Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mathematical Biosciences 186 (2003) 113–144
Mathematical Biosciences
www.elsevier.com/locate/mbs

Phylogenetic invariants for the general Markov model of sequence mutation

Elizabeth S. Allman [a,*], John A. Rhodes [b]

[a] Department of Mathematics and Statistics, University of Southern Maine, 96 Falmouth Street, Portland, ME, 04104, USA

[b] Department of Mathematics, Bates College, 3 Andrews Road, Lewiston, MA 04240, USA

Received 5 November 2002; received in revised form 1 August 2003; accepted 26 August 2003

Abstract

A phylogenetic invariant for a model of biological sequence evolution along a phylogenetic tree is a polynomial that vanishes on the expected frequencies of base patterns at the terminal taxa. While the use of these invariants for phylogenetic inference has long been of interest, explicitly constructing such invariants has been problematic.

We construct invariants for the general Markov model of κ-base sequence evolution on an n-taxon tree, for any κ and n. The method depends primarily on the observation that certain matrices defined in terms of expected pattern frequencies must commute, and yields many invariants of degree $\kappa + 1$, regardless of the value of n. We define strong and parameter-strong sets of invariants, and prove several theorems indicating that the set of invariants produced here has these properties on certain sets of possible pattern frequencies. Thus our invariants may be sufficient for phylogenetic applications.

Keywords: Phylogenetic invariants; Tree; Sequence evolution

1. Introduction

In [1], Cavender and Felsenstein and, in an independent work [2], Lake introduced an approach to phylogenetic tree construction from biological sequence data called *phylogenetic invariants*. We briefly and informally describe this method as applied to DNA sequences.

* Corresponding author. Tel.: +1-207 780 4930; fax: +1-207 780 5607.
E-mail addresses: eallman@maine.edu (E.S. Allman), jrhodes@bates.edu (J.A. Rhodes).

0025-5564/$ - see front matter
doi:10.1016/j.mbs.2003.08.004

Given a topological tree relating n terminal taxa and a particular parameterized model of molecular evolution along this tree, one can compute the expected pattern frequencies of each of the 4^n patterns of various bases at the terminal taxa, in terms of the parameters of the model. For simple yet natural models, these expected pattern frequencies will be polynomials in the model parameters.

A *phylogenetic invariant* for the topological tree and parameterized model is another polynomial, in 4^n variables, which becomes zero when the expected pattern frequencies are substituted for the variables, regardless of the values of the model parameters. If such phylogenetic invariants can be found, then one might use them to choose topological trees (and/or models of evolution) consistent with sequence data as follows: From aligned DNA sequences, compute the observed frequencies of patterns. Assuming these observed frequencies are good estimators of the expected frequencies for some choice of model parameters, they should cause the phylogenetic invariants to vanish, or at least be small. The optimal topological tree is chosen as the one for which the invariants come the closest to vanishing on the observed frequency data. Thus one has a model-based method of choosing topological trees which, unlike current maximum likelihood approaches, does not require the full estimation of all model parameters.

For such a scheme to be useful, however, many issues require better understanding. First, one must have a practical, efficient way of producing phylogenetic invariants. One might also hope to learn which of the invariants distinguish among different topological trees (i.e., are topologically informative), and which give no such information. Then, in order to apply invariants to real data, one must decide what it means for an invariant to be 'close to vanishing' on observed frequencies. A statistical understanding of the behavior of these polynomials on noisy data is highly desirable. Moreover, as there are infinitely many invariants, choosing a finite set of generators with good statistical properties is necessary. Finally, robustness of the method under violation of model assumptions is critical to applications, since models of sequence evolution are only approximations of reality. While much work remains to implement such a plan, the approach has intrigued a number of researchers. (See [3] for a survey and further references.)

A recent work of Chor et al. [4] makes use of phylogenetic invariants (for a two-state symmetric model) in another way. Though the focus of their paper is the construction of examples of observed pattern frequencies that lead to non-unique maximum likelihood trees, the approach illustrates the potential usefulness of invariants in maximum likelihood calculations. One can maximize the likelihood as a function of expected pattern frequencies, subject to the constraints that these expected frequencies satisfy the phylogenetic invariants, rather than searching directly for parameter values to maximize the likelihood function.

In this paper, we address only the question of finding phylogenetic invariants. Several previous approaches to this question exist. In the work of Cavender and Felsenstein [1] clever arguments based on the 4-point condition with log-det metric, and on statistical independence of evolutionary processes along different parts of a tree are used to produce invariants for the Jukes–Cantor 2-base model with 4 terminal taxa.

In later work of Ferretti and Sankoff [5–7] invariants are found empirically for a variety of models by looking for algebraic relationships among expected frequencies for particular parameter values, and then proving these empirically-found polynomials to be true invariants. A weakness of this method is that while 'all' invariants of a given low degree can be found, one has little understanding of what 'new' invariants of higher degree might remain unknown.

Evans and Speed [8] were able to make a significant leap for the Kimura 3-parameter model, by introducing the use of harmonic analysis on a certain abelian group that reflects the structure of this model. While this method finds all invariants, it seems to depend strongly on the particular model structure, so that it is limited in its ability to be generalized. A related approach to invariants for the Kimura 3-parameter model was found by Steel et al. in [9], building on the work of Hendy and Penny [10]. See also [11].

For the general Markov model, Steel gave a single invariant in [12] from consideration of the 4-point condition with log-det metric. Further work of Semple and Steel [13] gave many additional invariants for this model: each subtree relating either 3 or 4 terminal taxa gave rise to invariants expressible as the entries in certain matrix equations. However, since all these invariants are deduced through considering frequencies of patterns at only two terminal taxa at a time, one might suspect other invariants exist that were not found. Another construction of a large number of invariants is reported by Hagedorn in [14].

The language of algebraic geometry is of course the natural one for discussing polynomial phylogenetic invariants. That Gröbner basis techniques from computational algebraic geometry could in principal produce all invariants has been pointed out several times, including in [15,16]. However, the number of variables involved in such computations seems to place them well beyond the reach of current technology, except in the simplest model situations.

In this paper we present several methods of finding phylogenetic invariants for the general Markov model of base substitution along any topological tree. We place no restrictions on either the number n of terminal taxa the tree relates, nor on the number κ of bases from which sequences are made. Our approach requires nothing more than linear algebra, and even for large n allows one to easily produce many invariants that do not simply arise from subtrees relating fewer taxa. While the invariants found previously in [12,13] are included among those constructed here, many new invariants are produced also.

Two of our constructions, one based on the commutation of certain matrix expressions in the expected pattern frequencies and one based on the symmetry of other matrix expressions, yield invariants of degree $\kappa + 1$ for sequences composed of κ bases. (Note that in [14] Hagedorn reports that this is the lowest degree at which one should expect to find non-trivial invariants.) Other constructions yield invariants of degree 2κ. In all cases there is no dependency of the degree on the number of taxa. Furthermore, the nature of the constructions allow one to associate invariants to branching features of the tree, so that one can design invariants to test for certain phylogenetic relationships.

The work of Chang [17] contains one of our key insights, on the diagonalizations of certain matrix expressions in the expected pattern frequencies. However that work was directed at proving that model parameters could be recovered, and did not exploit these diagonalizations for finding invariants. The fact that simultaneously diagonalizable matrices must commute lies at the heart of our approach.

As anyone who has explicitly calculated invariants knows, one can quickly be overwhelmed by staring at polynomials in a large number of variables with many terms. The invariants we find, however, have the rather welcome feature that they are expressible by equating entries in certain matrix products. Not only is this psychologically pleasant, it also allows for simple implementations in software.

While we are not able to prove that we have found all phylogenetic invariants for the general Markov model, we prove several theorems that give some assurance that a 'sufficiently large' set of invariants is in hand. The proofs of these theorems focus attention on what are perhaps the most interesting of our invariants, which are those $(\kappa+1)$-degree ones deduced from the commutation of certain matrix expressions, as mentioned before. Indeed, the other invariants we construct play no role in our sufficiency proofs.

In the 3-taxon case with $\kappa \leqslant 4$ bases, we show (Theorem 5 and its corollaries) that we have a large enough set of polynomials that, provided a certain non-singularity condition is met, the vanishing of our invariants at a point implies the vanishing of any invariant, including those we might have failed to find. We also show (Theorems 11, 13), that in the n-taxon case with any number κ of bases, any point for which our invariants vanish that is 'nearly-diagonal' arises from model parameters, and thus will also result in the vanishing of any invariant, including those we might have failed to find.

Since in biological applications one expects the relevant points to both satisfy the specified non-singularity condition and to be nearly-diagonal, these results indicate that our invariants should be a large enough set for use in phylogenetic applications.

We close with several concrete examples showing possible pitfalls in the use of invariants for phylogenetic inference. There are arrays which satisfy all invariants for the general Markov model, yet do not arise as the pattern frequency arrays for any choice of model parameters. Lest this be interpreted too negatively, we point out that the use of the log-det distance and the four-point condition to infer a 4-taxon phylogeny can be interpreted as the use of a single, specific invariant for phylogenetic inference. Therefore, despite these examples, there is strong evidence that in practice invariants may be valuable.

2. Phylogenetic models and invariants

We denote by κ the number of letters (or bases) in the alphabet from which sequences are constructed, and use $1, 2, 3, \ldots, \kappa$ to denote the letters. Thus for DNA sequences $\kappa = 4$, and we might identify the bases A, C, G and T with the numbers 1–4.

Since much of the analysis in this paper focuses on considering only 3 terminal taxa, one of which is assumed to be the root of the tree, we first describe our model in that situation.

2.1. The general Markov model: three taxa

Let a, b, and c denote three terminal taxa, or *leaves*. There is only one unrooted bifurcating 3-leaf tree topology which can describe their phylogeny, as shown in Fig. 1.

The general Markov model of mutation we consider includes the following assumptions. All mutations are assumed to be base substitutions. Along each edge of the tree proceeding away from the root, substitutions occur at each site in a sequence, independent of other sites, but following an identical process that depends only on the edge (the i.i.d. assumption). Furthermore, substitution probabilities along various edges of the tree depend only on the immediate ancestor sequence (the Markov assumption).

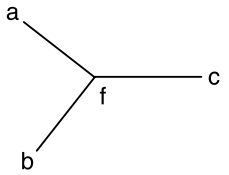

Fig. 1. The 3-taxon tree.

Let f denote the central vertex of the 3-leaf tree, as in Fig. 1. Assume taxon a is the root of the tree. (This assumption will be discussed further in Section 3 below.)

We specify a *root distribution vector* at a, as a row vector

$$\mathbf{p}_a = (\mathbf{p}_a(1), \mathbf{p}_a(2), \ldots, \mathbf{p}_a(\kappa)),$$

with $\mathbf{p}_a(i) \geqslant 0$ for all i and $\sum_{i=1}^{\kappa} \mathbf{p}_a(i) = 1$. We interpret $\mathbf{p}_a(i)$ as the probability of base i occurring at a site in a sequence at taxon a.

For each directed edge $a \to f$, $f \to b$ and $f \to c$ of the tree leading away from the root, a $\kappa \times \kappa$ Markov matrix M_{af}, M_{fb}, M_{fc} is given. That is, the entries of each matrix are non-negative, and each row sums to 1. The entries of these matrices are interpreted as probabilities of various base substitutions occurring at any particular site in the sequence. For instance, $M_{af}(i,j)$, the i, j entry of M_{af}, is the conditional probability that if base i appears at vertex a at a particular site, then base j will appear at vertex f at that site.

The vector $\mathbf{p}_a$ and the matrices M_{af}, M_{fb}, M_{fc} constitute the parameters of our model, which we denote by

$$\mathscr{M} = (\mathbf{p}_a, M_{af}, M_{fb}, M_{fc}).$$

Since $\mathbf{p}_a$ has $\kappa - 1$ degrees of freedom, and each of the matrices has $\kappa(\kappa - 1)$, there are a total of $(3\kappa + 1)(\kappa - 1)$ scalar parameters. To be more specific, we specify the scalar parameters as the first $\kappa - 1$ entries of $\mathbf{p}_a$ and the non-diagonal entries of the Markov matrices. Then the remaining entries are given by formulas

$$\mathbf{p}_a(\kappa) = 1 - \sum_{i=1}^{\kappa-1} \mathbf{p}_a(i), \quad M_{af}(j,j) = 1 - \sum_{i \neq j} M_{af}(j,i),$$

$$M_{fb}(j,j) = 1 - \sum_{i \neq j} M_{fb}(j,i), \quad M_{fc}(j,j) = 1 - \sum_{i \neq j} M_{fc}(j,i).$$

In particular, all entries of $\mathbf{p}_a$, M_{af}, M_{fb} and M_{fc} are (linear) polynomials in our chosen scalar parameters.

Because we take an algebraic approach in this work, at times we will need to allow the scalar parameters to be any complex numbers. Note that then the entries of $\mathbf{p}_a$, M_{af}, M_{fb}, and M_{fc} are also allowed to be any complex numbers, as long as each row sums to 1. In such situations we refer to *complex parameters*. When the additional criteria that all entries of $\mathbf{p}_a$, M_{af}, M_{fb}, and M_{fc} be non-negative real numbers holds, we refer to the parameters as being *stochastic*.

2.2. Expected frequencies of patterns

From a choice of parameters

$$\mathscr{M} = (\mathbf{p}_a, M_{af}, M_{fb}, M_{fc}),$$

we can compute the expected frequency of any pattern in sequences whose mutation is described by our model. Let $E_{abcf}(i,j,k,l)$ denote the expected frequency of the pattern with i at a, j at b, k at c, and l at f. Then E_{abcf} is a four-dimensional array, the *expected frequency array*, with entries

$$E_{abcf}(i,j,k,l) = \mathbf{p}_a(i)M_{af}(i,l)M_{fb}(l,j)M_{fc}(l,k).$$

Note that the entries of E_{abcf} are monomials in the entries of $\mathbf{p}_a$ and M_{af}, M_{fb}, M_{fc}, and hence are polynomials in the scalar parameters of the model described above.

We will also need notation for various subarrays of E_{abcf}, as well as for the marginal arrays found by summing over various indices. For instance, we let E_{a2cf} be the three-dimensional array defined by

$$E_{a2cf}(i,j,k) = E_{abcd}(i,2,j,k),$$

while $E_{abc\Sigma}$ is the three-dimensional array defined by

$$E_{abc\Sigma}(i,j,k) = \sum_{l=1}^{\kappa} E_{abcf}(i,j,k,l).$$

More generally, replacing one or more of the subscripts a, b, c, f with a number indicates the subarray whose entries are those entries of E_{abcf} with the given numbers occuring in the corresponding indices, while replacing one or more of a, b, c, f with a Σ indicates the marginal array obtained by summing over the corresponding index.

