT. All four of these have equal expected frequencies. One test of the symmetry which the Jukes-Cantor model predicts is to test whether these four patterns are significantly unequal in frequency. This can easily be done by a chi-square test of the equality of the numbers of times the four patterns occur. The test has 3 degrees of freedom.

There are similar tests within each pattern type. Table 1 shows an accounting of the different pattern types, how many patterns each contains, and thus how many degrees of freedom are available within each pattern type for testing the symmetry of the model of base substitution. There are 64 patterns in all. The pattern frequencies thus have 63 degrees of freedom, which is equivalent to the statement that the 64-dimensional space of expected pattern frequencies is constrained to a 63-dimensional subspace, namely those that add up to 1 (they are also confined to a simplex in that subspace, namely the sets of frequencies that have no negative frequencies).

The accounting of symmetry test degrees of freedom shows that there are 59 equations, in fact linear equations, which are consequences of the symmetry. For example, $p_{AAA} = p_{CCC}$ and $p_{ACG} = p_{AGT}$ are two of them. If the expected pattern frequencies satisfy these 59 linear equations as well, they must lie in a

64 - 1 - 59 = 4-dimensional linear subspace. The 4-dimensional subspace in fact corresponds precisely to the frequencies of the 5 pattern types, less one for the fact that the total of the frequencies of pattern types is 1.

The sufficient statistics for estimating the phylogeny in this case are not the observed numbers of the 64 patterns, but the observed numbers of occurrences of the 5 pattern types. Let us represent by n_{xxx} the frequency of any one of the patterns of type xxx, and by N_{xxx} the total frequency of all patterns of type xxx, and analogously for the other patterns. Let p_{xxx} and P_{xxx} be the expected frequencies of one pattern of type xxx and the total expected frequency of all patterns of type xxx, and analogously for the other patterns.

We can note that in this model the expected frequencies of the five pattern types can be written as functions of the unknown net probabilities of change in the three branches of the unrooted tree, which we will call q_1, q_2, and q_3:

$$P_{xxx} = (1-q_1)(1-q_2)(1-q_3) + q_1q_2q_3/9 \tag{7}$$

$$P_{xxy} = (1-q_1)(1-q_2)q_3 + 1/3\ q_1q_2(1-q_3) + 2/9\ q_1q_2q_3 \tag{8}$$

$$P_{xyx} = (1-q_1)q_2(1-q_3) + 1/3\ q_1(1-q_2)q_3 + 2/9\ q_1q_2q_3 \tag{9}$$

$$P_{yxx} = q_1(1-q_2)(1-q_3) + 1/3\ (1-q_1)q_2q_3 + 2/9\ q_1q_2q_3 \tag{10}$$

$$P_{xyz} = 2/3\ q_1q_2(1-q_3) + 2/3\ q_1(1-q_2)q_3 + 2/3\ (1-q_1)q_2q_3 + 2/9\ q_1q_2q_3 \tag{11}$$

The expected frequencies of the five pattern types add up to 1. Thus we have four equations in three unknowns. This implies that there is one algebraic relationship between the five quantities (although it does not prove it – see section 10 below).

After some tedious algebra (for which see Appendix 1), one can find this relationship, a cubic polynomial equation:

$$\begin{aligned} &3\ (2P_{xxx} - 2P_{xxy} - 2P_{xyx} - 2P_{yxx} + 1)^2 \\ &-[4(P_{xxx} + P_{xxy}) - 1][4(P_{xxx} + P_{xyx}) - 1][4(P_{xxx} + P_{yxx}) - 1] \quad = \quad 0. \end{aligned} \tag{12}$$

I suspect that this cubic equation is a consequence of the independence of substitution in different lineages.

With three species, the Jukes-Cantor model, the expressions (7)-(11) do not depend on where the root of the tree is, and there is only one possible unrooted tree topology. The cubic polynomial is an invariant for this case, in that it is an expression in the expected pattern type frequencies which has the same value for all trees of a given topology, whatever their branch lengths. In fact, since the placement of the root of the tree does not affect the pattern frequencies, and since there is only one possible unrooted tree topology, this cubic invariant is not a phylogenetic invariant. We have thus accounted for all the degrees of freedom in this case, and find invariants, but of course no phylogenetic invariants.

5. The Jukes-Cantor model with 3 species and a clock.

The first signs of phylogenetic invariants occur when we impose the constraint of a molecular clock. If the tree topology is the second tree in Fig. 1, this corresponds to requiring that $q_1 = q_2$. and that $q_3 > q_1$. There are still 64 patterns and 63 degrees of freedom, and still 59 of these which are accounted for by the symmetries of the Jukes-Cantor model. Of the 4 remaining degrees of freedom, 2 of them are accounted for by the branch lengths of the tree. With a molecular clock, the constraint that $q_1 = q_2$ implies that there are only two independent branch lengths (q_1 and q_3).