With this notation, $E^T_{a\Sigma\Sigma\Sigma} = \mathbf{p}_a$, since both are simply the vectors of expected frequencies at a of various bases $1,2,\ldots,\kappa$, regardless of what appears at b, c and f. For another example, $E_{ab2\Sigma}$ is the frequency matrix of patterns with various bases at a and b, but with 2 at c and any base at f.

Since the three-dimensional array $E_{abc\Sigma}$ will play a particularly important role, we also denote it by E_{abc}. It is the expected frequency array of patterns at the leaves (and thus its entries can be estimated from sequence data for the terminal taxa alone). The same notational conventions on replacing a, b or c by numbers or Σ's will be used for subarrays and marginal arrays of E_{abc}. When we need to be explicit about choices of model parameters we write $E_{abc}(\mathscr{M})$ with $\mathscr{M} = (\mathbf{p}_a, M_{af}, M_{fb}, M_{fc})$.

Note all the entries in all of the subarrays and marginal arrays associated to $E_{abcf}(\mathscr{M})$ and $E_{abc}(\mathscr{M})$ will be polynomials (of degree at most 4) in the scalar parameters specifying $\mathscr{M}$.

2.3. Trees relating n taxa

Our notation naturally generalizes to trees and models relating more than 3 terminal taxa.

We adopt the phrase *n-taxon tree* as shorthand for an unrooted topological bifurcating tree with n leaves labeled by the taxa. There are thus three 4-taxon trees relating taxa a, b, c, and d, as shown in Fig. 2.

Consider the 4-taxon tree T_1. Assuming we root T_1 at a, with internal vertices labeled as shown, our model parameters will be $\mathscr{M} = (\mathbf{p}_a, M_{ae}, M_{eb}, M_{ef}, M_{fc}, M_{fd})$, where $\mathbf{p}_a$ is again the root distri-

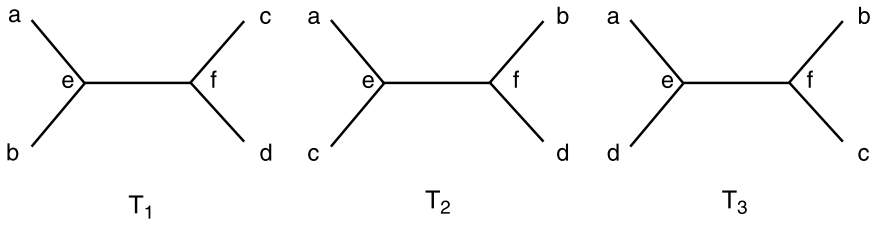

Fig. 2. The 4-taxon trees.

bution vector, and the various M_{xy} are Markov matrices for each edge in the tree directed away from the root.

Since a bifurcating tree with n leaves has $2n-3$ edges, we specify a mutation model by indicating a topological tree with n labeled leaves, a root distribution vector at leaf a, and $2n-3$ Markov matrices assigned to the edges directed away from a. As above, this gives $(\kappa-1)+\kappa(\kappa-1)(2n-3)$ scalar parameters. While strictly speaking one should call the tree topology a model parameter, we find it more convenient in this paper to reserve the word 'parameter' for numerical quantities.

In the 4-leaf case with tree T_1, one has a six-dimensional array E_{abcdef} of expected pattern frequencies at all vertices, where

$$E_{abcdef}(i,j,k,l,m,n) = \mathbf{p}_a(i)M_{ae}(i,m)M_{eb}(m,j)M_{ef}(m,n)M_{fc}(n,k)M_{fd}(n,l).$$

The four-dimensional array E_{abcd} of expected pattern frequencies at the leaves is defined by $E_{abcd} = E_{abcd\Sigma\Sigma}$ where

$$E_{abcd\Sigma\Sigma}(i,j,k,l) = \sum_{m=1}^{\kappa}\sum_{n=1}^{\kappa} E_{abcdef}(i,j,k,l,m,n).$$

Of course the parameters $\mathscr{M}$ and the array $E_{abcd}(\mathscr{M})$ reflect the labeled topological tree specified, as one sees upon imitating our development for either of the other labeled 4-taxon trees T_2 and T_3.

The n-taxon case is similar. Obvious modifications describe model parameters for trees rooted at internal vertices.

2.4. Phylogenetic invariants and varieties

Returning to the 3-leaf case for concreteness, let X_{abc} denote a $\kappa\times\kappa\times\kappa$ array of indeterminants, and $\mathbb{C}[X_{abc}] = \mathbb{C}[X_{111},\ldots X_{\kappa\kappa\kappa}]$ the polynomial ring in its κ^3 indeterminant entries, with complex coefficients. Then a *phylogenetic invariant* for the 3-leaf general Markov model is a polynomial $p\in\mathbb{C}[X_{abc}]$ such that $p\equiv 0$ under the substitution $X_{abc}\leftarrow E_{abc}(\mathscr{M})$ of the polynomial expressions (in terms of the scalar parameters) for the expected pattern frequencies at the leaves.

If two polynomials p_1 and p_2 vanish under this substitution, then so does any $\mathbb{C}[X_{abc}]$-linear combination of p_1 and p_2. Thus the set of all phylogenetic invariants for the 3-leaf general Markov model forms an ideal in $\mathbb{C}[X_{abc}]$. We denote this *phylogenetic invariant ideal* by $\mathfrak{A}_T$, where T is the tree of Fig. 1, the only bifurcating topological tree that can relate 3 terminal taxa.

Similarly, for the 4-leaf case, and the tree T_1 of Fig. 2, one has an ideal $\mathfrak{A}_{T_1}\in\mathbb{C}[X_{abcd}]$ of all polynomials that vanish identically under the substitution $X_{abcd}\leftarrow E_{abcd}(\mathscr{M})$. Since there are three

possible 4-taxon trees, we in fact have three such 4-taxon ideals, which we might denote $\mathfrak{A}_{T_1}$, $\mathfrak{A}_{T_2}$, and $\mathfrak{A}_{T_3}$. Any polynomial p such that $p \in \mathfrak{A}_{T_i}$ but $p \notin \mathfrak{A}_{T_j}$ for some i, j is *topologically-informative*, since it will vanish on all expected frequencies arising from tree i, but not vanish on most of those arising from tree j.

For the n-taxon case we similarly have an ideal of phylogenetic invariants for each n-taxon tree, and a concept of topologically-informative invariants.

For each topological tree, one would like to be able to explicitly give the ideal of phylogenetic invariants for the general Markov model. Since these ideals are finitely generated (by the Hilbert basis theorem) this would mean to give an explicit list of generators of the ideal. Even in the 3-leaf case, where there is only one topology and thus no topologically-informative invariants, determining generators of the phylogenetic invariant ideal might be valuable for measuring the fit of the general Markov model to data.

Finding phylogenetic invariants can be viewed as a problem in computational algebraic geometry. In principle, Gröbner basis methods can be used to find generators of the ideal $\mathfrak{A}_T$ from the polynomial entries of the expected frequency arrays. However, the computation seems to be beyond current capabilities, due to the large number of variables involved in the elimination process. Nonetheless, we will find it convenient to use some of the language of algebraic geometry.

For any ideal $\mathfrak{a} \in \mathbb{C}[x_1, \ldots, x_m]$, the affine algebraic variety associated to $\mathfrak{a}$ is

$$V(\mathfrak{a}) = \{\mathbf{x} \in \mathbb{C}^m \,|\, p(\mathbf{x}) = 0 \text{ for all } p \in \mathfrak{a}\}.$$

For an n-taxon tree T with phylogenetic invariant ideal $\mathfrak{A}_T$, the *phylogenetic variety* associated to T is $V(\mathfrak{A}_T) \subseteq \mathbb{C}^{\kappa^n}$. For any choice $\mathscr{M}$ of model parameters for the chosen topological tree, whether stochastic or complex, the array $E(\mathscr{M})$ of expected frequencies of patterns at the terminal taxa will produce a point in $V(\mathfrak{A}_T)$. However, $V(\mathfrak{A}_T)$ will typically contain many other points as well (see Section 9 for examples).

For a tree T, let m denote the number of scalar parameters in the general Markov model on T. Then we can view complex parameters as points in $\mathbb{C}^m$, and the stochastic parameters as points in a subset of $[0,1]^m \subset \mathbb{R}^m$. Several times we will use the observation that if $p(X)$ is a polynomial that vanishes under the substitution $X \leftarrow E(\mathscr{M})$ for all choices of $\mathscr{M}$ in a non-empty open subset of either $\mathbb{C}^m$ or $\mathbb{R}^m$, then $p \in \mathfrak{A}_T$. This is a consequence of viewing $p(E(\mathscr{M}))$ as a polynomial in the scalar parameters, and the fact that a multivariable polynomial which vanishes on a non-empty open set in $\mathbb{C}^n$ or $\mathbb{R}^n$ must be identically zero.

In particular, we can characterize the phylogenetic variety $V(\mathfrak{A}_T)$ as the smallest algebraic variety containing all points of the form $E(\mathscr{M})$ when $\mathscr{M}$ is allowed to range over any non-empty open subset of either $\mathbb{C}^m$ or $\mathbb{R}^m$.

3. Alternative root locations

Although we assume throughout most of this paper that our trees are rooted at leaf a, this assumption is in fact not essential for studying phylogenetic invariants of the general Markov model. More specifically, the phylogenetic invariant ideal associated to an n-taxon tree is independent of any choice of root location, whether at a leaf, at an internal node of valence 3, or at a node of valence 2 inserted along some edge. This follows from the following proposition, which is a slight variation on Theorem 2 of [18], with a similar proof.

Proposition 1. *Suppose an n-taxon tree T is given. Let r be some choice of root for T (which may be a leaf, an internal node of valance* 3*, or along some edge). Let stochastic parameters $\mathscr{M}_r = \{\mathbf{p}_r, M_{xy}, \ldots\}$ for the general Markov model on the rooted tree T_r be given.*

Suppose further that all entries of $\mathbf{p}_r$ are positive, and no column of any of the M_{xy} is zero. Then for any other choice of a root s for T at either a leaf or an internal node of valance 3*, there is a uniquely determined choice of general Markov model parameters $\mathscr{M}_s$ with the same properties producing the same expected frequency array as $\mathscr{M}_r$ at the leaves and internal nodes of valence* 3*.*

Since the proposition equates the expected frequency array at leaves and internal nodes of valance 3 for different roots under certain conditions, it certainly implies the equality of the expected frequency array at the leaves alone. This implies a result on phylogenetic invariants:

Corollary 2. *Let T be an n-taxon tree and let a be one of the taxa labeling the leaves. Then the phylogenetic invariant ideal $\mathfrak{A}_T$ for the general Markov model on T rooted at a is identical to the phylogenetic invariant ideal $\mathfrak{A}_{T_r}$ for the general Markov model on T rooted at r, where r is any other leaf, internal node of valance* 3*, or new node inserted on an edge of T.*

Proof. To see $\mathfrak{A}_T \subseteq \mathfrak{A}_{T_r}$, suppose $p(X) \in \mathfrak{A}_T$. Let $E(\mathscr{M})$ be the expected frequency array at the leaves for the model rooted at a with parameters $\mathscr{M}$, and $E(\mathscr{M}_r)$ the expected frequency array at the leaves for the model rooted at r with parameters $\mathscr{M}_r$.

Let m_r denote the number of scalar parameters associated to $\mathscr{M}_r$. The real scalar parameters for $\mathscr{M}_r$ resulting in $\mathbf{p}_r$ having positive entries, all $M_{xy} \in \mathscr{M}_r$ having non-zero columns, and $\mathscr{M}_r$ being stochastic form a non-empty open set in $\mathbb{R}^{m_r}$. By Proposition 1, we see that for all $\mathscr{M}_r$ in this set, $E(\mathscr{M}_r) = E(\mathscr{M})$ for some $\mathscr{M}$. Thus $E(\mathscr{M}_r) \in V(\mathfrak{A}_T)$, and so $p(E(\mathscr{M}_r)) = 0$ on this open set in parameter space. We conclude $p \in \mathfrak{A}_{T_r}$.

We similarly see $\mathfrak{A}_{T_r} \subseteq \mathfrak{A}_T$ if the root r is at either a leaf or an internal node of valence 3 of T.

If, however r lies on an edge of T, say the edge from f to c, we need one additional observation to complete the argument. Let $E(\mathscr{M}_f)$ denote the expected frequency array at the leaves for the tree rooted at f, with parameters $\mathscr{M}_f$. By Proposition 1 again, if m is the number of scalar parameters in $\mathscr{M}$, we know that for any $\mathscr{M}$ in a certain non-empty open set in $\mathbb{R}^m$ there is an $\mathscr{M}_f$ such that $E(\mathscr{M}) = E(\mathscr{M}_f)$. But if $\mathscr{M}_f = \{\mathbf{p}_f, M_{fc}, \ldots\}$, then $E(\mathscr{M}_f) = E(\mathscr{M}_r)$ where $\mathscr{M}_r = \{\mathbf{p}_r, M_{rf}, M_{rc}, \ldots\}$ is defined by letting $\mathbf{p}_r = \mathbf{p}_f$, $M_{rf} = I$, and $M_{rc} = M_{fc}$, and retaining for $\mathscr{M}_r$ all Markov matrices in $\mathscr{M}_f$ associated to edges other than $f \to c$. Thus for all $\mathscr{M}$ in some open set in the parameter space of the model rooted at a, $E(\mathscr{M}) \in V(\mathfrak{A}_{T_r})$ and we can proceed as above. □

This corollary justifies our assumption throughout the rest of this paper that trees be rooted at a leaf.

4. Recovering parameters from E_{abc}

Our basic viewpoint leading to the construction of invariants focuses first on the tree relating 3 terminal taxa. It is intimately tied to the question of when and how one can recover the parameters $\mathscr{M}$ from a numerical array $E_{abc}(\mathscr{M})$, which was addressed by Chang in [17]. In order to

both motivate our approach, and provide insight into our construction, we summarize some of the ideas and issues raised in that paper, in the context of a 3-taxon tree.

Of course, considering more than three taxa is important for most real applications. Generalizations to four and more taxa will be discussed in subsequent sections.

4.1. Obstacles to recovering parameters

Suppose for the 3-taxon tree numerical parameters $\mathscr{M} = (\mathbf{p}_a, M_{af}, M_{fb}, M_{fc})$ are given, thus determining an array $E_{abc} = E_{abc}(\mathscr{M})$. Several simple observations indicate $\mathscr{M}$ can not always be recovered from E_{abc}, at least without imposing additional conditions or assumptions.

If some of the Markov matrix parameters are singular, then there may be infinitely many choices of parameters giving the same array E_{abc}. To illustrate this, consider an extreme example where all entries of both M_{fb} and M_{fc} are $1/\kappa$. Then we find

$$E_{abc}(i,j,k) = \sum_{l=1}^{\kappa} \mathbf{p}_a(i) M_{af}(i,l) M_{fb}(l,j) M_{fc}(l,k) = \sum_{l=1}^{\kappa} \mathbf{p}_a(i) M_{af}(i,l) \frac{1}{\kappa^2} = \frac{1}{\kappa^2} \mathbf{p}_a(i).$$

Thus the array E_{abc} is independent of the choice of M_{af}. More subtle examples of arrays E_{abc} arising from infinitely many choices of parameters can be constructed in which two of the Markov matrices are non-singular.

There is another issue preventing the unique recovery of $\mathscr{M}$ from E_{abc}. Informally, one can insert a permutation of the bases at the internal node of the tree without affecting any of the expected frequencies in E_{abc}.

To be more specific, let σ be a permutation of the set $\{1, 2, \ldots, \kappa\}$. Then there is an associated permutation matrix P with the property that for any row vector $v = (v_1, v_2, \ldots, v_\kappa)$, we have

$$vP = (v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(\kappa)}).$$

Then if M is any matrix with κ columns, MP has the same columns as M, but reordered by σ. Similarly for a matrix M with κ rows, $P^T M$ has the same rows as M, but reordered by σ.

Proposition 3. *Suppose σ is a permutation of $\{1, 2, 3 \ldots, \kappa\}$ with associated permutation matrix P. For any choice of model parameters $\mathscr{M} = (\mathbf{p}_a, M_{af}, M_{fb}, M_{fc})$, let*

$$\mathscr{M}^\sigma = (\mathbf{p}_a, M_{af}P, P^T M_{fb}, P^T M_{fc}).$$

Then $E_{abc}(\mathscr{M}^\sigma) = E_{abc}(\mathscr{M})$.

Proof

$$\begin{aligned} E_{abc}(\mathscr{M})(i,j,k) &= \sum_{l=1}^{\kappa} \mathbf{p}_a(i) M_{af}(i,l) M_{fb}(l,j) M_{fc}(l,k) \\ &= \sum_{l=1}^{\kappa} \mathbf{p}_a(i) M_{af}(i,\sigma(l)) M_{fb}(\sigma(l),j) M_{fc}(\sigma(l),k) \\ &= \sum_{l=1}^{\kappa} \mathbf{p}_a(i) (M_{af}P)(i,l) (P^T M_{fb})(l,j) (P^T M_{fc})(l,k) \\ &= E_{abc}(\mathscr{M}^\sigma)(i,j,k). \qquad \square \end{aligned}$$

Of course for an n-taxon tree with parameters $\mathscr{M}$, for any choice of a permutation for each of the $n-2$ internal nodes one can define a similar action of the permutations on $\mathscr{M}$ to produce new parameters $\mathscr{M}'$ with $E(\mathscr{M}') = E(\mathscr{M})$.

Remark 1. For biological applications, we expect our Markov matrices to have their largest entries along the main diagonal, since most sites should not mutate along an edge, or we would not have been able to align sequences. This means both that the Markov matrices are non-singular, since they are fairly close to the identity matrix, and that we can single out a single biologically-reasonable ordering of the columns of M_{af}. Thus the issues of non-uniqueness of parameters raised here are primarily of theoretical importance.

4.2. Recovering parameters

We turn now to the deduction of parameters $\mathscr{M}$ from the array E_{abc}.

Consider the expected frequency array E_{abk}, where k is a particular base at c. Then we can express its entries as

$$E_{abk}(i,j) = E_{abc}(i,j,k) = \sum_{l=1}^{\kappa} \mathbf{p}_a(i) M_{af}(i,l) M_{fb}(l,j) M_{fc}(l,k).$$

Letting

$$C_{fc,k} = \operatorname{diag}(M_{fc}(1,k), M_{fc}(2,k), \ldots, M_{fc}(\kappa,k))$$

denote the diagonal matrix formed from the kth column of M_{fc}, and $D_a = \operatorname{diag}(\mathbf{p}_a)$, this becomes

$$E_{abk} = D_a M_{af} C_{fc,k} M_{fb}. \tag{1}$$

Similarly,

$$E_{ab\Sigma}(i,j) = \sum_{l=1}^{\kappa} \mathbf{p}_a(i) M_{af}(i,l) M_{fb}(l,j),$$

so

$$E_{ab\Sigma} = D_a M_{af} M_{fb}. \tag{2}$$

Assuming D_a, M_{af}, and M_{fb} are non-singular, then

$$(E_{ab\Sigma})^{-1} E_{abk} = M_{fb}^{-1} C_{fc,k} M_{fb}.$$

Now the expression on the right side of this equation is simply a diagonalization of a matrix. That is, the rows of M_{fb} are the left eigenvectors of $(E_{ab\Sigma})^{-1} E_{abk}$, and the diagonal entries of $C_{fc,k}$ are the corresponding eigenvalues.

As long as the eigenvalues are distinct, the eigenvectors of a diagonalizable matrix are uniquely determined up to scalar multiples. Since M_{fb} is a Markov matrix, its rows must each sum to 1, so the particular scalar multiple is thus uniquely determined. Therefore the collection of rows of M_{fb} can be found from the eigenvectors; only the order in which those rows appear is not deducible from $(E_{ab\Sigma})^{-1} E_{abk}$. This, however, is precisely the issue described in Proposition 3. More formally, we obtain the following partial converse of that proposition, which is essentially Lemma 4.1 of [17], and whose proof we therefore omit.

Proposition 4. *Let*

$$\mathscr{M} = (\mathbf{p}_a, M_{af}, M_{fb}, M_{fc}) \quad \text{and} \quad \mathscr{M}' = (\mathbf{p}'_a, M'_{af}, M'_{fb}, M'_{fc}).$$

Suppose that M_{af} and M_{fb} are non-singular, no pair of rows of M_{fc} are identical, and $\mathbf{p}_a$ has all non-zero entries. Then $E_{abc}(\mathscr{M}') = E_{abc}(\mathscr{M})$ implies that $\mathscr{M}' = \mathscr{M}^{\sigma}$ for some permutation σ.

Note the proposition indicates that, provided all Markov parameters are non-singular, there are only finitely many parameter choices leading to the same expected frequency array.

The argument behind this last proposition in fact provides a constructive way of recovering parameters $\mathscr{M}$ from an array $E_{abc}(\mathscr{M})$. Provided $E_{ab\Sigma}$ is non-singular, one simply computes the matrices $(E_{ab\Sigma})^{-1}E_{abi}$ from E_{abc}, and then using standard algorithms computes their common eigenvectors, scaling appropriately to give choices of Markov matrix parameters. If $\kappa \leqslant 4$, one could in principle even find exact formulas for the eigenvectors, since the characteristic equation which must be solved is polynomial of degree κ. For any κ, one can compute the eigenvectors numerically.

5. Phylogenetic invariants for 3 taxa

As the last section has shown, if $X_{abc} = E_{abc}(\mathscr{M})$, then the matrices $X^{-1}_{ab\Sigma}X_{abi}$ must have a full set of common eigenvectors. Since matrices with common eigenvectors commute, this observation leads to the construction of phylogenetic invariants.

5.1. Commutation relations

We work initially under the assumption that any matrix whose inverse we need actually exists. After using this to deduce invariants, we will show the invariants found in this way are valid even if the assumption is not met.

First, for $i, j = 1, 2, \ldots, \kappa$, define matrices

$$Y_{c;i} = (X_{abi})(X_{ab\Sigma})^{-1}, \tag{3}$$

$$Y_{b;j} = (X_{ajc})(X_{a\Sigma c})^{-1}. \tag{4}$$

If $X_{abc} = E_{abc}(\mathscr{M})$, then from equations like (1) and (2) we deduce that

$$Y_{c;i} = (D_a M_{af})C_{fc;i}(D_a M_{af})^{-1},$$

$$Y_{b;j} = (D_a M_{af})C_{fb;j}(D_a M_{af})^{-1}.$$

Thus these matrices are simultaneously diagonalizable, and hence commute:

$$Y_{c;i}Y_{c;j} = Y_{c;j}Y_{c;i}, \tag{5}$$

$$Y_{c;i}Y_{b;j} = Y_{b;j}Y_{c;i}. \tag{6}$$

We first focus on Eq. (5) in the case where $i \neq j$. Expressing it in terms of X_{abc} yields

$$X_{abi}X^{-1}_{ab\Sigma}X_{abj}X^{-1}_{ab\Sigma} = X_{abj}X^{-1}_{ab\Sigma}X_{abi}X^{-1}_{ab\Sigma}.$$

Multiplying on the right by $X_{ab\Sigma}$, yields

$$X_{abi}(X_{ab\Sigma})^{-1}X_{abj} = X_{abj}(X_{ab\Sigma})^{-1}X_{abi}. \tag{7}$$

Now the inverse of a non-singular matrix A can be expressed as

$$A^{-1} = \frac{1}{\det(A)}\mathrm{Cof}(A)^T,$$

where $\mathrm{Cof}(A)$ denotes the matrix of cofactors of A. For a $\kappa \times \kappa$ matrix A, this is another $\kappa \times \kappa$ matrix whose entries are explicitly known polynomials in the entries of A, found by taking $\pm$determinants of $(\kappa-1)\times(\kappa-1)$ submatrices. As such, each entry of the cofactor matrix is a polynomial of degree $\kappa - 1$ in the entries of A, with $(\kappa-1)!$ terms.

Multiplying Eq. (7) by $\det(X_{ab\Sigma})$ gives

$$X_{abi}\mathrm{Cof}(X_{ab\Sigma})^T X_{abj} = X_{abj}\mathrm{Cof}(X_{ab\Sigma})^T X_{abi}. \tag{8}$$

This is an identity of matrices, yielding κ^2 scalar identities from the various entries. Furthermore, each side of the equation has entries that are polynomials of degree $\kappa+1$ in the entries of X_{abc}, so each identity, after possible cancelation, is of degree at most $\kappa+1$. While all κ^2 identities obtained this way may not be independent of one another, we do at least have a collection of invariants that must be satisfied if X_{abc} is of the form $E_{abc}(\mathscr{M})$ and $X_{ab\Sigma}$ is non-singular.

Writing each of these identities in the form

$$p(X_{abc}) = p(X_{111}, X_{112}, \ldots, X_{\kappa\kappa\kappa}) = 0,$$

we have a set of polynomial invariants

$$\mathscr{S}_{c;i,j} = \{p_n(X_{abc}) \,|\, n = 1, \ldots, \kappa^2\}.$$

We get such a set for each choice of the pair $i \neq j$, and there are $\kappa(\kappa-1)/2$ such pairs. This leads to a set of invariants

$$\mathscr{S}_c = \bigcup_{i,j} \mathscr{S}_{c;i,j}.$$

Similar arguments yield sets $\mathscr{S}_a$ and $\mathscr{S}_b$ of invariants arising from

$$X_{jbc}\mathrm{Cof}(X_{\Sigma bc})^T X_{ibc} = X_{ibc}\mathrm{Cof}(X_{\Sigma bc})^T X_{jbc},$$

$$X_{ajc}\mathrm{Cof}(X_{a\Sigma c})^T X_{aic} = X_{aic}\mathrm{Cof}(X_{a\Sigma c})^T X_{ajc}.$$

Each of the sets $\mathscr{S}_a$, $\mathscr{S}_b$, and $\mathscr{S}_c$ has at most $\kappa^3(\kappa-1)/2$ elements, each of which is a polynomial of degree at most $\kappa+1$.

Returning to Eq. (6) to obtain additional invariants, we express it in terms of X_{abc} and multiply it by $\det(X_{ab\Sigma})\det(X_{a\Sigma c})$ to get

$$X_{ajc}\mathrm{Cof}(X_{a\Sigma c})^T X_{abi}\mathrm{Cof}(X_{ab\Sigma})^T = X_{abi}\mathrm{Cof}(X_{ab\Sigma})^T X_{ajc}\mathrm{Cof}(X_{a\Sigma c})^T. \tag{9}$$

The entries of this matrix identity give κ^2 scalar invariants, each of which is of degree at most 2κ. We denote the set of these invariants by $\mathscr{S}_{bc;ji}$. There are κ^2 such sets as i, j range over $1, 2, \ldots, \kappa$, which we combine to form a set of cardinality at most κ^4:

$$\mathscr{S}_{bc} = \bigcup_{i,j} \mathscr{S}_{bc;ji}.$$

We can also construct sets $\mathscr{S}_{ac}$ and $\mathscr{S}_{ab}$, based on the formulas

$$X^T_{abj}\mathrm{Cof}(X_{ab\Sigma})X_{ibc}\mathrm{Cof}(X_{\Sigma bc})^T = X_{ibc}\mathrm{Cof}(X_{\Sigma bc})^T X^T_{abj}\mathrm{Cof}(X_{ab\Sigma}),$$

$$X^T_{ajc}\mathrm{Cof}(X_{a\Sigma c})X^T_{ibc}\mathrm{Cof}(X_{\Sigma bc}) = X^T_{ibc}\mathrm{Cof}(X_{\Sigma bc})X^T_{ajc}\mathrm{Cof}(X_{a\Sigma c}),$$

which can be derived similarly. Alternately, one may obtain these equations from Eq. (9) by interchanging the roles of a, b, and c, being careful to transpose matrices when necessary.

5.2. Symmetry relations

Another source of invariants is the fact that certain expressions defined in terms of the expected frequency array must produce symmetric matrices. For example, since

$$X_{abi} = D_a M_{af} C_{fc;i} M_{fb},$$

$$X_{\Sigma bc} = M^T_{fb}\mathrm{diag}(\mathbf{p}_a M_{af}) M_{fc},$$

$$X_{ajc} = D_a M_{af} C_{fb;j} M_{fc},$$

we see that

$$X_{abi}(X^T_{\Sigma bc})^{-1}X^T_{ajc} = D_a M_{af} C_{fc;i}\,\mathrm{diag}(\mathbf{p}_a M_{af}) C_{fb;j} M^T_{af} D_a$$

is symmetric. Thus invariants arise from the equation

$$X_{abi}\mathrm{Cof}(X_{\Sigma bc})X^T_{ajc} = X_{ajc}\mathrm{Cof}(X_{\Sigma bc})^T X^T_{abi}. \tag{10}$$

Interchanging the roles of a, b, c also yields

$$X_{ibc}\mathrm{Cof}(X_{a\Sigma c})^T X_{abj} = X^T_{abj}\mathrm{Cof}(X_{a\Sigma c})X^T_{ibc},$$

$$X^T_{ibc}\mathrm{Cof}(X_{ab\Sigma})^T X_{ajc} = X^T_{ajc}\mathrm{Cof}(X_{ab\Sigma})X_{ibc}.$$

Since an equation stating that a matrix is symmetric yields $\kappa(\kappa-1)/2$ non-trivial scalar equalities, these symmetry conditions yield at most $3\kappa(\kappa-1)/2$ invariants, all of degree at most $\kappa+1$. However, one can show that these invariants are a subset of those arising from commutation relations.