One degree of freedom is still accounted for by the cubic equation (12), which still holds in this case, as this case of a clock is a subcase of the preceding one. This leaves us with one degree of freedom unaccounted for. We have 5 expected pattern frequencies which must sum to 1, and two parameters, implying that two equations can be written in the expected pattern type frequencies. One of these is (12). The other is not hard to find. The equality of the branch lengths q_1 and q_2 implies that patterns xyx and yxx are expected to be equally frequent. Examination of equations (8) and (9) verifies that if $q_1 = q_2$ then $P_{xyx} = P_{yxx}$. This is the missing invariant. It is not simply a consequence of the symmetry

of the model of base substitution or of the independence of substitutions in different branches of the tree, for it will not hold in models that have both of these, but lack a clock. It will also exist in models that lack the symmetry and independence assumptions but have a clock. This invariant is a phylogenetic invariant, as in the third tree in Fig. 1 it is no longer true that $P_{xyx} = P_{yxx}$, but now instead $P_{xyx} = P_{xxy}$. We have thus identified a phylogenetic invariant, and a linear one at that. Strictly speaking, the invariants are the expressions

$$C_1 = P_{yxx} - P_{xyx}, \tag{13}$$

$$C_2 = P_{xyx} - P_{xxy}, \tag{14}$$

and

$$C_3 = P_{xxy} - P_{yxx}. \tag{15}$$

Note that $C_1 + C_2 + C_3 = 0$. For the second tree shown in Fig. 1, $C_1 = 0$, and we must then have $C_2 + C_3 = 0$. For each of the other two bifurcating tree topologies, there are similar relationships with $C_2 = 0$ or $C_3 = 0$.

6. Jukes-Cantor model with 4 species and no clock.

When we reach 4 species things become more complicated. There are 256 patterns of nucleotides. The situation is shown in Table 2. Taking into account the exchangeability of the four nucleotides, there are now 15 pattern types. This means that after the symmetry invariants have been taken out, there must remain 15 degrees of freedom, so that there are fully 241 degrees of freedom accounted for by symmetry of the nucleotides. These 15 degrees of freedom can be reduced by one since the expected pattern frequencies must add to 1. There is, in an unrooted 4-species tree, one parameter per branch and there are five branches. This leaves 15-1-5 = 9 degrees of freedom. Fortunately, we can make use at this point of invariants found by Cavender (Cavender & Felsenstein, 1987), Lake (1987), and Drolet & Sankoff (1990) to account for some of these.

a. Cavender's Invariants

Cavender investigated a model of two states, 0 and 1, and found for each of the three possible unrooted tree topologies with four species and no clock that there were two quadratic expressions in the expected frequencies that were phylogenetic invariants. In the present case we can group the four bases into two groups of two in any way (it does not matter how because of the symmetry of the bases). We may, for example, code bases into R and Y (purine and pyrimidine). R and Y will evolve as two symmetric states, the model Cavender considered. We can then classify the 256 site patterns into sixteen classes: RRRR, RRRY, ... YYYY.

The symmetry between the Y and R symbols reduces these further to eight: 0000, 0001, 0010, 0011, 0100, 0101, 0110, and 0111, where 0 and 1 are place holders of which one stands for an R and the other a Y. The expected frequencies of these eight classes of patterns must satisfy Cavender's two phylogenetic invariants (his K and L invariants), as the evolution of states R and Y follows his assumptions exactly. We shall here call the frequencies of these eight pattern types $S_{0000}, S_{0001}, S_{0010}, S_{0011}, S_{0100}, S_{0101}, S_{0110}$, and S_{0111}, as they aggregate the patterns into types in a way different from the classes whose frequencies are indicated above by the P's.

For the first tree topology in Fig. 2, Cavender's K-invariant is

$$K_1 = (S_{0100} - S_{0111})(S_{0010} - S_{0001}) - (S_{0110} - S_{0101})(S_{0000} - S_{0011}) \quad (16)$$

and his L-invariant is:

$$L_1 = (S_{0001} + S_{0010})(S_{0100} + S_{0111}) - (S_{0000} + S_{0011})(S_{0101} + S_{0110}) \quad (17)$$

Both of these invariants are zero, and both are phylogenetic invariants. It is worth noting that the Cavender K invariant can be considered to be a consequence of the four-point metric condition of Buneman (1974). Buneman pointed out that for a tree of this topology if there is a distance d_{ij} that is additive along

branches of the first tree in Fig. 2, it must satisfy

$$d_{14} + d_{23} = d_{13} + d_{24}. \tag{18}$$

In the derivation of Cavender's result we note that the branch lengths are additive, and that a branch length t may be expressed in terms of the probability D that the states of the species at the two ends of the branches are different,

$$t = -\frac{1}{2}\, ln\, (1 - 2D). \tag{19}$$

If D_{ij} is the probability that species i differs in state from species j, in the two-state case

$$D_{14} = S_{0001} + S_{0011} + S_{0101} + S_{0111}, \tag{20}$$

$$D_{23} = S_{0010} + S_{0011} + S_{0100} + S_{0101}, \tag{21}$$

$$D_{13} = S_{0010} + S_{0011} + S_{0110} + S_{0111} \tag{22}$$

and

$$D_{24} = S_{0001} + S_{0011} + S_{0100} + S_{0110}. \tag{23}$$

We can express the D_{ij} in terms of the S's in this way, and use these and equation (19) to express the total branch lengths between species i and j in terms of the S's. Since these total branch lengths can be used in place of the d_{ij} to satisfy Buneman's condition, we end up with an expression in the S's. This turns out to be precisely Cavender's K invariant.