Finally, we have the trivial invariant that the entries of X_{abc}, being frequencies of all possible patterns, must sum to 1, so let

$$\mathscr{S}_0 = \{X_{111} + X_{112} + \cdots + X_{\kappa\kappa\kappa} - 1\}.$$

In total, we have found a set of invariants

$$\mathscr{S} = \mathscr{S}_0 \cup \mathscr{S}_a \cup \mathscr{S}_b \cup \mathscr{S}_c \cup \mathscr{S}_{ab} \cup \mathscr{S}_{bc} \cup \mathscr{S}_{ac}.$$

As one would expect, explicit computations for specific values of κ show these are usually non-trivial. Of course we should also expect many of these invariants to be consequences of others, so that they are not independent. For instance, identities such as

$$\sum_{i=1}^{\kappa} Y_{c;i} = I$$

can be used to explicitly produce some dependencies.

Actually, we have not yet given a complete argument as to why the elements of $\mathscr{S}$ are invariants. Our derivation of $\mathscr{S}$ above from commutation and symmetry relations required assumptions that $X_{ab\Sigma}$, $X_{a\Sigma c}$ and $X_{\Sigma bc}$ all be invertible, or equivalently that

$$\det(X_{ab\Sigma}) \neq 0, \quad \det(X_{a\Sigma c}) \neq 0, \quad \det(X_{\Sigma bc}) \neq 0.$$

However, we can certainly find a non-empty open set of parameters on which these conditions hold (for instance, a neighborhood of $\{(1/\kappa, \ldots, 1/\kappa), I, I, I\}$). But since any polynomial vanishing on $E(\mathscr{M})$ for all $\mathscr{M}$ in such an open set must lie in $\mathfrak{A}_T$, this shows $\mathscr{S} \subseteq \mathfrak{A}_T$.

Remark 2. We could obtain many other invariants, by varying the steps above slightly and temporarily assuming different matrices are invertible. For instance, in Eq. (3) we may replace $X_{ab\Sigma}$ by any linear combination of the X_{abk}s and in Eq. (4) we may replace $X_{a\Sigma c}$ by any linear combination of the X_{alb}s and then reason similarly under the assumption that these linear combinations are non-singular. We do not explicitly list the invariants so produced here for three reasons. First, there are no essentially new ideas in producing them. Second, all our results on the inferential power of the invariants we have found will use only some of the invariants already explicitly listed. Third, if one were to use these invariants on sequence data, one might choose to implement the invariants in a form using inverses rather than cofactors. Since one can expect matrices such as $X_{ab\Sigma}$ to be far from singular while the individual X_{abk}s should be more nearly singular, the numerical computation of the inverse of $X_{ab\Sigma}$ should be better behaved.

Nonetheless, using a symbolic computation package it is straightforward to produce an explicit list of all invariants discussed here. For example, one finds that for $\kappa = 4$, the invariants of degree 5 so produced form a 1728-dimensional vector space. By a partially-computational argument of Hagedorn [14], or a representation-theoretic argument of Landsberg and Manivel [19], this must be the full space of degree 5 invariants.

5.3. Additional invariants through saturation and radical

The invariants obtained through the commutation relations can, in principle, be refined through two additional steps. Though explicit computations of these steps seems beyond current capabilities of Gröbner basis packages, they still provide useful theoretical understanding.

Let $\mathfrak{I} = \langle \mathscr{S} \rangle$ be the ideal generated by $\mathscr{S}$ in the polynomial ring $\mathbb{C}[X_{abc}]$, so $\mathfrak{I} \subseteq \mathfrak{A}_T$.

Let $d_a(X_{abc})$, $d_b(X_{abc})$, and $d_c(X_{abc})$ be polynomials in the entries of X_{abc} defined by

$$d_a(X_{abc}) = \det(X_{\Sigma bc}), \quad d_b(X_{abc}) = \det(X_{a\Sigma c}), \quad d_c(X_{abc}) = \det(X_{ab\Sigma})$$

and consider the set of polynomials

$$\mathscr{T} = \mathscr{S} \cup \{1 - td_a(X_{abc}), 1 - ud_b(X_{abc}), 1 - vd_c(X_{abc})\},$$

where t, u, and v are 3 new indeterminants. In the ring $\mathbb{C}[X_{abc}, t, u, v]$, $\mathscr{T}$ generates an ideal $\mathfrak{T}$, and defines a variety $V(\mathfrak{T})$ in $\mathbb{C}^{\kappa^3+3}$. The projection of this variety onto $\mathbb{C}^{\kappa^3}$ is precisely the subset of $V(\mathfrak{I})$ comprised of those points at which none of d_a, d_b, d_c vanish. The ideal

$$\widetilde{\mathfrak{J}} = \mathfrak{T} \cap \mathbb{C}[X_{abc}]$$

is called the *saturation* of $\mathfrak{J}$ with respect to d_a, d_b, d_c. The variety $V(\widetilde{\mathfrak{J}})$ contains all the points of $V(\mathfrak{J})$ at which all these determinants are non-zero, and is, in fact, the smallest variety to do so.

Clearly $\mathfrak{J} \subseteq \widetilde{\mathfrak{J}}$, but it is also the case that $\widetilde{\mathfrak{J}} \subseteq \mathfrak{A}_T$. To see this we argue as before that there is a non-empty open set of parameters $\mathscr{M}$ for which these polynomials will vanish on $E_{abc}(\mathscr{M})$ (for instance, a neighborhood of $\{(1/\kappa, \ldots, 1/\kappa), I, I, I\}$). But since any polynomial vanishing on $E(\mathscr{M})$ for all $\mathscr{M}$ in such an open set must lie in $\mathfrak{A}_T$, this shows $\widetilde{\mathfrak{J}} \subseteq \mathfrak{A}_T$.

Remark 3. In principle, computing a Gröbner basis for $\mathfrak{T}$ using a monomial term ordering of '*tuv*-elimination type' (that is, any monomial ordering in which monomials involving any of t, u, or v are greater than all monomials in $\mathbb{C}[X_{abc}]$) would allow one to find a basis for the elimination ideal $\widetilde{\mathfrak{J}} = \mathfrak{T} \cap \mathbb{C}[X_{abc}]$.

Unfortunately a Gröbner basis calculation even for $\mathfrak{J}$, much less for $\mathfrak{T}$, when $\kappa = 4$ seems beyond the capability of current standard software packages. To see the difficulty, note that we are dealing with polynomials in $64 + 3$ variables, and the polynomials arising from Eqs. (8) and (10) and their analogs are homogeneous of degree 5 and have hundreds of terms. Those arising from Eq. (9) are more complex. However this elimination problem involves far fewer variables than that involved in the direct calculation of all invariants by Gröbner methods as outlined, for example, in [16].

Example. For $\kappa = 2$, things are simple. One readily sees that Eq. (8) yields only the zero polynomial, and thus $\mathscr{S}_c$, and similarly $\mathscr{S}_b$ and $\mathscr{S}_a$, contain only the zero invariant. In fact, one can see this without even doing a calculation since

$$Y_{c;1} + Y_{c;2} = I,$$

so the commutation of the $Y_{c;i}$ is guaranteed. Eqs. (9) and (10) are more opaque, but a calculation shows they also yield only the zero polynomial. Thus $\mathscr{S}_{bc}$, $\mathscr{S}_{ab}$, $\mathscr{S}_{ac}$ and $\mathscr{S}_{sym}$ also only contain zero, and $\mathscr{S} = \mathscr{S}_0$. That is, $\mathfrak{J}$ is generated by the trivial invariant. One then deduces that $\widetilde{\mathfrak{J}} = \mathfrak{J}$.

Of course, in this case our model has 7 scalar parameters, while $E(\mathscr{M})$ is a point in $\mathbb{C}^8$, so we would have expected $\mathfrak{A}_T = \langle \mathscr{S}_0 \rangle$, as a Gröbner basis calculation can confirm.

If the saturation $\widetilde{\mathfrak{J}}$ could be found for larger κ, one additional step could, in principal, produce a potentially larger ideal of invariants. By the Strong Nullstellensatz, the full ideal of all polynomials vanishing on $V(\widetilde{\mathfrak{J}})$ is the radical $\sqrt{\widetilde{\mathfrak{J}}}$, and of course $\sqrt{\widetilde{\mathfrak{J}}} \subseteq \mathfrak{A}_T$.

We summarize our results so far by the chain of inclusions

$$\mathfrak{J} \subseteq \widetilde{\mathfrak{J}} \subseteq \sqrt{\widetilde{\mathfrak{J}}} \subseteq \mathfrak{A}_T,$$

which implies

$$V(\mathfrak{J}) \supseteq V(\widetilde{\mathfrak{J}}) = V\left(\sqrt{\widetilde{\mathfrak{J}}}\right) \supseteq V(\mathfrak{A}_T).$$

For any κ we have an explicit list of generators of $\mathfrak{J}$, but do not know if we have explicit generators for any of the other ideals.

6. Sufficiency of the invariants arising from commutation relations: 3 taxa

While our method is able to produce a large set S of explicit invariants which generate an ideal $\mathfrak{I}$, due to current computational limits we have not be able to explictly find $\widetilde{\mathfrak{I}}$ or $\sqrt{\widetilde{\mathfrak{I}}}$. Moreover, at this point one might speculate that a large gap still lies between $\sqrt{\widetilde{\mathfrak{I}}}$ and $\mathfrak{A}_T$. One may thus reasonably ask what we have gained over the direct, yet infeasible, Gröbner basis calculation of generators of $\mathfrak{A}_T$ described in [16]. Is there a sense in which we have found 'enough' invariants?

We address this question in two ways. In this section we give a result for the 3-taxon case with $\kappa \leqslant 4$. Later in Section 8, after discussing constructing invariants for n-taxon trees, we shall give a different result that has no restriction on κ or n.

We will need some terminology which will be used in the n-taxon case also.

Definition. Let T be an n-taxon tree relating taxa $a_1, a_2, \ldots, a_n$ with phylogenetic invariant ideal $\mathfrak{A}_T$. Suppose $\mathscr{D} \subseteq \mathbb{C}^{\kappa^n}$ and $\mathscr{R} \subseteq \mathbb{C}[X_{a_1 a_2 \ldots a_n}]$. Then we say $\mathscr{R}$ is a *strong set of invariants* on $\mathscr{D}$ for the tree T if it has the properties

$$\mathscr{R} \subseteq \mathfrak{A}_T,$$

$$V(\langle \mathscr{R} \rangle) \cap \mathscr{D} \subseteq V(\mathfrak{A}_T).$$

Thus a set of invariants $\mathscr{R}$ being strong on $\mathscr{D}$ means that any point in $\mathscr{D}$ satisfying the invariants in $\mathscr{R}$ satisfies all possible invariants for the tree. As long as we only consider points in $\mathscr{D}$, a strong set of invariants has as much distinguishing power as all of $\mathfrak{A}_T$.

For any set $\mathscr{D}$, there is value in identifying small sets of invariants that are strong on $\mathscr{D}$. With this in mind, we focus on the 3-taxon case and let

$$\mathscr{S}' = \mathscr{S}_0 \cup \mathscr{S}_c \subseteq \mathscr{S}.$$

Thus $\mathscr{S}'$ contains the trivial invariant and only certain of the $(\kappa+1)$-degree invariants arising from commutation relations.

Theorem 5. *Suppose $\kappa \leqslant 4$ in the 3-taxon case. Let $\mathscr{O}$ denote the open set that is the complement of $V(\langle d_c \rangle)$. Then $\mathscr{S}'$ is a strong set of invariants on $\mathscr{O}$.*

Before giving the proof of this theorem, we focus on its implications.

Since an ordering of the taxa a, b, c is arbitrary, one can consider more invariants to get a set of strong invariants on a larger set.

Corollary 6. *Suppose $\kappa \leqslant 4$ in the 3-taxon case. Let $\mathscr{Q}$ denote the open set that is the complement of*

$$V(\langle d_a, d_b, d_c \rangle) = V(\langle d_a \rangle) \cap V(\langle d_b \rangle) \cap V(\langle d_c \rangle)$$

and let $\mathscr{S}'' = \mathscr{S}_0 \cup \mathscr{S}_a \cup \mathscr{S}_b \cup \mathscr{S}_c$. Then $\mathscr{S}''$ is a strong set of invariants on $\mathscr{Q}$, and thus $\mathscr{S}$ is a strong set of invariants on $\mathscr{Q}$.

Proof. If $X_{abc} \in \mathscr{Q}$, then at least one of the d_a, d_b, d_c is non-zero at X_{abc}. Permuting the taxa if necessary, we may assume $d_c(X_{abc}) \neq 0$. Then Theorem 5 applies to show $X_{abc} \in V(\mathfrak{A}_T)$.

Since $\mathscr{S}'' \subseteq \mathscr{S}$, $\mathscr{S}$ is strong on any set on which $\mathscr{S}''$ is strong. □

Let $\mathfrak{I}' = \langle \mathscr{S}' \rangle$, $\widetilde{\mathfrak{I}}'$ its saturation with respect to the polynomial d_c, and $\sqrt{\widetilde{\mathfrak{I}}'}$ the radical of the saturation. Then

$$\mathfrak{I}' \subseteq \mathfrak{I}, \quad \widetilde{\mathfrak{I}}' \subseteq \widetilde{\mathfrak{I}}, \quad \sqrt{\widetilde{\mathfrak{I}}'} \subseteq \sqrt{\widetilde{\mathfrak{I}}}.$$

We can now show that if we were able to find explicit generators for $\widetilde{\mathfrak{I}}'$ and $\widetilde{\mathfrak{I}}$, these sets would be quite powerful.

Corollary 7. *For $\kappa \leqslant 4$ in the 3-taxon case,*

$$\sqrt{\widetilde{\mathfrak{I}}'} = \sqrt{\widetilde{\mathfrak{I}}} = \mathfrak{A}_T$$

and $V(\widetilde{\mathfrak{I}}') = V(\widetilde{\mathfrak{I}}) = V(\mathfrak{A}_T)$. Thus $\widetilde{\mathfrak{I}}'$ and $\widetilde{\mathfrak{I}}$ are strong sets of invariants on all of $\mathbb{C}^{\kappa^3}$.

Proof. With $\mathscr{O}$ as in Theorem 5, we have that

$$V(\mathfrak{I}') \cap \mathscr{O} \subseteq V(\mathfrak{A}_T).$$

But $V(\widetilde{\mathfrak{I}}')$ is the smallest variety containing $V(\mathfrak{I}') \cap \mathscr{O}$, so

$$V(\widetilde{\mathfrak{I}}') \subseteq V(\mathfrak{A}_T).$$

Since $\widetilde{\mathfrak{I}}' \subseteq \widetilde{\mathfrak{I}} \subseteq \mathfrak{A}_T$, we also have

$$V(\widetilde{\mathfrak{I}}') \supseteq V(\widetilde{\mathfrak{I}}) \supseteq V(\mathfrak{A}_T).$$

Thus these three varieties must be equal, and so the corresponding radical ideals are equal. □

Example. For $\kappa = 2$, we have observed that $\widetilde{\mathfrak{I}} = \langle \mathscr{S}_0 \rangle$, from which it follows that $\sqrt{\widetilde{\mathfrak{I}}} = \langle \mathscr{S}_0 \rangle$. Therefore $\mathfrak{A}_T = \langle \mathscr{S}_0 \rangle$, and all invariants of the 3-taxon tree are multiples of the trivial invariant.

Our proof of Theorem 5 will depend on the irreducibility of certain varieties of commuting matrices.

Proposition 8. *Let $\mathscr{C}(n,m)$ denote the ane variety of commuting n-tuples of $m \times m$ matrices over $\mathbb{C}$. Then for $\kappa \leqslant 4$, $\mathscr{C}(\kappa - 1, \kappa)$ is irreducible.*

This is proved for $\kappa = 4$ by Guralnick and Sethuraman in [20]; the easier case of $\kappa = 3$ is shown by Guralnick in [21], with references given there to earlier proofs. This last paper also shows that for $\kappa > 4$ the variety is not irreducible.

Lemma 9. *For $\kappa \leqslant 4$, $(\kappa - 1)$-tuples of $\kappa \times \kappa$ simultaneously diagonalizable matrices are Zariski dense in $\mathscr{C}(\kappa - 1, \kappa)$.*

Proof. Consider the smaller set of $(\kappa - 1)$-tuples of $\kappa \times \kappa$ simultaneously diagonalizable matrices where the first matrix has κ distinct eigenvalues. This is the same as the set of $(\kappa - 1)$-tuples of

commuting matrices where the first matrix has distinct eigenvalues. Thus it is an open set in $\mathscr{C}(\kappa-1,\kappa)$, since the distinct eigenvalue condition can be specified as $p(x)\neq 0$ for a certain polynomial in the entries of the matrices. But a non-empty open subset of an irreducible set is dense. □

Proof of Theorem 5. Suppose $X_{abc}\in V(\langle\mathscr{S}'\rangle)\cap\mathscr{O}$ and $p\in\mathfrak{A}_T$.

First note that it is enough to show that $p(X_{abc})=0$ under the additional assumption that $X_{a\Sigma\Sigma}$ has no zero entries. For if some of the entries of $X_{a\Sigma\Sigma}$ are zero, we can find a matrix M arbitrarily close to I (in the Euclidean sense), whose columns add to 1, with $MX_{a\Sigma\Sigma}$ having no zero entry. Then defining $\widetilde{X}_{abc}$ by $\widetilde{X}_{abi}=MX_{abi}$, one checks that $\widetilde{X}_{abc}\in V(\langle\mathscr{S}'\rangle)\cap\mathscr{O}$, and so X_{abc} is in the closure of arrays satisfying the additional assumption as well.

Assuming, then, that $X_{a\Sigma\Sigma}$ has no zero entries, for any $(\kappa-1)$-tuple of $\kappa\times\kappa$ matrices $(M_1,\ldots,M_{\kappa-1})$, let $M_\kappa=I-\sum M_i$, and define an array Y_{abc} by $Y_{abi}=X_{ab\Sigma}M_i$. Let $\widetilde{p}(M_1,\ldots,M_{\kappa-1})=p(Y_{abc})$.

We first show $\widetilde{p}$ vanishes on all $(\kappa-1)$-tuples of simultaneously diagonalizable matrices. Writing $M_i=S^{-1}D_iS$, it is enough to consider those tuples where S has no row summing to 0, since these are dense in the full set. But then we may assume S has rows summing to 1. But for such M_i, by applying the ideas of Section 4, we see $Y_{abc}=E(\mathscr{M})$. In fact, $\mathscr{M}$ is composed of $\mathbf{p}_a=X^T_{a\Sigma\Sigma}$, $M_{af}=D_a^{-1}X_{ab\Sigma}S^{-1}$, $M_{fb}=S$, and M_{fc} with ith column coming from the diagonal of the D_i. Thus $\widetilde{p}(M_1,\ldots M_{\kappa-1})=p(E(\mathscr{M}))=0$.

Now by the preceeding lemma, $\widetilde{p}$ must therefore vanish on all $(\kappa-1)$-tuples of commuting matrices. In particular, for the commuting matrices $M_i=X^{-1}_{ab\Sigma}X_{abi}$, we find $p(X_{abc})=0$. □

Remark 4. Since $\mathscr{C}(m-1,m)$ is not irreducible for $m>4$, it is not hard to see that Theorem 5 cannot be extended to larger κ. Indeed explicit arrays $X_{abc}\notin V(\mathfrak{A}_T)$ can be constructed which satisfy all polynomials in $\mathscr{S}'$, but not d_c. However, whether Corollary 6 can be extended by a different proof to larger κ is not known.

7. Phylogenetic invariants for four or more taxa

In light of the last section, there are two reasonable goals in producing phylogenetic invariants. First, one might hope to produce as many invariants as possible, keeping in mind that their statistical behavior on noisy data is currently unknown and thus the more invariants we have to investigate, the more likely we may be to find ones that behave well. Second, one might hope to produce as small a set of invariants as possible that, on some set $\mathscr{D}$ of possible data, is in some sense sufficient to stand-in for all invariants (e.g., the notion of a strong set of invariants on $\mathscr{D}$).

In this section we pursue both goals for the general Markov model on an n-taxon tree. We first define a large number of invariants through commutation and symmetry relations and other approaches. We then single out a smaller subset of these which in the next section we will prove has a sufficiency property.

Consider the case of four taxa a, b, c, and d, related according to the tree T_1 in Fig. 2. Model parameters are specified as

$$\mathscr{M}=\{\mathbf{p}_a,M_{ae},M_{eb},M_{ef},M_{fc},M_{fd}\},$$

with the expected frequency of patterns at the terminal taxa given by $E_{abcd}=E_{abcd}(\mathscr{M})$.

Of course a large number of invariants can be found by considering subtrees connecting any three of the leaves. (By a *subtree* of T we mean a topological bifurcating tree whose leaves are a subset of those of T and which is obtained from T by deleting some leaves, edges and internal nodes, 'merging' edges if necessary.) For instance, the entries in the three-dimensional array $E_{abc\Sigma}$ must satisfy all of the invariants discussed earlier in the three taxa case. These polynomials, then, are simply the ones found earlier, with the sums appearing as entries of $X_{abc\Sigma}$ substituted in for the variables used earlier. Similarly one can substitute the entries of $X_{\Sigma bcd}$, $X_{a\Sigma cd}$, and $X_{ab\Sigma d}$ to obtain more invariants associated to 3-taxon subtrees.

However, we must expect other invariants not coming from 3-taxon subtrees, as a simple example shows. Since the only 3-taxon invariants in the $\kappa = 2$ case are multiples of the trivial one, any array X_{abcd} whose entries sum to 1 will satisfy all the 3-taxon subtree invariants. However, with only 11 scalar parameters in the $\kappa = 2$, 4-taxon model, the points of the form $E(\mathscr{M})$ in $\mathbb{C}^{16}$ must be on a variety of dimension at most 11. Thus there must be additional invariants. (In fact, from any 4-taxon invariant not arising from a 3-taxon invariant one could easily construct an example of a four-dimensional array that satisfies all invariants induced by 3-taxon subtrees, yet does not satisfy all invariants of the 4-taxon model.)

We will focus on producing invariants for n-taxon trees that do not come from considering smaller subtrees.

7.1. First construction: (3+)-taxon identities

Our first construction of invariants for n-taxon trees is one that has no analog in the 3-taxon case. As an example of it, for the 4-taxon tree T_1 note that

$$E_{aic\Sigma} = D_a M_{ae} C_{eb,i} M_{ef} M_{fc},$$

$$E_{a\Sigma c\Sigma} = D_a M_{ae} M_{ef} M_{fc},$$

$$E_{ai\Sigma d} = D_a M_{ae} C_{eb,i} M_{ef} M_{fd},$$

$$E_{a\Sigma\Sigma d} = D_a M_{ae} M_{ef} M_{fd}.$$

Thus

$$E_{aic\Sigma} E^{-1}_{a\Sigma c\Sigma} = E_{ai\Sigma d} E^{-1}_{a\Sigma\Sigma d}$$

and for each choice of i we obtain κ^2 invariants from the entries of

$$X_{aic\Sigma} \mathrm{Cof}(X_{a\Sigma c\Sigma})^T \det(X_{a\Sigma\Sigma d}) = X_{ai\Sigma d} \mathrm{Cof}(X_{a\Sigma\Sigma d})^T \det(X_{a\Sigma c\Sigma}). \tag{11}$$

Note that these invariants are of degree 2κ.

In fact, the polynomial identities obtained from Eq. (11) are topologically informative; that is, identity (11) holds precisely because a and b are neighbors. To see this, suppose we consider the tree T_2 of Fig. 2 in which a and c are neighbors, with parameters $\mathscr{M} = \{\mathbf{p}_a, M_{ae}, M_{ec}, M_{ef}, M_{fb}, M_{fd}\}$. Then we find:

$$E_{aic\Sigma} = D_a M_{ae} C'_{eb,i} M_{ec},$$

$$E_{a\Sigma c\Sigma} = D_a M_{ae} M_{ec},$$

$$E_{ai\Sigma d} = D_a M_{ae} M_{ef} C_{fb,i} M_{fd},$$

$$E_{a\Sigma\Sigma d} = D_a M_{ae} M_{ef} M_{fd},$$

where $C'_{eb,i}$ is the diagonal matrix constructed from the ith column of the product $M_{ef}M_{fb}$. Thus

$$E_{aic\Sigma} E^{-1}_{a\Sigma c\Sigma} = (D_a M_{ae}) C'_{eb,i} (D_a M_{ae})^{-1},$$

$$E_{ai\Sigma d} E^{-1}_{a\Sigma\Sigma d} = (D_a M_{ae}) M_{ef} C_{fb,i} M^{-1}_{ef} (D_a M_{ae})^{-1}.$$

Generically at least, $C'_{eb,i} \neq M_{ef} C_{fb,i} M^{-1}_{ef}$, as one can choose parameters so the matrix product on the right is not diagonal. Thus the generic $E_{abcd}(\mathscr{M})$ for tree T_2 will not satisfy Eq. (11).

Returning to tree T_1, there are invariants analogous to those in Eq. (11) again specifying that a and b are neighbors, but in which the roles of a and b are reversed:

$$X_{ibc\Sigma} \mathrm{Cof}(X_{\Sigma bc\Sigma})^T \det(X_{\Sigma b\Sigma d}) = X_{ib\Sigma d} \mathrm{Cof}(X_{\Sigma b\Sigma d})^T \det(X_{\Sigma bc\Sigma}),$$

as well as invariants specifying that c and d are neighbors:

$$X^T_{a\Sigma ci} \mathrm{Cof}(X_{a\Sigma c\Sigma}) \det(X_{\Sigma bc\Sigma}) = X^T_{\Sigma bci} \mathrm{Cof}(X_{\Sigma bc\Sigma}) \det(X_{a\Sigma c\Sigma}),$$

$$X^T_{a\Sigma id} \mathrm{Cof}(X_{a\Sigma\Sigma d}) \det(X_{\Sigma b\Sigma d}) = X^T_{\Sigma bid} \mathrm{Cof}(X_{\Sigma b\Sigma d}) \det(X_{a\Sigma\Sigma d}).$$

This construction of invariants generalizes to an n-taxon tree T as follows: Choose three of the taxa, and denote them by a_1, a_2, and a_3. Let v be the internal node of T that is the only internal node of the 3-taxon subtree relating these a_i. Let $a_4, \ldots, a_m$ denote those taxa for which the path $a_i \to a_1$ in T does not pass through v. (The construction assumes at least one such a_i exists, and thus that $n \geqslant 4$.) Let $a_{m+1}, \ldots, a_n$ denote the remaining taxa. Then order the indexing of the expected frequency array so taxa appear in the order $a_1, a_2, \ldots, a_n$.

Now for any $1 \leqslant i_4, \ldots, i_m \leqslant \kappa$, similar reasoning to the above yields

$$X_{a_1 a_2 \Sigma i_4 \ldots i_m \Sigma \ldots \Sigma} \mathrm{Cof}(X_{a_1 a_2 \Sigma \ldots \Sigma})^T \det(X_{a_1 \Sigma a_3 \Sigma \ldots \Sigma}) = X_{a_1 \Sigma a_3 i_4 \ldots i_m} \mathrm{Cof}(X_{a_1 \Sigma a_3 \Sigma \ldots \Sigma})^T \det(X_{a_1 a_2 \Sigma \ldots \Sigma}). \quad (12)$$

Varying the choices of a_1, a_2, a_3 yields many more invariants, all of degree 2κ. These are also topologically informative for any $n \geqslant 4$. That at least some of them are topologically informative is perhaps most easily seen by summing equations of the form (12) over all indices $i_5, \ldots, i_m$ which reduces them to the invariants (11) for a 4-taxon subtree.

7.2. Second construction: (4+)-taxon identities

A similar source of invariants which did not arise in the 3-taxon case is the 4-point condition, with the essential idea appearing in [1]. For the 4-taxon tree T_1 with parameters $\mathscr{M} = \{\mathbf{p}_a, M_{ae}, M_{eb}, M_{ef}, M_{fc}, M_{fd}\}$, for convenience define

$$\mathbf{p}_e = \mathbf{p}_a M_{ae}, \quad \mathbf{p}_b = \mathbf{p}_e M_{eb}, \quad M_{be} = \mathrm{diag}(\mathbf{p}_b)^{-1} M^T_{eb} \mathrm{diag}(\mathbf{p}_e), \quad D_b = \mathrm{diag}(\mathbf{p}_b).$$

Then

$$E_{a\Sigma\Sigma d} = D_a M_{ae} M_{ef} M_{fd}, \quad E_{a\Sigma c\Sigma} = D_a M_{ae} M_{ef} M_{fc},$$

$$E_{\Sigma b\Sigma d} = D_b M_{be} M_{ef} M_{fd}, \quad E_{\Sigma bc\Sigma} = D_b M_{be} M_{ef} M_{fc}$$

and so

$$E_{a\Sigma\Sigma d}(E_{\Sigma b\Sigma d})^{-1} = E_{a\Sigma c\Sigma}(E_{\Sigma bc\Sigma})^{-1}. \tag{13}$$

Thus

$$X_{a\Sigma\Sigma d}\mathrm{Cof}(X_{\Sigma b\Sigma d})^T \det(X_{\Sigma bc\Sigma}) = X_{a\Sigma c\Sigma}\mathrm{Cof}(X_{\Sigma bc\Sigma})^T \det(X_{\Sigma b\Sigma d}) \tag{14}$$

yields invariants of degree 2κ. Variations on this, by interchanging the role of the pair a and b with c and d, or taking the matrix inverse of each side of Eq. (13) are possible also.