One might imagine that another classification of the four bases into two sets of two bases each would lead to a different pair of invariants based on Cavender's invariants. Such different invariants will exist, but they can be derived from (16) and (17) using the symmetry conditions, and so they provide no extra information about tree shape and count for nothing in the accounting of degrees of freedom.

b. Lake's Invariants

Lake (1987) found two linear invariants in a model of base change which had balanced transversions (so that if an A changed, it was equally likely to change to a C or a T). The Jukes-Cantor model has this property, among others. Thus Lake's two linear invariants must also apply to the present model.

In the present notation, Lake's invariants are

$$\frac{2}{3}P_{xyxy} + \frac{1}{3}P_{xyzw} - \frac{1}{3}P_{xyxz} - \frac{1}{3}P_{xyzx} = 0 \tag{24}$$

and

$$\frac{2}{3}P_{xyyx} + \frac{1}{3}P_{xyzw} - \frac{1}{3}P_{xyzx} - \frac{1}{3}P_{yxxz} = 0. \tag{25}$$

Taking Cavender's and Lake's invariants into account reduces the 9 degrees of freedom by 4 so that we have 5 degrees of freedom to account for.

c. Three-species cubic invariants

In the three-species Jukes-Cantor case we found one cubic invariant. In the present case we can always consider three of the four species and ignore the remaining one. The 256 patterns then reduce to 64 in the obvious way: if the first species is being ignored, the pattern AAA refers to any pattern which has A in all of the last three species. Its expected frequency is the sum of the frequencies of AAAA, CAAA, GAAA and TAAA. The expected frequencies of these 64 classes will satisfy the cubic polynomial (12). There are four different ways in which we can drop one of the species (one for each species we could drop), hence four such cubic invariants. None of them is a phylogenetic invariant. It will not take readers long to satisfy themselves that these four quantities are independent. Each depends on a different three-species marginal distribution, and none of those distributions can be computed from each other. We have thus reduced the number of degrees of freedom from 5 to 1.

d. Drolet-Sankoff quadratic invariant

That one remaining degree of freedom is the four-state quadratic invariant discovered by Drolet & Sankoff (1989). They investigated the case of four species, without a clock, and with the a symmetric model of change among n states. The Jukes-Cantor model is the $n = 4$ case of the one they consider. The quantity they found is a phylogenetic invariant which is quadratic. This too must be satisfied by the expected frequencies in the present case. Drolet and Sankoff's first quadratic phylogenetic invariant is:

$$F_2 - F_3, \tag{26}$$

where

$$\begin{aligned} F_2 \quad = \quad & [4(P_{xxxx} + P_{xyxy} + P_{xyxz} + P_{xyzy}) - 1]\times \\ & [4(P_{xxxx} + P_{xyxy} + P_{xyyy} + P_{xxyx} + P_{xyxx} + P_{xxxy}) - 1] \\ & + 4(P_{xyyy} + P_{xxyx} - P_{xyxz})4(P_{xyxx} + P_{xxxy} - P_{xyzy}) \end{aligned} \tag{27}$$

and

$$\begin{aligned} F_3 \quad = \quad & [4(P_{xxxx} + P_{xyyx} + P_{xyzx} + P_{xyyz}) - 1]\times \\ & [4(P_{xxxx} + P_{xyyx} + P_{xyyy} + P_{xxxy} + P_{xxyx} + P_{xyxx}) - 1] \\ & + 4(P_{xyyy} + P_{xxxy} - P_{xyzx})4(P_{xxyx} + P_{xyxx} - P_{xyyz}) \end{aligned} \tag{28}$$

for which the invariant is zero. They also found another quadratic invariant. One might think that this is one too many. Actually, it implies the L invariant of Cavender. In the case of the Jukes-Cantor model it can be shown (Appendix 2) that when the Drolet-Sankoff L invariant has the value it is predicted to, and when the symmetry invariants do also, that the Cavender L invariant must also have its predicted value.

It is interesting and important to note that we have now completely accounted for the degrees of freedom:

5	branch length parameters
1	Drolet-Sankoff quadratic phylogenetic invariant
2	Lake linear phylogenetic invariants
2	Cavender 2-state quadratic phylogenetic invariants
4	three-species cubic invariants
241	linear invariants testing symmetry of base substitution
1	since the expected frequencies add to 1
256	

Note that only 5 of these 256 degrees of freedom are phylogenetic invariants.

7. Jukes-Cantor model with four species and a molecular clock.

When we constrain the preceding case so that the tree is clocklike, the picture changes slightly. The 5 branch lengths are replaced by 3 divergence times. All the other invariants continue to be zero, as this case is a subcase of the preceding one. Thus we have 2 degrees of freedom unaccounted for. These must test the clockness of the tree,

Fig. 3 shows the two forms of possible clocklike unlabelled tree topologies. The 15 possible bifurcating tree topologies are all of one or the other of these two kinds. The enumeration of degrees of freedom is the same as before except that the five degrees of freedom for branch lengths are replaced by 3 for branch lengths and 2 for the phylogenetic invariants for clockness.

3	branch length parameters
2	linear phylogenetic invariants for clockness
1	Drolet-Sankoff quadratic phylogenetic invariant
2	Lake linear phylogenetic invariants
2	Cavender 2-state quadratic phylogenetic invariants
4	three-species cubic invariants
241	linear invariants testing symmetry of base substitution
1	since the expected frequencies add to 1
256	

(and a partridge in a pear tree).

There is one surprise in the clock case. For the first tree topology in Fig. 3, consideration of the symmetries will immediately suggest that the following are invariants:

$$P_{xyxx} = P_{yxxx}, \tag{29}$$

$$P_{xyxy} = P_{xyyx}, \tag{30}$$

$$P_{xyxz} = P_{xyyz}, \tag{31}$$

and

$$P_{xyzx} = P_{xyzy}. \tag{32}$$

The problem is that there are too many of them. We are supposed to have 2 degrees of freedom to test clockness, not 4. The dilemma could be resolved if some of these invariants were not independent, if they were implied by combinations of others. In fact, this is the case. We can use (29)-(32) to show straightforwardly that when these hold, Lake's two linear invariants (24) and (25) are equal. We can also show that Cavender's K invariant must equal 0, and we can also show that the three species cubic invariant for species 1, 3, and 4 necessarily equals that for species 2, 3, and 4. These conclusions are explained in Appendix 3. Therefore equations (29)-(32) represent only one new

clock invariant, not four. This is one too few clock invariants. In Appendix 3 it is demonstrated that there is one more clock invariant:

$$2P_{xxyx}+P_{xyxy}+P_{xyyx}+P_{yxzx}+P_{xyzx}-P_{xyxx}-P_{yxxx}-2P_{xxyy}-2P_{yzxx}=0 \quad (33)$$

which is written more simply in another form in that Appendix.

For the second tree topology in Fig. 3, the same approach immediately suggests 6 invariants: the ones in (29)-(32) plus two more:

$$P_{xxxy} = P_{xxyx} \quad (34)$$

and

$$P_{xyxz} = P_{xyzx} \quad (35)$$

These can also be used to prove the equivalence of the two Lake linear invariants, and to prove that the Cavender K invariant is zero. They also prove the equivalence of two pairs of three-species cubic invariants, the one mentioned above plus the invariants for species 1, 2 and 3 and for 1, 2, and 4. This leaves us with two clock invariants. Equation (33) is not an invariant for this tree topology.

8. Cavender's multiple linear invariants

Cavender (1989) has found all linear invariants in a four-species case far more general than the present model. The Jukes-Cantor 4-species case presented here is a special case of the model he considers. The present calculations shed some light on his invariants. Without an evolutionary clock we have 243 linear invariants. 241 of them are symmetry tests, and the two of those that are phylogenetic invariants are the Lake invariants. Cavender finds 68 linear invariants. In the present case most of these correspond to the symmetry tests. In the Jukes-Cantor case they provide no information about the phylogeny. The pattern frequencies are continuous functions of the parameters of Cavender's model. We can invoke continuity to argue that when Cavender's model is near

the Jukes-Cantor model, that most of his linear invariants will have very little information on the phylogeny. Most of the phylogenetic information expressed in the linear invariants will thus be in the Lake invariants, except possibly when the model is far from the Jukes-Cantor assumptions.

9. Properties of invariants in different cases

It is useful to tabulate a number of properties of the invariants. We have seen that some invariants (such as Cavender's K and L) are present in models that have two states, while others (such as Lake's linear invariants) are present only when there are 4 or more states. In the table below we call this the *state level*. Invariants also differ according to how many species must be present in the tree before they exist. The cubic invariants discussed above are present whenever there are 3 or more species, but the others all require 4 species. We call this number the *species level*. The invariants are all polynomials in the expected pattern frequencies; they differ according to the degree of the polynomial, which we indicate by *degree*. Some are phylogenetic, some not. Finally, they differ in one more way. If we consider the patterns as being in a 4-way table, we can compute the various marginal sums of this table. Some invariants can be computed using only these marginal sums. For example Cavender's K can be computed using only two-species marginals, but Lake's linear invariants cannot be computed from marginals. We call this the *interaction level* of the invariant. It is also true that different n-species marginals cannot be computed from each other. Thus the four 3-species cubic invariants in a 4-species case are independent.