Note that by taking the determinant of each side of Eq. (13) and rearranging terms we obtain

$$\det(E_{a\Sigma\Sigma d})\det(E_{\Sigma bc\Sigma}) = \det(E_{a\Sigma c\Sigma})\det(E_{\Sigma b\Sigma d}). \tag{15}$$

This is nothing more than an exponentiated form of the 4-point condition applied to the log-det distance. However the invariants in Eq. (14) are potentially more powerful than the single invariant to which Eq. (15) gives rise.

Actually, more careful reasoning shows one can strengthen Eq. (14) to

$$X_{ai\Sigma d}\mathrm{Cof}(X_{\Sigma b\Sigma d})^T \det(X_{\Sigma bc\Sigma}) = X_{aic\Sigma}\mathrm{Cof}(X_{\Sigma bc\Sigma})^T \det(X_{\Sigma b\Sigma d}). \tag{16}$$

Notice that summing Eq. (16) over i gives Eq. (14) again.

This construction of invariants generalizes for the n-taxon case as follows: For a tree T choose any four taxa and denote them by a_1, a_2, a_3, a_4 in such a way that in the 4-taxon subtree relating them a_1 and a_2 are neighbors, and so a_3 and a_4 are also neighbors. Let $a_5, \ldots, a_m$ denote those taxa other than a_1, a_2, a_3, a_4 for which the path $a_i \to a_1$ first joins the subtree anywhere except along the subtree edges containing a_3 and a_4. (If no such taxa exist, the construction will still make sense.) Let $a_{m+1}, \ldots, a_n$ denote the remaining taxa. Then order the indexing of the expected frequency array so taxa appear in the order $a_1, a_2, \ldots, a_n$.

Now for any $1 \leqslant i_2, i_5, \ldots, i_m \leqslant \kappa$ similar reasoning to the above yields

$$\begin{aligned} &X_{a_1 i_2 \Sigma a_4 i_5 \ldots i_m \Sigma \ldots \Sigma}\mathrm{Cof}(X_{\Sigma a_2 \Sigma a_4 \Sigma \ldots \Sigma})^T \det(X_{\Sigma a_2 a_3 \Sigma\Sigma \ldots \Sigma}) \\ &\quad = X_{a_1 i_2 a_3 \Sigma i_5 \ldots i_m \Sigma \ldots \Sigma}\mathrm{Cof}(X_{\Sigma a_2 a_3 \Sigma\Sigma \ldots \Sigma})^T \det(X_{\Sigma a_2 \Sigma a_4 \Sigma \ldots \Sigma}). \end{aligned} \tag{17}$$

Note that summing Eq. (17) over each of $i_2, i_5, \ldots, i_m$ produces Eq. (14) applied to the 4-taxon subtree.

Interestingly, our first construction of invariants by (3+)-taxon identities can be viewed as a degenerate case of the 4-point construction, with $a_1 = a_2$.

Varying the choices of a_1, a_2, a_3, a_4 yields many more invariants, all of degree 2κ. One can show that these will be topologically informative.

7.3. Third construction: commutation relations

Numerous invariants can be found by considering commutation relations, as in the 3-taxon case. For instance for the 4-taxon tree T_1, for any $1 \leqslant i, j, k, l \leqslant \kappa$

$$E_{abij} = D_a M_{ae} P_{e,cd;ij} M_{eb},$$

$$E_{abkl} = D_a M_{ae} P_{e,cd;kl} M_{eb},$$

$$E_{ab\Sigma\Sigma} = D_a M_{ae} M_{eb},$$

where $P_{e,cd;ij}$ is a diagonal matrix whose diagonal entries give the conditional probabilities that each base at e becomes i at c and j at d. That is,

$$P_{e,cd;ij}(k,k) = \sum_{m=1}^{\kappa} M_{ef}(k,m)M_{fc}(m,i)M_{fd}(m,j).$$

Thus

$$E_{abij}E_{ab\Sigma\Sigma}^{-1}E_{abkl} = E_{abkl}E_{ab\Sigma\Sigma}^{-1}E_{abij}$$

and so we have invariants from the entries of

$$X_{abij}\mathrm{Cof}(X_{ab\Sigma\Sigma})^T X_{abkl} = X_{abkl}\mathrm{Cof}(X_{ab\Sigma\Sigma})^T X_{abij}. \tag{18}$$

This used only that a and b were neighbors, so similar invariants are obtained focusing on c and d. One can also show that these invariants are topologically informative. Note that invariants in this class are all of degree $\kappa+1$. Thus they are of lower degree than those produced by either of the other constructions.

This construction, analogous to that of Eq. (8) in the 3-taxon case, generalizes to the n-taxon tree as follows: For an n-taxon tree T, choose any two taxa, and denote them by a_1 and a_2. Let v be any internal node of T that lies on the path $a_1 \to a_2$. Let $a_3, \ldots, a_m$ denote those taxa for which the path $a_i \to a_1$ first joins the path $a_2 \to a_1$ at v. Let $a_{m+1}, \ldots, a_n$ denote the remaining taxa. Then order the indexing of the expected frequency array so taxa appear in the order $a_1, a_2, \ldots, a_n$.

Now for any $1 \leqslant i_3, \ldots, i_m \leqslant \kappa$ and $1 \leqslant j_3, \ldots, j_m \leqslant \kappa$, similar reasoning to the above yields

$$X_{a_1a_2i_3\ldots i_m\Sigma\ldots\Sigma}\mathrm{Cof}(X_{a_1a_2\Sigma\ldots\Sigma})^T X_{a_1a_2j_3\ldots j_m\Sigma\ldots\Sigma} = X_{a_1a_2j_3\ldots j_m\Sigma\ldots\Sigma}\mathrm{Cof}(X_{a_1a_2\Sigma\ldots\Sigma})^T X_{a_1a_2i_3\ldots i_m\Sigma\ldots\Sigma}. \tag{19}$$

There is also an analog of the 3-taxon equation (9) for the n-taxon tree: Choose any three taxa and denote them by a_1, a_2, and a_3. Let v be the internal node of T that is the only internal node of the 3-taxon subtree relating these a_i. Let $a_4, \ldots, a_m$ denote those taxa except a_3 for which the path $a_i \to a_1$ in T first joins the path $a_2 \to a_1$ at v. Similarly, let $a_{m+1}, \ldots, a_l$ denote those taxa except a_2 for which the path $a_i \to a_1$ in T first joins the path $a_3 \to a_1$ at v. (Either of these sets may be empty, but if both are, then the construction reduces to one for a 3-taxon subtree.) Let $a_{l+1}, \ldots, a_n$ denote the remaining taxa. Then order the indexing of the expected frequency array so taxa appear in the order $a_1, a_2, \ldots, a_n$.

Now for any $1 \leqslant i_3, \ldots, i_m \leqslant \kappa$ and $1 \leqslant j_2, j_{m+1}, \ldots, j_l \leqslant \kappa$,

$$\begin{aligned} &X_{a_1a_2i_3\ldots i_m\Sigma\ldots\Sigma}\mathrm{Cof}(X_{a_1a_2\Sigma\ldots\Sigma})^T X_{a_1j_2a_3\Sigma\ldots\Sigma j_{m+1}\ldots j_l\Sigma\ldots\Sigma}\mathrm{Cof}(X_{a_1\Sigma a_3\Sigma\ldots\Sigma})^T \\ &\quad = X_{a_1j_2a_3\Sigma\ldots\Sigma j_{m+1}\ldots j_l\Sigma\ldots\Sigma}\mathrm{Cof}(X_{a_1\Sigma a_3\Sigma\ldots\Sigma})^T X_{a_1a_2i_3\ldots i_m\Sigma\ldots\Sigma}\mathrm{Cof}(X_{a_1a_2\Sigma\ldots\Sigma})^T. \end{aligned}$$

Varying the choices of a_1, a_2 and a_3 yields many more invariants, all of degree 2κ. One can also show these are topologically informative.

7.4. Fourth construction: symmetry relations

The construction of invariants through symmetry relations for the 3-taxon tree generalizes as follows to the n-taxon tree: Choose three of the taxa, and denote them by a_1, a_2, and a_3. Let v be the internal node of T that is the only internal node of the 3-taxon subtree relating these a_i. Let $a_4, \ldots, a_m$ denote those taxa for which the path $a_i \to a_1$ in T does not pass through v. (If no such taxa exist, the construction reduces to a 3-taxon subtree construction.) Let $a_{m+1}, \ldots, a_n$ denote the

remaining taxa. Then order the indexing of the expected frequency array so taxa appear in the order $a_1, a_2, \ldots, a_n$.

Now for any $1 \leqslant i_2, i_3, i_4, \ldots, i_m \leqslant \kappa$, similar reasoning to the 3-taxon case yields

$$X_{a_1 a_2 i_3 i_4 \ldots i_m \Sigma \ldots \Sigma} \mathrm{Cof}(X_{\Sigma a_2 a_3 \Sigma \ldots \Sigma}) X^T_{a_1 i_2 a_3 i_4 \ldots i_m \Sigma \ldots \Sigma} = X_{a_1 i_2 a_3 i_4 \ldots i_m \Sigma \ldots \Sigma} \mathrm{Cof}(X_{\Sigma a_2 a_3 \Sigma \ldots \Sigma})^T X^T_{a_1 a_2 i_3 i_4 \ldots i_m \Sigma \ldots \Sigma}.$$

Varying the choices of a_1, a_2, a_3 yields many more invariants, all of degree $\kappa + 1$. At least some of these are topologically informative for $n \geqslant 4$.

Just as in the 3-taxon case, there are variations on the constructions above in which, rather than inverting the specified matrices, one instead inverts any linear combination of certain matrices that sum to produce it. The comments in Remark 2 are relevant here as well.

Obviously one can also repeat in the n-taxon case the non-explicit saturation and radical steps done for 3-taxon invariants, and possibly obtain more invariants as a result.

7.5. A small set of invariants

For further study in the next section, for an n-taxon tree T we choose a particular subset $\mathscr{S}'(T)$ of the invariants above. Actually, as there will be some freedom in the definition of $\mathscr{S}'(T)$, there is a family of such $\mathscr{S}'(T)$.

The definition is an inductive one on the number of taxa n.

For the 3-taxon tree T, let $\mathscr{S}'(T) = \mathscr{S}' = \mathscr{S}_0 \cup \mathscr{S}_c$ as in Section 6, where c is an arbitrarily chosen leaf.

Now for an n-taxon tree T, choose any two taxa which are neighbors and denote them a_1 and a_2. Let T^- denote the $(n-1)$-taxon tree obtained by deleting a_2 and the edge $a_2 \leftrightarrow v$ leading from it, and replacing the two other edges $v \leftrightarrow a_1$ and $v \leftrightarrow w$ that connect to v with $a_1 \leftrightarrow w$. Now $\mathscr{S}'(T^-)$ is already defined, so let

$$\mathscr{S}^- = \{\widetilde{p}(X_{a_1 a_2 \ldots a_n}) \,|\, \widetilde{p}(X_{a_1 a_2 \ldots a_n}) = p(X_{a_1 \Sigma a_3 \ldots a_n}) \text{ for some } p \in \mathscr{S}'(T^-)\}.$$

Let $\mathscr{S}^+$ be all the invariants produced by Eq. (19), for all choices of i_k, j_l, but with the fixed choice of a_1 and a_2. Finally, let $\mathscr{S}'(T) = \mathscr{S}^+ \cup \mathscr{S}^-$.

Note $\mathscr{S}'(T)$ is composed of the trivial invariant and only certain of the invariants arising from the commutation relations construction which are of degree $\kappa + 1$.

8. Sufficiency of invariants arising from commutation relations: near-diagonal arrays

Producing an analog for the n-taxon case of the 3-taxon results in Section 6 would of course be desirable, though we have not done so. However, while those results identify powerful sets of invariants on certain sets, they also point out that there are points satisfying the invariants that do not come from any choice of model parameters, whether stochastic or complex. (See Section 9 for explicit examples.) In other words, polynomial invariants alone are not capable of distinguishing points of the form $E(\mathscr{M})$.

In this section we prove a different type of sufficiency result for the invariants constructed in this paper. In brief, there is an open set on which the invariants can test whether a point is of the form $E(\mathscr{M})$. Furthermore, this result requires no restriction on n or κ.

Again, we need some new terminology.

Definition. Let T be an n-taxon tree rooted at taxon a_1. Suppose $\mathscr{D} \subseteq \mathbb{C}^{\kappa^n}$ and $\mathscr{R} \subseteq \mathbb{C}[X_{a_1a_2\ldots a_n}]$. Then we say $\mathscr{R}$ is a *parameter-strong set of invariants* on $\mathscr{D}$ for the tree T if

$$X_{a_1a_2\ldots a_n} \in V(\langle \mathscr{R} \rangle) \cap \mathscr{D}$$

implies

$$X_{a_1a_2\ldots a_n} = E_{a_1a_2\ldots a_n}(\mathscr{M})$$

for some choice of complex parameters $\mathscr{M}$.

Informally, $\mathscr{R}$ is parameter-strong on $\mathscr{D}$ means any point in $\mathscr{D}$ at which the elements of $\mathscr{R}$ vanish 'comes from' some complex choice of model parameters.

Note that a choice of a terminal taxon as the root of T is implicit in the notion of a parameter-strong set of invariants, since the designation of the root is necessary for defining the parameters. Indeed, one can construct examples of even two-dimensional arrays X_{ab} for which X_{ab} is of the form $E_{ab}(\mathscr{M})$ if the 2-taxon tree is rooted at a, but not if it is rooted at b.

We immediately see

Proposition 10. *If a set R is a parameter-strong set of invariants on $\mathscr{D}$, then it is a strong set of invariants on $\mathscr{D}$.*

Definition. An n-dimensional array X is said to be *diagonal* if $X(i_1, i_2, \ldots, i_n) = 0$ for all n-tuples $(i_1, i_2, \ldots, i_n)$ except possibly those with $i_1 = i_2 = \cdots = i_n$.