Here is a table of these properties, for the 4-species case with a clock:

Invariant	State level	Species level	Degree	Phylogenetic	Interaction level
Symmetry	2	2	1	no	4
Clock	2	3	1	yes	2
Cubic	3	3	3	no	3
Lake	3	3	1	yes	4
Cavender K	2	4	2	yes	2
Cavender L	2	4	2	yes	4
Drolet-Sankoff	3	4	2	yes	4

The table gives an incomplete picture. In some cases we have denoted an invariant as present for a given number of species or a given number of states even though the number of that class of invariants rises as the number of states or species rises. Thus Cavender's K invariant, a single invariant, is present whenever there are two or more species. By contrast, one invariant related to Lake's linear invariants is present when there are 3 states and 4 species, but when the number of states increases to 4 there are then two Lake invariants. The three-state Lake-like invariant is

$$P_{xyxy} + P_{xyyx} - P_{yxzx} - P_{yxxz} = 0 \tag{36}$$

as can be verified by exact calculation of the probabilities of these four pattern types as functions of the branch lengths.

Table 3 gives an accounting of invariants with different numbers of species and different numbers of states with a clock. The corresponding table without the clock is the same, except that the degrees of freedom for the clock must be added to those for branch lengths, so that entries with 3 degrees of freedom for branch lengths and 2 for clock invariants have instead 5 degrees of freedom for branch lengths.

It is worth noting that one property, the interaction level, has implications for independence of the invariants. The invariants that have one interaction

level cannot depend on those that have another. It is a well-known fact that in a multi-way table (in the nucleic acid sequence case, a $4 \times 4 \times 4 \times 4$ table, for example), that an n-species marginal distribution cannot be computed from the $(n-1)$-species marginal distributions. Thus if one invariant is computed from 4-species pattern frequencies, for example, it cannot depend on others that are computed from 2- or 3-species pattern frequencies.

10. Accounting for all degrees of freedom

The present study accounts for all of the degrees of freedom in the Jukes-Cantor cases with 2, 3, or 4 species and 2, 3, or 4 states. In the absence of a molecular clock, no new phylogenetic invariants have been found. In the presence of the clock the linear clock phylogenetic invariants have been found.

However, there is a major limitation of these results. In cases like the present one, where we consider nonlinear functions of the expected pattern frequencies, we must be cautious about the notion of degrees of freedom. We have, for example, 256 pattern frequencies predicted by 5 branch lengths, but it is not immediately obvious that this means that there are 251 equations in the pattern frequencies if they are the ones predicted by the model. This is because the notion of degrees of freedom applies only to linear equations, and the present equations include quadratics and cubics. It is possible that there are more invariants to be found. I suspect not, but cannot prove this. The informal methods used here do not prove that the invariants shown here are all independent. A reviewer has pointed out that what is needed is to prove that the polynomials found here form a basis of space of the 256 pattern frequencies. This has not been done here. We cannot rule out the possibility that there are more to be found, or that some of these are redundant.

It would be of interest to have an analysis similar to the present one for the case of Kimura's (1980) two-parameter model. While the inequality of transition and transversion rates makes few realistic models of nucleotide substitution close to the Jukes-Cantor model, more might be close to Kimura's model, which

allows for this inequality. The conclusions from the Kimura model as to which invariants contain the information about the phylogenies might thus be much closer to being correct in more realistic models.

11. Likelihood, parsimony, and invariants

If we were to test all invariants at once for having their desired values, this would amount to a test of whether the observed pattern frequencies were in the low-dimensional subspace defined by varying all branch length parameters and for each computing the expected pattern frequencies. For example, for the 4-species case without a clock, the subspace is 5-dimensional, as there are then 5 branch lengths. We have not presented such a test; its full elaboration is a matter for future work. However a straightforward approach would be to take the likelihood ratio between the best fitting arbitrary expected frequencies, which will be the same as the observed frequencies, and the best fitting expected frequencies from a phylogeny. For any one tree topology twice the log of the likelihood ratio should be distributed as χ^2 with $255 - 5 = 250$ degrees of freedom. This is an asymptotic distribution, valid as the number of sites becomes large.

One difficulty with this neat picture is that we are not simply finding the best-fitting point in the 5-dimensional subspace defined by one tree topology, but are picking the best tree from three different tree topologies. It is not clear whether there is some way to correct for this. If the three tests were statistically independent we could do so by a Bonferroni correction for multiple tests.

A more serious issue is how to compare different tree topologies, when we are willing to assume that one or another of them provides a correct model for the data. We cannot do a simple likelihood ratio test, as the hypotheses are not nested one within another. It is in this case that the strengths of the invariants approach are clearest. Many of the invariants, we have seen, test the symmetries of the model. If we assume that symmetry, we can focus our test on the phylogenetic invariants, and will not lose power by wasting effort testing

those symmetries. The different invariants test somewhat different aspects of the model, and this allows us to have a clearer idea what is being accepted and what rejected. The linear invariants are readily tested (Lake, 1987), and the Cavender L quadratic invariant is also (Cavender & Felsenstein, 1987). Drolet & Sankoff (1990) have given expressions for the variances of the other quadratic invariants and pointed out their asymptotic normality.

We do not yet have a complete picture of the statistical testing of invariants. What is clear is that they provide a more precise picture of the different kinds of evidence our data provides about tree topologies, branch lengths, and departures from the model. Although some invariants can be related to parsimony (Lake, 1987), they seem to me much more naturally related to likelihood methods, providing as they do an anatomical structure of the implications of the data.

Acknowledgments

I am indebted to James Cavender and David Sankoff for discussions of some of these issues and to James Cavender and the referees for comments on an earlier version of the manuscript. This research was supported by NSF grants BSR-8614807 and BSR-8918333 and by NIH grant 5-R01 GM41716.

REFERENCES

Buneman, P. 1974. The recovery of trees from measurements of dissimilarity. pp. 387-395 in *Mathematics in the Archaeological and Historical Sciences*, ed. F. R. Hodson, D. G. Kendall, and P. Tautu. Edinburgh University Press, Edinburgh.

Cavender, J. A. 1978. Taxonomy with confidence. *Mathematical Biosciences* **40**: 271-280.

Cavender, J. A. and J. Felsenstein. 1987. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification* **4**: 57-71.

Cavender, J. A. 1989. Mechanized derivation of linear invariants. *Molecular Biology and Evolution* **6**: 301-316.

Drolet, S. and D. Sankoff. 1990. Quadratic tree invariants for multivalued characters. *Journal of Theoretical Biology* **144**: 117-129.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. pp. 21-123 in *Mammalian Protein Metabolism III*, ed. H. N. Munro. Academic Press, New York.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**: 111-120.

Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Molecular Biology and Evolution 4: 167-191.

Table 1

The pattern types with a Jukes-Cantor model with 3 species and no molecular clock.

Pattern type	Patterns	Symmetry d.f.
xxx	4	3
xxy	12	11
xyx	12	11
yxx	12	11
xyz	24	23
	—	—
Total	64	59

Table 2

The pattern types with a Jukes-Cantor model with 4 species and no molecular clock.

Pattern type	Patterns	Symmetry d.f.
xxxx	4	3
xxxy	12	11
xxyx	12	11
xyxx	12	11
yxxx	12	11
xxyy	12	11
xyxy	12	11
xyyx	12	11
xxyz	24	23
xyxz	24	23
xyzx	24	23
yxxz	24	23
yxzx	24	23
yzxx	24	23
xyzw	24	23
	—	—
	256	241

Table 3

Non-symmetry invariants with different numbers of states and of species in the case of a molecular clock. The corresponding table for no clock is the same except that the degrees of freedom for clock invariants instead become degrees of freedom for more branch lengths.

Species	States		
	2	3	4
2	2 classes = 1 for sum + 1 branch length	(same as 2 states)	(same as 2 states)
3	4 classes = 1 for sum + 2 branch lengths + 1 clock	5 classes = 1 for sum + 2 branch lengths + 1 clock + 1 cubic	(same as 3 states)
4	8 classes = 1 for sum + 3 branch lengths + 2 clock + 2 Cavender K and L	14 classes = 1 for sum + 3 branch lengths + 2 clock + 2 Cavender K and L + 4 cubic + 1 Lake-like + 1 Drolet-Sankoff	15 classes = 1 for sum + 3 branch lengths + 2 clock + 2 Cavender K and L + 4 cubic + 2 Lake + 1 Drolet-Sankoff

FIGURE CAPTION

Figure 1. The three-species unrooted tree with the branch lengths shown, and the three possible kinds of rooted bifurcating trees showing a molecular clock.

Figure 2. The three different unrooted four-species bifurcating trees.

Figure 3. The two different shapes of rooted bifurcating trees with 4 species showing a molecular clock. There are 15 bifurcating tree topologies in all, each of one or the other of these forms.

Appendix 1

The Three-Species Cubic Invariant

If we let

$$q_i = \frac{3}{4}(1 - f_i), \tag{37}$$

then

$$1 - q_i = \frac{1}{4} + \frac{3}{4} f_i. \tag{38}$$

These are in effect the substitution in equation (5) of

$$f = e^{-4/3ut} \tag{39}$$

Substituting these for the q_i in equations (7)-(11) we get the equations

$$P_{xxx} = \frac{1}{16} + \frac{3}{16} f_1 f_2 + \frac{3}{16} f_2 f_3 + \frac{3}{16} f_1 f_3 + \frac{3}{8} f_1 f_2 f_3, \tag{40}$$

$$P_{xxy} = \frac{3}{16} + \frac{9}{16} f_1 f_2 - \frac{3}{16} f_2 f_3 - \frac{3}{16} f_1 f_3 - \frac{3}{8} f_1 f_2 f_3, \tag{41}$$

$$P_{xyx} = \frac{3}{16} - \frac{3}{16} f_1 f_2 - \frac{3}{16} f_2 f_3 + \frac{9}{16} f_1 f_3 - \frac{3}{8} f_1 f_2 f_3, \tag{42}$$

$$P_{yxx} = \frac{3}{16} - \frac{3}{16} f_1 f_2 + \frac{9}{16} f_2 f_3 - \frac{3}{16} f_1 f_3 - \frac{3}{8} f_1 f_2 f_3, \tag{43}$$

and

$$P_{xyz} = \frac{6}{16} - \frac{6}{16} f_1 f_2 - \frac{6}{16} f_2 f_3 - \frac{6}{16} f_1 f_3 + \frac{6}{8} f_1 f_2 f_3. \tag{44}$$