Let T be any n-taxon tree. If parameters for the general Markov model on T are chosen so the identity matrix is assigned to each edge and some base distribution vector is chosen for taxon a_1, then the corresponding expected pattern frequency array will be diagonal. Conversely, given an n-dimensional $\kappa \times \cdots \times \kappa$ diagonal array of non-negative numbers that sum to 1, there are stochastic parameters on T, with the base distribution vector composed of the diagonal entries and all Markov matrices the identity, such that the array is the expected pattern frequency array for these parameters. Informally, diagonal arrays are the expected frequency arrays of models in which no mutations occur, and thus they fit all topological trees.

This motivates the following definition. Note that we strengthen the contents of the last paragraph slightly by requiring all bases appear with positive probability.

Definition. An n-dimensional $\kappa \times \kappa \times \cdots \times \kappa$ array X is *phylogenetically trivial* if it is diagonal, with positive entries on the diagonal that sum to 1.

Remark 5. In biological circumstances expected frequency arrays, whether theoretical or estimated from data, are typically close to phylogenetically trivial ones. In the theoretical framework this is because model parameters describe base substitution processes where most sites are left unchanged. When working with data, this is because most sites must be identical in order to identify sequences as related and to align them.

We first focus on the 3-taxon case, with $\mathscr{S}' = \mathscr{S}_0 \cup \mathscr{S}_c \subseteq \mathscr{S}$ the sets of invariants defined earlier.

Theorem 11. *For the* 3*-taxon tree, there exists an open set* $\mathcal{O} \subset \mathbb{C}^{\kappa^3}$ *which contains all phylogenetically trivial arrays and on which* $\mathcal{S}'$ *is parameter-strong regardless of which leaf is taken to be the root. Furthermore, if* $X_{abc} \in \mathcal{O}$*, and for a fixed choice of a terminal taxon as the root,* $X_{abc} = E_{abc}(\mathcal{M}) = E_{abc}(\mathcal{M}')$*, then* $\mathcal{M}' = \mathcal{M}^{\sigma}$ *for some permutation* σ*.*

While we do not identify such a set $\mathcal{O}$ in a quantitative way, the existence of such a set means that arrays that satisfy the invariants in $\mathcal{S}'$ and are 'sufficiently-close' to phylogenetically trivial ones arise from model parameters, and these parameters are 'essentially' unique.

To prove Theorem 11 we need the following technical lemma, whose proof we defer to Appendix A.

Lemma 12. *For* $i = 1, \ldots, \kappa$*, let* $E_{i,i}$ *denote the* $\kappa \times \kappa$ *matrix with all entries* 0*, except the* (i,i) *entry which is* 1*. Then for any open neighborhood* $\mathcal{Q} \subseteq \mathbb{C}^{\kappa^2}$ *of the identity matrix* I*, there exist open neighborhoods* $\mathcal{N}_i \subset \mathbb{C}^{\kappa^2}$ *of* $E_{i,i}$ *with the following property: If* $M_i \in \mathcal{N}_i$*,* $i = 1, \ldots, \kappa$ *are commuting matrices, then there exists a matrix* $S \in \mathcal{Q}$ *with rows summing to* 1 *such that* SM_iS^{-1} *is diagonal for all* i*. Furthermore, the simultaneous left eigenspaces of the* M_i *are all one-dimensional.*

Proof of Theorem 11. Let X^0_{abc} denote any phylogenetically trivial array. It is enough to find a neighborhood $\mathcal{O}$ of X^0_{abc} on which $\mathcal{S}'$ is parameter-strong and the other statements of the theorem hold.

Let $\mathcal{O}_1 \ni X^0_{abc}$ and $\mathcal{Q} \ni I$ be open sets with the properties that if $X_{abc} \in \mathcal{O}_1$ and $S \in \mathcal{Q}$ then: (1) $X_{ab\Sigma}$ is invertible, (2) $X_{a\Sigma\Sigma}$, $X_{\Sigma\Sigma c}$ have all non-zero entries, and 3) $X^T_{\Sigma b\Sigma}S^{-1}$ has all non-zero entries. Such sets exist since the first two conditions are met by requiring polynomial inequalities in the entries of X_{abc} hold, while for the third we know $X^T_{\Sigma b\Sigma}S^{-1}$ is continuous in S and X_{abc}, and has non-zero entries at $S = I$, $X_{abc} = X^0_{abc}$.

For arrays $X_{abc} \in \mathcal{O}_1 \cap V(\mathfrak{I}')$, we can define the matrices

$$Z_{c;i} = (X_{ab\Sigma})^{-1}X_{abi}$$

and the $Z_{c;i}$ will commute.

For the chosen $\mathcal{Q}$, let $\mathcal{N}_i$ be the open sets whose existence is asserted by Lemma 12. We will now find an open set $\mathcal{O} \ni X^0_{abc}$ such that $X_{abc} \in \mathcal{O}$ implies each $Z_{c;i} \in \mathcal{N}_i$. To do this, note the map

$$\varphi : X_{abc} \mapsto (Z_{c;1}, \ldots, Z_{c;\kappa})$$

is continuous on $\mathcal{O}_1$. Moreover, $\varphi(X^0_{abc}) = (E_{1,1}, \ldots, E_{\kappa,\kappa})$. Thus $\mathcal{O} = \varphi^{-1}(\mathcal{N}_1 \times \cdots \times \mathcal{N}_\kappa)$ has the desired properties.

Now if $X_{abc} \in \mathcal{O} \cap V(\mathfrak{I}')$, then Lemma 12 implies that the $Z_{c;i}$ are simultaneously diagonalizable by a matrix $S \in \mathcal{Q}$ whose rows sum to 1. Using the ideas of Section 4, this allows us to deduce parameters $\mathcal{M}$ (rooted at a) with $X_{abc} = E_{abc}(\mathcal{M})$. Since Lemma 12 also implies the eigenspaces used in the deduction of $\mathcal{M}$ are all one-dimensional, $\mathcal{M}$ is uniquely determined up to the action of a permutation.

Note also that $\mathbf{p}_b = X^T_{\Sigma b\Sigma}$, $\mathbf{p}_c = X^T_{\Sigma\Sigma c}$, and $\mathbf{p}_f = \mathbf{p}_b M^{-1}_{fb} = X^T_{\Sigma b\Sigma}S^{-1}$ have only non-zero entries. Thus proceeding as in Section 3, from $\mathcal{M}$ we can deduce parameters $\mathcal{M}_b$ (resp., $\mathcal{M}_c$) with root at taxon b (resp., c), unique up to the action of a permutation, so that $X_{abc} = E_{abc}(\mathcal{M}_b) = E_{abc}(\mathcal{M}_c)$. □

The result also extends to n-taxon trees.

Theorem 13. *For an n-taxon tree T, let $\mathscr{S}'(T)$ denote the set of invariants defined in Section 7. Then there exists an open set $\mathscr{O} \subset \mathbb{C}^{\kappa^n}$ which contains all phylogenetically trivial arrays and on which $\mathscr{S}'(T)$ is parameter-strong, regardless of which leaf is taken as the root. Moreover, for a fixed choice of root, if $X_{a_1\ldots a_n} \in \mathscr{O}$ and $X_{a_1\ldots a_n} = E_{a_1\ldots a_n}(\mathscr{M}) = E_{a_1\ldots a_n}(\mathscr{M}')$, then $\mathscr{M}'$ can be obtained from $\mathscr{M}$ through the action of some choice of permutations at each internal node of T.*

Proof. We proceed by induction, with the base case $n = 3$ being Theorem 11. For future use, let $\mathscr{O}_3$ denote the open set constructed in Theorem 11 for this 3-taxon case.

For T an n-taxon tree with $n > 3$, let a_1 and a_2 be the two neighboring leaves of T that were used in the inductive definition of $\mathscr{S}'(T)$, with v the vertex both are connected to by an edge. Now the map

$$\varphi_1 : X_{a_1a_2\ldots a_n} \mapsto X_{a_1a_2a_3\Sigma\ldots\Sigma}$$

is continuous, so $\mathscr{Q}_1 = \varphi_1^{-1}(\mathscr{O}_3)$ is open, and is easily seen to contain all phylogenetically trivial arrays, since $\mathscr{O}_3$ has that property.

With T^- the $(n-1)$-taxon tree used in the definition of $\mathscr{S}'(T)$, let $\mathscr{O}_{T^-}$ be an open set which the inductive hypothesis asserts has the properties stated in the theorem. The map

$$\varphi_2 : X_{a_1a_2\ldots a_n} \mapsto X_{a_1\Sigma a_3\ldots a_n}$$

is continuous, so $\mathscr{Q}_2 = \varphi_2^{-1}(\mathscr{O}_{T^-})$ is open and contains all phylogenetically trivial arrays.

Let $\mathscr{O}_T = \mathscr{Q}_1 \cap \mathscr{Q}_2$. To show $\mathscr{S}'(T)$ is parameter-strong on $\mathscr{O}_T$ for parameters rooted at a_1, suppose

$$X_{a_1\ldots a_n} \in V(\langle \mathscr{S}'(T) \rangle) \cap \mathscr{O}_T.$$

Now summing Eq. (19) over all choices of $i_4, \ldots, i_n$ and $j_4, \ldots, j_n$ shows $\varphi_1(X_{a_1a_2\ldots a_n})$ satisfies Eq. (8) for each choice of i and j at a_3. So, since $\varphi_1(X_{a_1a_2\ldots a_n})$ satisfies the trivial invariant and lies in $\mathscr{O}_3$, by Theorem 11 we know

$$\varphi_1(X_{a_1a_2\ldots a_n}) = E_{a_1a_2a_3}(\mathscr{M}_3),$$

for some $\mathscr{M}_3 = \{\mathbf{p}_{a_1}, M_{a_1v}, M_{va_2}, M_{va_3}\}$. Moreover, $\mathscr{M}_3$ is uniquely determined up to the action of a permutation.

Now $X^{-1}_{a_1a_2\Sigma\ldots\Sigma}$ exists since $\varphi_1(X_{a_1a_2\ldots a_n}) \in \mathscr{O}_3$. Thus since $X_{a_1\ldots a_n}$ satisfies Eq. (19) for all i_k, j_l, one sees that all the matrices

$$Y_{i_3\ldots i_n} = X_{a_1a_2i_3\ldots i_n}(X_{a_1a_2\Sigma\ldots\Sigma})^{-1}$$

must commute.

Now the $Y_{i_3\ldots i_n}$ must also commute with all the

$$Y_{i_3} = X_{a_1a_2i_3\Sigma\ldots\Sigma}(X_{a_1a_2\Sigma\ldots\Sigma})^{-1},$$

which are simply sums of $Y_{i_3\ldots i_n}$. However, since the parameters $\mathscr{M}_3$ are unique up to the action of permutations, we know the Y_{i_3} have κ one-dimensional simultaneous eigenspaces, and are diagonalized by

$$Y_{i_3} = (D_{a_1}M_{a_1v})C_{va_3;i_3}(D_{a_1}M_{a_1v})^{-1}.$$

Thus one also has that

$$Y_{i_3\ldots i_n} = (D_{a_1}M_{a_1v})C_{i_3\ldots i_n}(D_{a_1}M_{a_1v})^{-1},$$

for some diagonal matrices $C_{i_3\ldots i_n}$. Since $X_{a_1a_2\Sigma\ldots\Sigma} = D_{a_1}M_{a_1v}M_{va_2}$, we find

$$X_{a_1a_2i_3\ldots i_n} = D_{a_1}M_{a_1v}C_{i_3\ldots i_n}M_{va_2}. \tag{20}$$

By the definition of $\mathscr{S}'(T)$, we know $\varphi_2(X_{a_1a_2\ldots a_n})$ satisfies all invariants in $\mathscr{S}'(T^-)$ and also lies in $\mathscr{O}_{T^-}$. Thus by induction we know

$$\varphi_2(X_{a_1a_2\ldots a_n}) = E_{a_1a_3\ldots a_n}(\mathscr{M}_{T^-}),$$

for some $\mathscr{M}_{T^-} = \{\mathbf{p}_{a_1}, M_{a_1w}, \ldots\}$, unique up to action of permutations at internal nodes of T^-. Here w is as in the definition of $\mathscr{S}'(T^-)$. Also this $\mathbf{p}_{a_1}$ must agree with that in $\mathscr{M}_3$, since both are given by $X^T_{a_1\Sigma\ldots\Sigma}$.

Let

$$M_{vw} = M^{-1}_{a_1v}M_{a_1w}.$$

Finally, let $\mathscr{M}_T$ be comprised of $\mathbf{p}_{a_1}$, M_{a_1v}, M_{va_2}, M_{vw}, and all parameters in $\mathscr{M}_{T^-}$ except M_{a_1w}.

With this choice of parameters, we claim $X_{a_1\ldots a_n} = E_{a_1\ldots a_n}(\mathscr{M}_T)$. To establish this, we need only show

$$X_{a_1a_2i_3\ldots i_n} = E_{a_1a_2i_3\ldots i_n}(\mathscr{M}_T),$$

or that

$$(D_{a_1}M_{a_1v})^{-1}X_{a_1a_2i_3\ldots i_n}M^{-1}_{va_2} = (D_{a_1}M_{a_1v})^{-1}E_{a_1a_2i_3\ldots i_n}(\mathscr{M}_T)M^{-1}_{va_2}.$$

But by Eq. (20) for the left hand side, and the structure of the expected frequency array for the right hand side, both sides of this equation are diagonal matrices. Thus to prove they are equal, its enough to show equality upon multiplying on the right by a column vector of 1s. Now the Markov matrix M_{va_2} has this vector as right eigenvector of eigenvalue 1, so $M^{-1}_{va_2}$ does as well. Thus it is enough to show

$$(D_{a_1}M_{a_1v})^{-1}X_{a_1a_2i_3\ldots i_n}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = (D_{a_1}M_{a_1v})^{-1}E_{a_1a_2i_3\ldots i_n}(\mathscr{M}_T)\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix},$$

which is simply

$$(D_{a_1}M_{a_1v})^{-1}X_{a_1\Sigma i_3\ldots i_n} = (D_{a_1}M_{a_1v})^{-1}E_{a_1\Sigma i_3\ldots i_n}(\mathscr{M}_T).$$

But this follows from the fact that

$$X_{a_1\Sigma a_3\ldots a_n} = E_{a_1a_3\ldots a_n}(\mathscr{M}_{T^-}) = E_{a_1\Sigma a_3\ldots a_n}(\mathscr{M}_T)$$

by the definition of $\mathscr{M}_T$.

The uniqueness of parameters up to the actions of permutations follows easily from that property for the 3-taxon and $(n-1)$-taxon subtrees used in the argument.

While this proves all the claims assuming parameters are rooted at a_1, it is straightforward to modify the arguments for other choices of a leaf to serve as the root. □

Remark 6. For biological understanding, Theorems 11, 13 are of course primarily of interest when applied to an arrays X of non-negative real numbers. In that case, one can modify the proofs

above to show the eigenvalues and eigenvectors of the $Z_{c;i}$ in Theorem 11 must be real, and so the parameters $\mathscr{M}$ must be real as well. Unfortunately, that is not sufficient to conclude that the parameters must be stochastic. See Section 9 for an example.