Note that there are no linear terms in the f_i in these expressions. Adding the first two of these equations

$$P_{xxx} + P_{xxy} = \frac{1}{4} + \frac{3}{4} f_1 f_2, \tag{45}$$

from which

$$f_1 f_2 = [4(P_{xxx} + P_{xxy}) - 1]/3, \tag{46}$$

and in analogous fashion from the first and third equations,

$$f_1 f_3 = [4(P_{xxx} + P_{xyx}) - 1]/3 \tag{47}$$

and from the first and fourth,

$$f_2 f_3 = [4(P_{xxx} + P_{yxx}) - 1]/3. \tag{48}$$

Substituting these into the last of the equations, we can eliminate all the terms $f_1 f_2$, $f_1 f_3$, and $f_2 f_3$, leaving only $f_1 f_2 f_3$ for which we then can solve:

$$f_1 f_2 f_3 = \frac{2}{3}(P_{xxx} - P_{xxy} - P_{xyx} - P_{yxx} + \frac{1}{2}). \tag{49}$$

Squaring this equation and comparing it to the product of equations (46), (47), and (48), we find that both have $f_1^2 f_2^2 f_3^2$ on the left-hand side, and therefore we can equate the right-hand sides and get equation (12).

Appendix 2

Equivalence of Drolet-Sankoff L Invariant and other invariants

Drolet and Sankoff's second quadratic phylogenetic invariant is

$$L_1 = Q_1 Q_2 - Q_3 Q_4 = 0, \tag{50}$$

where

$$Q_1 = P_{xxxx} + P_{xxyy}, \tag{51}$$

$$Q_2 = P_{xxxy} + P_{xxyx} + P_{xxyz}, \tag{52}$$

$$Q_3 = P_{xyxx} + P_{yxxx} + P_{yzxx}, \tag{53}$$

and

$$Q_4 = P_{xyxy} + P_{xyyx} + P_{xyxz} + P_{xyzx} + P_{yxzx} + P_{yxxz} + P_{xyzw}. \tag{54}$$

We will see that this is not a separate invariant but implies the Cavender L invariant in the Jukes-Cantor case, given the symmetries of that model, and the Lake invariants. The Jukes-Cantor model was the one Drolet and Sankoff (1990) were considering.

An alternative form of (50) is obtained by noting that since each of the 15 P's occur in exactly one of equations (51) through (54),

$$Q_1 = 1 - Q_2 - Q_3 - Q_4, \tag{55}$$

and substituting this into (50) gives

$$Q_4 = (Q_2 + Q_4)(Q_3 + Q_4). \tag{56}$$

We know that in a pattern type like xxyz each of the symbols stands for a distinct nucleotide. There are then $4 \times 3 \times 2 = 24$ possible assignments of nucleotides to the symbols x, y, and z. In the Jukes-Cantor case the frequencies of all of these patterns must be equal. Of these 24 equally frequent patterns, 1/3 have both x and y both pyrimidines or both purines. Making a similar argument for the other pattern types, we find that on classifying the bases into Y and R and those into 0 and 1, as done above in the discussion of Cavender's K invariant,

$$\begin{aligned} U_1 &= S_{0000} + S_{0011} \\ &= P_{xxxx} + P_{xxyy} + \tfrac{1}{3}(P_{xxxy} + P_{xxyx} \\ &\quad + P_{xxyz} + P_{xyxx} + P_{yxxx} + P_{yzxx} \\ &\quad + P_{xyxy} + P_{xyyx} + P_{xyzw}), \end{aligned} \tag{57}$$

$$\begin{aligned} U_2 &= S_{0001} + S_{0010} \\ &= \tfrac{2}{3}(P_{xxxy} + P_{xxyx} + P_{xxyz}) \\ &\quad + \tfrac{1}{3}(P_{xyxz} + P_{yxxz} + P_{yxzx} + P_{xyzx}), \end{aligned} \tag{58}$$

$$\begin{aligned} U_3 &= S_{0100} + S_{0111} \\ &= \tfrac{2}{3}(P_{xyxx} + P_{yxxx} + P_{yzxx}) \\ &\quad + \tfrac{1}{3}(P_{xyxz} + P_{yxxz} + P_{yxzx} + P_{xyzx}), \end{aligned} \tag{59}$$

and

$$\begin{aligned} U_4 &= S_{0101} + S_{0110} \\ &= \tfrac{1}{3}(P_{xyxz} + P_{yxxz} + P_{yxzx} + P_{xyzx}) \\ &\quad + \tfrac{2}{3}(P_{xyxy} + P_{xyyx} + P_{xyzw}). \end{aligned} \tag{60}$$

so that the Cavender L-invariant is

$$U_1U_4 - U_2U_3 = 0, \tag{61}$$

which also can be written as

$$.U_4 = (U_2 + U_4)(U_3 + U_4) \tag{62}$$

Adding equations (58) and (60) gives

$$U_2 + U_4 = \frac{2}{3}(Q_2 + Q_4), \tag{63}$$

and adding equations (59) and (60) gives

$$U_3 + U_4 = \frac{2}{3}(Q_3 + Q_4). \tag{64}$$

If we could show that

$$U_4 = \frac{4}{9}Q_4, \tag{65}$$

we would have completed the demonstration that the Drolet-Sankoff L invariant is is not independent of the Cavender L invariant. This can be shown, but requires use of the Lake invariants, equations (24) and (25). Adding those two equations shows that the two terms in (60) satisfy:

$$P_{xyxz} + P_{yxxz} + P_{yxzx} + P_{xyzx} = 2(P_{xyxy} + P_{xyyx} + P_{xyzw}). \tag{66}$$

If we define

$$V = P_{xyxy} + P_{xyyx} + P_{xyzw}, \tag{67}$$

then (66) and (60) show that

$$U_4 = \frac{1}{3}(2V) + \frac{2}{3}V = \frac{4}{3}V. \tag{68}$$

Since (54) shows that

$$Q_4 = 3V, \tag{69}$$

this, together with (68) immediately establishes (65). Thus the Drolet- Sankoff L invariant is not autonomous but is a consequence of the Lake linear invariants, Cavender's L invariant, and the symmetry invariants.

Appendix 3

Equivalence of some clock invariants and other invariants

For the two tree shapes in Fig. 3, in the four-state case respectively 5 and 6 clock invariants are given in section 7. This Appendix derives one of them, and also shows how all but two of them are equivalent to other invariants.

Equivalence of the two Lake invariants. Examining equations (24) and (25) and comparing them termwise, equations (30)-(32) can be used to turn equation (24) into (25), proving that one implies the other.

Equivalence of cubic invariants. Equations (29)-(32) were obtained by noting that any two patterns that can be obtained from each other by transposing the bases in species 1 and 2 must have the same expected frequency. This also immediately establishes that the three-species marginal pattern frequencies for species 1, 3, and 4 must be the same as those for species 2, 3, and 4. The cubic invariants for these two triples of species are the same functions of those three-species marginal pattern frequencies, and hence must be equal. For the second tree topology in Fig. 3, the principles are the same, except that species 3 and 4 may also be transposed (with or without transposing species 1 and 2) without changing the pattern frequency. The cubic invariant for species 1, 2, and 3 must then equal that for species 1, 2, and 4.

Clock invariants imply the Cavender K invariant. If the D_{ij} are the two-state distances between species i and j (the probabilities that species i is in a different one of the two states than species j), section 6a above mentions how it may be shown that Cavender's K invariant is

$$K = (D_{14} + D_{23} - D_{24} - D_{13}) - 2(D_{14}D_{23} - D_{24}D_{13}). \tag{70}$$

From the equivalence of patterns that have the first two species transposed, and using equations (20) and (23), we can show that $S_{0101} = S_{0110}$ and that $S_{0100} = S_{0111}$, which leads immediately to

$$D_{14} = D_{24}. \tag{71}$$

In similar fashion it can be shown using (21) and (22) that

$$D_{13} = D_{23}. \tag{72}$$

When these are substituted into equation (70) they immediately establish that Cavender's K is 0. Since equations (71) and (72) hold for both of the trees in Fig. 3, they establish that $K = 0$ for both of them.

Proof of equation (33). For the first tree topology in Fig. 3, in the four-species case, we can imagine dropping species 1 or species 2 from the tree. For the two remaining three-species trees there must be a clock invariant analogous to equation (13). The three-species pattern types can obviously be written in terms of the four-species pattern types. For example for the three-species tree that drops species 1,

$$P_{yxx} = P_{xyxx} + P_{xxyy} + P_{yzxx}, \tag{73}$$

and there is a similar equation for P_{xyx}. Equation (13) can then be written in terms of the four-species pattern type frequencies. Similarly, for the three-species tree in which species 2 is dropped the same thing can be done. The resulting two equations define clock invariants; neither is equivalent to any of the equations (29)-(32). This would seem to define two more clock invariants when we needed only one. However, it can be shown that equations (29)-(32) do establish the equivalence of the two new clock invariants, so that they account for only one degree of freedom between them. Adding the two new invariants to express the new clock invariant in the most symmetrical form, we get equation (33).