9. Some cautionary examples

In this section we give examples of specific arrays that illustrate the pitfalls of using phylogenetic invariants to attempt to conclude that an array is of the form $E(\mathscr{M})$ for some choice of $\mathscr{M}$. While Theorem 13 tells us that can be valid for arrays sufficiently close to diagonal, in general it is not correct, and even when it is, we have no assurances that the parameters are stochastic.

Though our examples are contrived, and perhaps not of the sort that would appear in biological applications, nevertheless they are instructive.

Consider the $2\times2\times2$ array X_{abc} with

$$X_{ab1} = \begin{pmatrix} 0.25 & \epsilon \\ 0 & 0.25 \end{pmatrix} \quad X_{ab2} = \begin{pmatrix} 0.25 & -\epsilon \\ 0 & 0.25 \end{pmatrix},$$

for any $\epsilon \neq 0$. Then since for 3 taxa and $\kappa = 2$, the phylogenetic invariant ideal is generated by the trivial invariant, we have that $X_{abc} \in V(\mathfrak{A}_T)$. However, $(X_{ab\Sigma})^{-1}X_{abi} = 2X_{abi}$ is not diagonalizable, and thus by Section 4, $X_{abc} \neq E_{abc}(\mathscr{M})$ for any $\mathscr{M}$.

While in this example X_{abc} had a negative entry which is obviously not biologically meaningful, we can eliminate that feature and preserve the essential nature of the example by letting

$$X_{ab1} = \begin{pmatrix} 0.25 & \epsilon \\ 0 & 0.25 \end{pmatrix}\begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}, \quad X_{ab2} = \begin{pmatrix} 0.25 & -\epsilon \\ 0 & 0.25 \end{pmatrix}\begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}.$$

Choosing any $1/12 > \epsilon > 0$ makes all entries of X_{abc} positive, yet $(X_{ab\Sigma})^{-1}X_{ab1}$ is again not diagonalizable.

One can construct similar examples with larger κ, using larger non-diagonalizable matrices. If, for instance, these have a single Jordan block, one can relatively easily show that the point X_{abc} so constructed is in the closure of those points of the form $E_{abc}(\mathscr{M})$. Thus we can ensure X_{abc} lies in $V(\mathfrak{A}_T)$ even though we do not explicitly know all invariants for $\kappa \geqslant 3$.

One can even extend this example to more taxa by beginning with such a 3-taxon example and then adding to the tree additional edges (with Markov matrices I, for instance) off of taxon c, thus making c an internal node in the final tree. To see that such an example lies in $V(\mathfrak{A}_T)$, we need only note that it lies in the closure of the set of points of the form $E(\mathscr{M})$.

We thus obtain

Proposition 14. *For $n \geqslant 3$, let T be any n-taxon tree and $\mathfrak{A}_T$ the ideal of phylogenetic invariants for the general Markov model on T. Then there exist arrays $X_{a_1a_2\ldots a_n}$ of non-negative numbers with $X_{a_1a_2\ldots a_n} \in V(\mathfrak{A}_T)$ such that $X_{a_1a_2\ldots a_n} \neq E_{a_1a_2\ldots a_n}(\mathscr{M})$ for all choices of parameters $\mathscr{M}$, including complex ones. Thus there are no parameter-strong sets of invariants on all of $\mathbb{C}^{\kappa^n}$.*

Even when an array X is of the form $E(\mathscr{M})$, it is generally not obvious whether the parameters $\mathscr{M}$ are stochastic, or merely complex. Remark 6 indicates that if an array is sufficiently close to a phylogenetically trivial one, the parameters will at least be real. One might hope that imposing the

additional condition that all entries of X be non-negative would ensure the parameters be stochastic.

Unfortunately, choosing some random examples of positive arrays X_{abc} near a diagonal one for $\kappa = 2$ and solving for the parameters shows these sometimes turn out to involve negative quantities.

To illuminate this issue further, for the 3-taxon case with $\kappa = 2$ and any $1/2 > \epsilon > 0$ consider the parameters

$$\mathbf{p}_a = (0.5, 0.5), \quad M_{af} = M_{fb} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}, \quad M_{fc} = \begin{pmatrix} 1-\epsilon & \epsilon \\ -\epsilon^2 & 1+\epsilon^2 \end{pmatrix}$$

in which one negative entry appears. A calculation of $E_{abc}(\mathscr{M})$ shows it to be an array of positive numbers which of course lies in $V(\mathfrak{A}_T)$. Note that as $\epsilon \to 0$, $E_{abc}(\mathscr{M})$ approaches a phylogenetically trivial array.

Modifying this example, for more bases and more taxa, as above, shows

Proposition 15. *For $n \geqslant 3$, let T be any n-taxon tree and $\mathfrak{A}_T$ the ideal of phylogenetic invariants for the general Markov model on T. Then there is no open set $\mathscr{O} \subset [0,1]^{\kappa^n}$ containing the phylogenetically trivial arrays for which*

$$X_{a_1 a_2 \ldots a_n} \in V(\mathfrak{A}_T) \cap \mathscr{O}$$

implies

$$X_{a_1 a_2 \ldots a_n} = E_{a_1 a_2 \ldots a_n}(\mathscr{M}),$$

for a stochastic $\mathscr{M}$.

Note that in passing from considering only the vanishing of phylogenetic invariants to also considering the biologically natural condition that all entries of an array be non-negative, we have made an important step from only considering equalities to also considering inequalities. We have passed from algebraic geometry over $\mathbb{C}$, to algebraic geometry over $\mathbb{R}$. The question of what additional conditions involving inequalities can help distinguish those points of the form $E(\mathscr{M})$ for stochastic $\mathscr{M}$ is an interesting one.

Appendix A

Proof of Lemma 12. The key is to show there are neighborhoods of $E_{i,i}$ in which any matrix has an eigenvector close to the standard basis vector $\mathbf{e}_i$.

Each $E_{i,i}$ has as its characteristic polynomial $p(x) = (x-1)x^{\kappa-1}$. But for any $\delta_1 > 0$ (to be chosen later) there exists a $\delta_2 > 0$ such that if all coefficients of a κth degree polynomial $q(x)$ are within δ_2 of the corresponding coefficients of $p(x)$, then $q(x)$ will have a simple root $\lambda \in \mathbb{C}$ with $|1-\lambda| < \delta_1$ while all other roots ρ of $q(x)$ satisfy $|\rho| < \delta_1$. (This can be shown, for instance, by the argument principle of complex analysis.) There also exists an open set $\mathscr{O}_1^i \ni E_{i,i}$ such that if $M \in \mathscr{O}_1^i$, then the characteristic polynomial of M will have its coefficients within δ_2 of those of $p(x)$. Thus for all δ_1 with $0 < \delta_1 < 1/2$, there is an open set

$$\mathscr{O}_1^i \ni E_{i,i} \tag{21}$$

such that all matrices in $\mathcal{O}_1^i$ have exactly one eigenvalue λ (of algebraic multiplicity 1) satisfying $|\lambda - 1| < \delta_1$.

Now for any matrix M and scalar λ for which the expression makes sense, define the vector

$$\mathbf{v} = \mathbf{v}(M, \lambda) = \frac{1}{\sum_{j=1}^{n} c_{j,1}} (c_{1,1}, c_{2,1}, \ldots, c_{n,1}),$$

where $c_{j,1}$ is the $(j,1)$-cofactor of $\lambda I - M$ (i.e., $(-1)^{j+1}$ times the determinant of the matrix obtained by deleting the jth row and first column of $\lambda I - M$). Since the $c_{j,1}$ are polynomials in λ and the entries of M, the vector $\mathbf{v}$ is a continuous function of M and λ. Furthermore, if λ is an eigenvalue of M, then properties of cofactors imply $\mathbf{v}(\lambda I - M) = \mathbf{0}$. Thus, provided it is defined, $\mathbf{v}$ is an eigenvector of M with eigenvalue λ. Note also that $\mathbf{v}(E_{i,i}, 1) = \mathbf{e}_1$.

Let $\epsilon > 0$ be such that $\mathcal{Q}$ contains an ϵ-ball around I (using the Euclidean metric). Then by the continuity of $\mathbf{v}(M, \lambda)$, there are open sets $\mathcal{O}_2^i \ni E_{i,i,}$ and $\mathcal{O}_3^i \ni 1$ such that if $M \in \mathcal{O}_2^i$ and $\lambda \in \mathcal{O}_3^i$ then

$$\|\mathbf{v}(M, \lambda) - \mathbf{e}_i\| < \frac{\min(\epsilon, 1)}{\sqrt{\kappa}}.$$

Now choose $1/2 > \delta_1 > 0$ in the first paragraph sufficiently small that the δ_1-ball around 1 is contained in $\mathcal{O}_3^i$, and let $\mathcal{O}_1^i \ni E_{i,i,}$ be the open set whose existence is asserted in (21). Let $\mathcal{N}_i = \mathcal{O}_1^i \cap \mathcal{O}_2^i$. Thus $\mathcal{N}_i$ is a neighborhood of $E_{i,i}$ such that if $M \in \mathcal{N}_i$ then M has a left eigenvector $\mathbf{v}$, whose entries sum to 1, of eigenvalue λ, with $\|\mathbf{v} - \mathbf{e}_i\| < \min(\epsilon, 1)/\sqrt{\kappa}$. Furthermore, the λ-eigenspace of M is one-dimensional.

Now suppose we have matrices $M_i \in \mathcal{N}_i$, $i = 1, \ldots, \kappa$ which commute. For each M_i, let $\mathbf{v}_i$ be the eigenvector whose existence is asserted in the last paragraph, with λ_i its eigenvalue.

Since $\|\mathbf{v}_i - \mathbf{e}_i\| < 1/\sqrt{\kappa}$ for $i = 1, \ldots, \kappa$, the vectors $\mathbf{v}_i$ for $i = 1, \ldots, \kappa$ must be linearly independent. To see this, suppose they are dependent. Then there exists a vector $\mathbf{c} = (c_1, \ldots, c_\kappa)$ with $\|\mathbf{c}\| = 1$ such that the inner product $\langle \mathbf{c}, \mathbf{v}_i \rangle$ is zero for all i. Then

$$|c_i| = |\langle \mathbf{c}, \mathbf{e}_i \rangle| = |\langle \mathbf{c}, \mathbf{e}_i \rangle - \langle \mathbf{c}, \mathbf{v}_i \rangle| = |\langle \mathbf{c}, \mathbf{e}_i - \mathbf{v}_i \rangle| \leqslant \|\mathbf{c}\| \|\mathbf{e}_i - \mathbf{v}_i\| < \frac{1}{\sqrt{\kappa}}.$$

But since $\|\mathbf{c}\| = \sum_{i=1}^{\kappa} |c_i|^2$, this implies $\|\mathbf{c}\| < 1$, which is a contradiction.

Because the M_j commute, the $\mathbf{v}_i$ must in fact be simultaneous eigenvectors of all the M_j. For

$$\mathbf{v}_i M_j M_i = \mathbf{v}_i M_i M_j = \mathbf{v}_i M_j \lambda_i,$$

so $\mathbf{v}_i M_j$ lies in the one-dimensional λ_i-eigenspace of M_i, and hence $\mathbf{v}_i M_j = \mathbf{v}_i \lambda_{i,j}$ for some $\lambda_{i,j}$.

Since the M_j have a linearly independent set of κ common eigenvectors, they are simultaneously diagonalizable. Finally, if S is the matrix whose ith row is $\mathbf{v}_i$, then $S \in \mathcal{Q}$, its rows sum to 1, and it diagonalizes all the M_i. □

References

[1] J.A. Cavender, J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, J. Class. 4 (1987) 57.
[2] J.A. Lake, A rate independent technique for analysis of nucleic acid sequences: evolutionary parsimony, Mol. Bio. Evol. 4 (1987) 167.
[3] T.R. Hagedorn, L.F. Landweber, Phylogenetic invariants and geometry, J. Theor. Biol. 205 (2000) 365.

[4] B. Chor, M.D. Hendy, B.R. Holland, D. Penny, Multiple maxima of likelihood in phylogenetic trees: an analytic approach, Mol. Bio. Evol. 17 (2000) 1529.
[5] V. Ferretti, D. Sankoff, The empirical discovery of phylogenetic invariants, Adv. Appl. Probab. 25 (2) (1993) 290.
[6] V. Ferretti, D. Sankoff, A remarkable nonlinear invariant for evolution with heterogeneous rates, Math. Biosci. 134 (1) (1996) 71.
[7] V. Ferretti, D. Sankoff, Phylogenetic invariants for more general evolutionary models, J. Theor. Biol. 173 (1995) 147.
[8] S.N. Evans, T.P. Speed, Invariants of some probability models used in phylogenetic inference, Ann. Stat. 21 (1) (1993) 355.
[9] M. Steel, L. Székely, P.L. Erdös, P. Waddell, A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model, N.Z. J. Botany 31 (31) (1993) 289.
[10] M.D. Hendy, D. Penny, Spectral analysis of phylogenetic data, J. Class. 10 (1993) 1.
[11] M. Steel, M.D. Hendy, D. Penny, Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results, Discr. Appl. Math. 88 (1–3) (1998) 367.
[12] M. Steel, Recovering a tree from the leaf colourations it generates under a Markov model, Appl. Math. Lett. 7 (2) (1994) 19.
[13] C. Semple, M. Steel, Tree representations of non-symmetric group-valued proximities, Adv. Appl. Math. 23 (3) (1999) 300.
[14] T.R. Hagedorn, A combinatorial approach to determining phylogenetic invariants for the general model. Technical Report 2671, CRM, 2000.
[15] S.N. Evans, X. Zhou, Constructing and counting phylogenetic invariants, J. Comp. Bio. 5 (4) (1998) 713.
[16] T.R. Hagedorn, Determining the number and structure of phylogenetic invariants, Adv. Appl. Math. 24 (1) (2000) 1.
[17] J.T. Chang, Full reconstruction of Markov models on evolutionary trees: identifiability and consistency, Math. Biosci. 137 (1) (1996) 51.
[18] M.A. Steel, L. Székely, M.D. Hendy, Reconstructing trees from sequences whose sites evolve at variable rates, J. Comp. Bio. 1 (2) (1994) 153.
[19] J.M. Landsberg, L. Manivel, On the ideals of secant varieties of Segre varieties, 2003, preprint.
[20] R.M. Guralnick, B.A. Sethuraman, Commuting pairs and triples of matrices and related varieties, Lin. Alg. Appl. 310 (2000) 139.
[21] R.M. Guralnick, A note on commuting pairs of matrices, Lin. Multilin. Alg. 31 (1992) 71.